

SPLITTING THE SMOOTHED PRIMAL-DUAL GAP: OPTIMAL ALTERNATING DIRECTION METHODS

QUOC TRAN-DINH* AND VOLKAN CEVHER*

Abstract. We develop rigorous alternating direction optimization methods for a prototype constrained convex optimization template, which has broad applications in computational sciences. We build upon our earlier work on the model-based gap reduction (MGR) technique, which revolves around a smoothed estimate of the primal-dual gap. MGR allows us to simultaneously update a sequence of primal and dual variables as well as primal- and dual-smoothness parameters so that the smoothed gap function converges to the true gap, which in turn converges to zero—both at optimal rates. In contrast, this paper introduces a new split-gap reduction (SGR) technique as a natural counterpart of MGR in order to take advantage of additional splitting structures present in the prototype template. We illustrate SGR technique using the forward-backward and Douglas-Rachford splittings on the smoothed gap function and derive new alternating direction methods. The new methods obtain optimal convergence rates without heuristics and eliminate the infamous penalty parameter tuning issue in the existing alternating direction methods. Finally, we verify the performance of our methods in comparison to the existing state-of-the-art and the new theoretical performance bounds via numerical examples.

Key words. Alternating minimization algorithm (AMA), alternating direction method of multipliers (ADMM), augmented Lagrangian, primal-dual first-order method, constrained convex optimization.

1. Introduction. A broad set of applications in data sciences result in convex optimization problems that are concisely captured by the following template [6, 7, 9]:

$$(1.1) \quad f^* := \min_{x \in \mathbb{R}^p} \{f(x) : Mx = c, \quad x \in \mathcal{X}\},$$

where $f : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{+\infty\}$ is a proper, closed and convex function, $M \in \mathbb{R}^{n \times p}$, $c \in \mathbb{R}^n$, and \mathcal{X} is a nonempty, closed and convex set in \mathbb{R}^p . Intriguingly, many problems of practical interest feature dimensions so large that are beyond the grasp of the accurate interior point methods. As a result, there is a renewed interest in the flexible, primal-dual first order methods that provide additional scalability and accuracy trade-offs.

Recently, [32] introduced model-based gap reduction technique, which establishes a principled framework for developing tuning-free primal-dual algorithms with rigorous convergence guarantees for (1.1). The key ideas in [32] can be summarized as follows: We first measure the duality gap for (1.1) using a smoothed gap function parameterized by two smoothness parameters γ and β , and the primal variable x and the dual one λ :

$$(1.2) \quad G_{\gamma\beta}(x, \lambda) := \max_{\tilde{\lambda} \in \mathbb{R}^n} \left\{ f(x) + \langle \tilde{\lambda}, Mx - c \rangle - \frac{\beta}{2} \|\tilde{\lambda}\|_2^2 \right\} - \min_{\tilde{x} \in \mathcal{X}} \left\{ f(\tilde{x}) + \langle \lambda, M\tilde{x} - c \rangle + \frac{\gamma}{2} b(\tilde{x}) \right\},$$

where $b(\cdot)$ is a smoothing function for \mathcal{X} . Note that $G_{\gamma\beta}(w)$ with $w := (x, \lambda)$ itself is not fully smooth and has a composite form. When γ and β are both zeros, $G_{\gamma\beta}(\cdot)$ measures the duality gap, since the max problem estimates the primal objective whereas the remaining term provides us the dual objective. While [32] considered general smoothing functions, two special cases lead to salient computational trade-offs in primal-dual optimization: (i) $b(x) := \frac{1}{2} \|x\|_2^2$ and (ii) $b(x) := \frac{1}{2} \|Mx - c\|_2^2$. In the sequel, we refer to the former as proximity smoother and the latter as the augmented Lagrangian smoother.

Given the choice of the smoothing functions, the authors in [32] then constructed a sequence $(\gamma_k, \beta_k, \bar{x}^k, \bar{\lambda}^k)$ to decrease the smoothed gap function $G_{\gamma\beta}(w)$ at a rate, parameterized by τ_k^2 , while simultaneously reducing the smoothness parameters to obtain an optimal primal-dual solution (x^*, λ^*) . For instance, when we use the proximity smoothing function, the sequence $\{\bar{x}^k\} \subset \mathcal{X}$ satisfies $|f(\bar{x}^k) - f^*| = \mathcal{O}(\gamma_k)$ on the primal objective residual, and $\|M\bar{x}^k - c\| = \mathcal{O}(\beta_k)$ on the feasibility gap separately with explicit

*Laboratory for Information and Inference Systems (LIONS), École Polytechnique Fédérale de Lausanne (EPFL), CH1015 - Lausanne, Switzerland.
E-mail: {quoc.trandinh, volkan.cevher}@epfl.ch.

constants. When we instead use the augmented Lagrangian smoothing, we can obtain $|f(\bar{x}^k) - f^*| = \mathcal{O}(\beta_k)$ and $\|M\bar{x}^k - c\| = \mathcal{O}(\beta_k)$ with the fixed parameter $\gamma_k := \gamma_0 > 0$.

While it may appear that model-based gap reduction technique can drive the smoothing constants towards zero to obtain arbitrarily fast convergence rates, the iterates must in fact satisfy a fundamental *uncertainty principle*: $\gamma_k \beta_k = \Omega(\tau_k^2)$, where τ_k^2 is the rate at which the unconstrained functions (e.g., $G_{\gamma\beta}$) can be minimized. Due to the smoothing technique, $G_{\gamma\beta}$ is a composite convex function with a Lipschitz gradient smooth part, and hence, we have $\tau_k^2 = \Omega(\frac{1}{k^2})$ as the iteration complexity lower bound. Fortunately, the constructed sequences obtain this optimal rate for both choices of the smoothing functions. For instance, with the proximity smoother, we can choose $\gamma_k = \mathcal{O}(\frac{1}{k})$ and $\beta_k = \mathcal{O}(\frac{1}{k})$ and then each iteration requires two proximal operator computations (one each for the primal and the dual), and one application of A and A^T . For the augmented Lagrangian smoother, we can achieve faster rates $\beta_k = \mathcal{O}(\frac{1}{k^2})$ with *significantly better constants*; however, the per-iteration complexity increases commensurately.

Splitting the smoothed gap. Often times, the template (1.1) can be further refined as:

$$(1.3) \quad f^* := \min_{x:=(u,v)} \{f(x) := g(u) + h(v) : Au + Bv = c, u \in \mathcal{U}, v \in \mathcal{V}\},$$

where $g: \mathbb{R}^{p_1} \rightarrow \mathbb{R} \cup \{+\infty\}$ and $h: \mathbb{R}^{p_2} \rightarrow \mathbb{R} \cup \{+\infty\}$ are two proper, closed and convex functions ($p := p_1 + p_2$), $A \in \mathbb{R}^{n \times p_1}$, $B \in \mathbb{R}^{n \times p_2}$, $c \in \mathbb{R}^n$, and \mathcal{U} and \mathcal{V} are two nonempty, closed and convex sets in \mathbb{R}^{p_1} and \mathbb{R}^{p_2} , respectively. As compared to (1.1), the primal variable x and the objective function f are now *split* into two parts with dimensions p_1 and p_2 . While many problems naturally lend themselves to the formulation (1.3), we can also introduce the splitting artificially to take advantage of special computational properties of the linear operators A and B or the proximal operators of g and h .

To this end, this paper proposes a splitting counterpart of the model-based gap reduction technique of [32] in order to exploit the specific splitting structure in (1.3). As a result, we develop optimal alternating direction methods that benefit from the same type of rigorous convergence guarantees. The new algorithmic sequences revolve around the Fenchel dual of (1.3) essentially in the same manner as the classical alternating minimization algorithm (AMA) and the alternating direction method-of-multiplier (ADMM) methods:

$$(1.4) \quad \psi^* := \max_{\lambda \in \mathbb{R}^n} \{\psi(\lambda) := -g_{\mathcal{U}}^*(-A^T \lambda) - h_{\mathcal{V}}^*(-B^T \lambda) + \langle c, \lambda \rangle\},$$

where $g_{\mathcal{U}}^*$ and $h_{\mathcal{V}}^*$ are the Fenchel conjugate [30] of $g_{\mathcal{U}}(\cdot) := g(\cdot) + \delta_{\mathcal{U}}(\cdot)$ and $h_{\mathcal{V}}(\cdot) := h(\cdot) + \delta_{\mathcal{V}}(\cdot)$, and $\delta_{\mathcal{U}}$ and $\delta_{\mathcal{V}}$ are the indicator functions of \mathcal{U} and \mathcal{V} , respectively. Indeed, we will rigorously illustrate how our augmented Lagrangian smoothing technique plays a key role in eliminating the infamous parameter tuning issues in conjunction with powerful forward-backward as well as Douglas-Rachford splitting techniques.

1.1. Related work. The theory behind primal-dual methods is the minmax or the saddle point principle in convex analysis [30], which enables us to develop numerical methods for solving (1.3) [5, 15]. Intriguingly, primal-dual methods can also viewed within variational inequalities or maximal monotone inclusions so that mathematical tools from such fields can be applied [1, 8, 9, 11, 14, 19, 23]. Among the general primal-dual algorithms, simple alternating direction methods have gained popularity mostly due to their robustness in addition to their efficiency in large scale problems.

Alternating direction methods include AMA and ADMM as two important special cases. The interest in these methods have been rapidly growing in applications; however, to the best of our knowledge, the supporting theory remains largely incomplete. While it is impossible to exhaustively review the substantial amount of work that has gone into these methods, we would like to point out the recent reviews and progress: *cf.*, [6, 9, 12, 14, 19, 17, 28, 31, 34, 36] and the references quoted therein. We now recall the standard AMA and ADMM algorithms here to clarify our contributions in the sequel.

The standard AMA. The standard AMA solves (1.3) by using one Lagrange dual step and one augmented Lagrangian dual step between two groups of variable u and v [35]. The main steps of this algorithm can be presented as follows:

$$(1.5) \quad \begin{cases} \hat{u}^{k+1} & := \operatorname{argmin}_{u \in \mathcal{U}} \{g(u) + \langle \hat{\lambda}^k, Au \rangle\}, \\ \hat{v}^{k+1} & := \operatorname{argmin}_{v \in \mathcal{V}} \{h(v) + \langle \hat{\lambda}^k, Bv \rangle + \frac{\eta_k}{2} \|A\hat{u}^{k+1} + Bv - c\|_2^2\}, \\ \hat{\lambda}^{k+1} & := \hat{\lambda}^k + \eta_k (A\hat{u}^{k+1} + B\hat{v}^{k+1} - c), \end{cases}$$

where $\eta_k > 0$ is the penalty parameter. We can view AMA as the forward-backward splitting algorithm applied to the dual problem (1.4) (*cf.*, [17, 35]). The AMA is guaranteed to converge when g is strongly convex or when $g_{\mathcal{U}}^*$ has Lipschitz gradient [17].

The standard ADMM method. ADMM generates a primal-dual sequence as follows:

$$(1.6) \quad \begin{cases} \hat{u}^{k+1} & := \operatorname{argmin}_{u \in \mathcal{U}} \{g(u) + \langle \hat{\lambda}^k, Au \rangle + \frac{\eta_k}{2} \|Au + B\hat{v}^k - c\|_2^2\} \\ \hat{v}^{k+1} & := \operatorname{argmin}_{v \in \mathcal{V}} \{h(v) + \langle \hat{\lambda}^k, Bv \rangle + \frac{\eta_k}{2} \|A\hat{u}^{k+1} + Bv - c\|_2^2\} \\ \hat{\lambda}^{k+1} & := \hat{\lambda}^k + \eta_k (A\hat{u}^{k+1} + B\hat{v}^{k+1} - c), \end{cases}$$

where $\eta_k > 0$ is a penalty parameter. Unlike AMA, the ADMM algorithm splits the full augmented Lagrangian $\mathcal{L}_\eta(u, v, \lambda) := g(u) + h(v) + \langle \lambda, Au + Bv - c \rangle + \frac{\eta}{2} \|Au + Bv - c\|_2^2$ in order to construct the algorithm by alternating between the two variables u and v [14, 16]. The parameter η_k can be fixed at a certain level or adaptively updated with *heuristic strategies* to obtain a desired performance. The ADMM algorithm can be viewed as the Douglas-Rachford splitting method applying to the dual problem (1.4) (*cf.*, [14, 16]).

ADMM is more widely studied or applied as compared to AMA since AMA requires strong convexity of the term g . Convergence theory, modifications, accelerations, and extensions of the standard ADMM (1.6) have been actively studied in the literature: *cf.*, [6, 12, 14, 17, 19, 28, 31, 36, 9, 13]. However, none of these works shown the optimal convergence rate in the sense of first-order black box oracles [24] under mild assumptions for both the objective residual $|f(\hat{x}^k) - f^*|$ and the primal feasibility gap $\|A\hat{u}^k + B\hat{v}^k - c\|$ without using an averaging scheme and additional assumptions (e.g., gradient Lipschitz assumption). Perhaps, the main reason of this limitation is that the main schemes (1.5) and (1.6) remain essentially the same in virtually all these works.

1.2. Our contributions. The methods we develop based on the new split-gap reduction technique in this paper ultimately alters the primal-dual updates $\{(\bar{x}^k, \bar{\lambda}^k)\}$ by the new theory coming from the smoothed gap perspective. Note that while our approach is a natural counterpart of the model-based gap reduction technique of [32], it is also related but different from the excessive gap reduction technique by Nesterov in [26] (*cf.*, [32] for details). Our main contributions are summarized as follows:

- (a) We propose a new primal-dual approach for developing alternating direction optimization methods based on a new *split-gap reduction technique*. To this end, we unify the smoothing technique of Nesterov for structured, unconstrained minimization with the powerful forward-backward and Douglas-Rachford splitting techniques. As a result, we develop two new alternating direction methods for (1.3), which are different from the standard AMA and ADMM (1.5) and (1.6).
- (b) We derive update rules for all the algorithmic parameters including the penalty parameters in a heuristic-free fashion where the parameter choices have explicit impacts on the convergence guarantees.
- (c) We rigorously characterize the $\mathcal{O}(1/k)$ - convergence rate of the two algorithms for both the primal objective residual $|f(\bar{x}^k) - f^*|$ and the primal feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$. We show that this convergence rate is optimal under mild assumptions in the sense of first-order black-box models [24].
- (d) We also develop different variants of both algorithms to exploit additional assumptions on A or B , g and h whenever they are available.

Let us emphasize the following key advantages of the proposed algorithms. First, our algorithms can solve a wide class of constrained convex problems (1.3), where we do not require any particularly strong assumption on g , h , \mathcal{U} and \mathcal{V} , except for the convexity, proximal tractability, and the boundedness of \mathcal{U} and \mathcal{V} . Second, the computational cost-per-iteration of the algorithms is fundamentally the same as the standard AMA and ADMM versions (1.5) and (1.6). Third, the algorithms do not use the diameter of \mathcal{U} and \mathcal{V} and the desired accuracy ε at any step as in existing accelerated primal-dual methods, while updating all the parameters automatically at each iteration [3, 21, 27]. Finally, we show how the choice of the penalty parameters trades off the convergence guarantee in the objective residual $|f(\bar{x}^k) - f^*|$ and the primal feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$.

Paper organization. Section 2 briefly presents a primal-dual formulation of problem (1.3) under basic assumptions, and characterizes its optimality condition. Section 3 deals with a smoothing technique for the primal-dual gap function. Section 4 presents a new AMA-like algorithm and analyzes its convergence. The two special cases are also studied in this section. Section 5 is devoted to developing a new ADMM-like algorithm and analyzes its convergence. Section 6 provides implementation remarks and some extensions. Section 7 presents numerical experiments to verify the performance of our algorithms. We conclude with a summary of our main results. For clarity of exposition, several technical and new proofs are moved to the appendix.

Notations and Terminology. We work on the real 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 T for both the transpose and adjoint operators, and it can be recognized from the context. 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 $\text{ri}(\mathcal{X})$ for its relative interior. We also use \mathbb{R}_{++} for the set of positive real numbers. For a given symmetric matrix \mathbf{X} , $\lambda_{\min}(\mathbf{X})$ and $\lambda_{\max}(\mathbf{X})$ denote the smallest and the largest eigenvalues of \mathbf{X} , respectively.

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

$$(1.7) \quad \text{prox}_{\varphi}(x) := \underset{z}{\operatorname{argmin}} \{ \varphi(z) + (1/2)\|z - x\|^2 \}.$$

Generally, computing prox_{φ} is intractable. However, if prox_{φ} can be computed efficiently (e.g., in a closed form or in polynomial time), then we say that φ has a *tractable* proximity operator. Examples of such convex functions can be found, e.g., in [1, 29].

2. Preliminaries.

2.1. Lagrangian primal-dual formulation. Let $x := [u, v] \equiv (u^T, v^T)^T \in \mathbb{R}^p$ and $\mathcal{X} := \mathcal{U} \times \mathcal{V}$ be the joint variable and the joint domain of u and v , respectively. Let $\mathcal{D} := \mathcal{X} \cap \{(u, v) : Au + Bv = c\}$ be the feasible set of (1.3). We define the Lagrange function of (1.3) 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 vector of Lagrange multipliers.

2.2. The dual problem. Using \mathcal{L} , we can write the dual problem of (1.3) as:

$$(2.1) \quad d^* := \max_{\lambda \in \mathbb{R}^n} d(\lambda),$$

where d is the dual function defined by:

$$(2.2) \quad d(\lambda) := \min_{(u,v) \in \mathcal{X}} \{ g(u) + h(v) + \langle \lambda, Au + Bv - c \rangle \}.$$

Due to the splitting of $f = g + h$, the dual function d decomposes into the sum of three individual components, i.e.: $d(\lambda) = d_0^1(\lambda) + d_0^2(\lambda) - \langle c, \lambda \rangle$, where:

$$(2.3) \quad \begin{cases} d_0^1(\lambda) & := \min_{u \in \mathcal{U}} \{ g(u) + \langle A^T \lambda, u \rangle \}, \\ d_0^2(\lambda) & := \min_{v \in \mathcal{V}} \{ h(v) + \langle B^T \lambda, v \rangle \}. \end{cases}$$

Let us denote by $u^*(\lambda)$ and $v^*(\lambda)$ one solution of these subproblems, respectively. These dual components are concave, but generally nonsmooth. Numerical methods such as subgradient algorithms for directly solving (2.1) are inefficient [24, 25].

2.3. Our assumptions. For characterizing the relation between the primal problem (1.3) and the dual one (2.1), we require the following assumptions:

ASSUMPTION A. 1. *The functions g and h are proper, closed and convex. The domains \mathcal{U} and \mathcal{V} are nonempty, closed and convex. The solution set \mathcal{X}^* of (1.3) is nonempty. Either \mathcal{X} is a polytope or the Slater condition (2.4) holds. In addition, both \mathcal{U} and \mathcal{V} are bounded*

We say that the *Slater condition* holds for (1.3) if we have

$$(2.4) \quad \text{ri}(\mathcal{X}) \cap \{(u, v) : Au + Bv = c\} \neq \emptyset,$$

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

Except for the boundedness of \mathcal{U} and \mathcal{V} , the rest of the assumptions are standard for any primal-dual method. We argue that the boundedness assumption is not too restrictive. For instance, the diameters of the domains \mathcal{U} and \mathcal{V} do not enter into any step of the proposed algorithms. They only appear in the convergence guarantee bounds.

2.4. Zero duality gap. Under Assumption A.1, the solution set Λ^* of the dual problem (2.1) is nonempty and bounded. Moreover, *strong duality* holds, i.e., $f^* - d^* = 0$. From the classical duality theory, we have $d(\lambda) \leq f(x)$ for any feasible primal-dual point (x, λ) . Hence, the primal-duality gap function G defined by:

$$(2.5) \quad G(w) := f(x) - d(\lambda) \geq 0, \quad \forall x \in \mathcal{D}, \forall \lambda \in \mathbb{R}^n,$$

where $w := [x, \lambda]$. Clearly, $G(w^*) = 0$ (zero duality gap) for any primal-dual solution $w^* = [x^*, \lambda^*] \in \mathcal{X}^* \times \Lambda^*$. In addition, w^* is a saddle point of the Lagrange function. That is, we have $\mathcal{L}(x^*, \lambda) \leq \mathcal{L}(x^*, \lambda^*) = f^* = d^* \leq \mathcal{L}(x, \lambda^*)$ for all $x \in \mathcal{X}$ and $\lambda \in \mathbb{R}^n$.

3. Smoothing techniques. The dual function d defined by (2.2) is concave, but generally does not possess useful properties for designing algorithms. Our first idea is to replace the first component d_0^1 of d in (2.3) by the following modification so that it leads to a new approximation of d , which possesses necessary properties to develop the algorithms. More precisely, we define the following smoothed function:

$$(3.1) \quad d_\gamma^1(\lambda) := \min_{u \in \mathcal{U}} \{g_\gamma(u) := g(u) + \langle A^T \lambda, u \rangle + \frac{\gamma}{2} \|A(u - \bar{u}_c)\|^2\},$$

where $\gamma > 0$ is a smoothing parameter, and \bar{u}_c is a given center point in \mathcal{U} .

REMARK 3.1. *The presence of A in the smoothing term $\frac{\gamma}{2} \|A(u - \bar{u}_c)\|^2$ of (3.1) prevents us from using prox_g for evaluating d_γ^1 at this moment. But, we will show in the sequel when we can still use this proximity operator.*

We now define the following quantities that govern the global efficiency of algorithms:

$$(3.2) \quad \begin{cases} D_{\Lambda^*} & := \min\{\|\lambda^*\| : \lambda^* \in \Lambda^*\}, \\ D_{\mathcal{U}}^A & := \frac{1}{2} \max\{\|A(u - \bar{u}_c)\|^2 : u \in \mathcal{U}\}, \\ D_{\mathcal{X}} & := \frac{1}{2} \max\{\|Au + Bv - c\|^2 : u \in \mathcal{U}, v \in \mathcal{V}\}. \end{cases}$$

Clearly, under Assumption A.1, such quantities are well-defined and bounded. We first investigate the properties of d_γ^1 in the following lemma.

LEMMA 3.2. *The minimization subproblem in (3.1) always admits an optimal solution $u_\gamma^*(\lambda)$ for any $\lambda \in \mathbb{R}^n$. Moreover, the function d_γ^1 is well-defined, concave, and smooth. Its gradient is given by $\nabla d_\gamma^1(\lambda) = Au_\gamma^*(\lambda)$, which is Lipschitz continuous with the Lipschitz constant $L_{d_\gamma^1} := \gamma^{-1} > 0$.*

In addition, the function d^1 satisfies the following estimates:

$$(3.3) \quad d_\gamma^1(\lambda) - \gamma D_{\mathcal{U}}^A \leq d^1(\lambda) \leq d_\gamma^1(\lambda), \quad \forall \lambda \in \mathbb{R}^n,$$

$$(3.4) \quad d_{\hat{\gamma}}^1(\bar{\lambda}) \leq d_\gamma^1(\bar{\lambda}) + \frac{1}{2}(\hat{\gamma} - \gamma) \|A(u_\gamma^*(\bar{\lambda}) - \bar{u}_c)\|^2, \quad \forall \hat{\gamma}, \gamma > 0,$$

where $u_\gamma^*(\cdot)$ is the solution in (3.1), and $D_{\mathcal{U}}^A$ is defined by (3.2).

Proof. Let us consider $\varphi(s) := \inf \{g(u) : Au = s, u \in \mathcal{U}\}$. Under Assumption A.1, the set $\mathcal{S} := \{s : Au = s, u \in \mathcal{U}\}$ is nonempty, closed, convex, and bounded. Hence, φ is proper, closed and convex. We can write the function d_γ^1 defined by (3.1) as follows:

$$(3.5) \quad d_\gamma^1(\lambda) = \min_{s \in \mathcal{S}} \left\{ \varphi(s) + \langle \lambda, s \rangle + \frac{\gamma}{2} \|s - \bar{s}_c\|^2 \right\},$$

where $s := Au \in \mathcal{S}$ and $\bar{s}_c := A\bar{u}_c$. Clearly, the function inside the min operator is strongly convex with the convexity parameter $\gamma > 0$. Hence, d_γ^1 is well-defined, concave and smooth. Its gradient is $\nabla d_\gamma^1(\lambda) = s_\gamma^*(\lambda) = Au_\gamma^*(\lambda)$. The Lipschitz continuity of ∇d_γ^1 can be proved similarly as Theorem 1 in [27].

The estimate (3.3) is obvious from the definition of (2.3) and (3.1). We consider the function $\psi(u, \gamma; \lambda) := f(u) + \langle \lambda, Au \rangle + (\gamma/2) \|A(u - \bar{u}^c)\|^2$. Then, for fixed $\lambda \in \mathbb{R}^n$, $\psi(\cdot, \cdot; \lambda)$ is convex w.r.t. u and linear w.r.t. $\gamma > 0$. Since $d_\gamma^1(\bar{\lambda}) = \min \{\psi(u, \gamma; \bar{\lambda}) : u \in \mathcal{U}\}$, $d_\gamma^1(\bar{\lambda})$ is concave w.r.t. $\gamma > 0$ and smooth. Moreover, $\frac{d_\gamma^1(\bar{\lambda})}{d\gamma} = \frac{1}{2} \|A(u_\gamma^*(\bar{\lambda}) - \bar{u}^c)\|^2$. Hence, by the concavity of $d_\gamma^1(\bar{\lambda})$ w.r.t. $\gamma > 0$, we have $d_{\hat{\gamma}}^1(\bar{\lambda}) \leq d_\gamma^1(\bar{\lambda}) + \frac{1}{2}(\hat{\gamma} - \gamma) \|A(u_\gamma^*(\bar{\lambda}) - \bar{u}^c)\|^2$, which is indeed (3.4). \square

Let $\gamma > 0$, $\beta > 0$, d_γ^1 be defined by (3.1) and d_0^2 be defined by (2.3). We consider the following functions:

$$(3.6) \quad \begin{cases} d_\gamma(\lambda) & := d_\gamma^1(\lambda) + d_0^2(\lambda) - \langle c, \lambda \rangle, \\ f_\beta(x) & := f(x) + \frac{1}{2\beta} \|Au + Bv - c\|^2, \\ G_{\gamma\beta}(w) & := f_\beta(x) - d_\gamma(\lambda). \end{cases}$$

Clearly, since \mathcal{U} is bounded, if $\gamma \downarrow 0^+$, then $d_\gamma(\lambda) \rightarrow d(\lambda)$. Hence, d_γ is an approximation of d . For any feasible point $x = [u, v] \in \mathcal{D}$, we have $f_\beta(x) = f(x)$. Hence, f_β is an approximation to f near the feasible set \mathcal{D} . Consequently, the function $G_{\gamma\beta}(\cdot)$ can be considered as an approximation of the primal-dual gap function $G(\cdot)$ defined by (2.5). Moreover, $G_{\gamma\beta}(\cdot)$ is convex w.r.t. (x, λ) .

The following lemma shows us how to use the approximate gap function $G_{\gamma\beta}$ to characterize the primal-dual solution for (1.3)-(2.1).

LEMMA 3.3. *Let $\{\bar{w}^k\}_{k \geq 0}$ be an arbitrary sequence in $\mathcal{X} \times \mathbb{R}^n$ and $\{(\gamma_k, \beta_k)\}_{k \geq 0}$ be a sequence in \mathbb{R}_{++}^2 . Then the following estimates hold:*

$$(3.7) \quad \begin{cases} -D_{\Lambda^*} \|A\bar{u}^k + B\bar{v}^k - c\| \leq f(\bar{x}^k) - f^* \leq \bar{G}_k + \gamma_k D_{\mathcal{U}}^A, \\ \|A\bar{u}^k + B\bar{v}^k - c\| \leq 2\beta_k D_{\Lambda^*} + \sqrt{2\beta_k (\bar{G}_k + \gamma_k D_{\mathcal{U}}^A)}, \\ d^* - d(\bar{\lambda}^k) \leq 2\beta_k D_{\Lambda^*}^2 + D_{\Lambda^*} \sqrt{2\beta_k (\bar{G}_k + \gamma_k D_{\mathcal{U}}^A)}, \end{cases}$$

where $\bar{G}_k := G_{\gamma_k \beta_k}(\bar{w}^k)$, and D_{Λ^*} and $D_{\mathcal{U}}^A$ are defined by (3.2).

Proof. We note that the function d_γ defined by (3.6) satisfies $d_\gamma(\lambda) - \gamma D_{\mathcal{U}}^A \leq d(\lambda) \leq d_\gamma(\lambda)$ for any $\lambda \in \mathbb{R}^n$ due to (3.3). Using this inequality and the definition (3.6) of f_β we have:

$$(3.8) \quad \begin{aligned} f(x) - d(\lambda) & \stackrel{(3.6)+(3.3)}{\leq} f_\beta(x) - d_\gamma(\lambda) + \gamma D_{\mathcal{U}}^A - \frac{1}{2\beta} \|Au + Bv - c\|^2 \\ & = G_{\gamma\beta}(w) + \gamma D_{\mathcal{U}}^A - \frac{1}{2\beta} \|Au + Bv - c\|^2. \end{aligned}$$

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

$$(3.9) \quad -\|\lambda^*\| \|Au + Bv - c\| \leq f(x) - f^* \leq f(x) - d(\lambda).$$

Combining (3.8) and (3.9), we obtain the first inequality of (3.7). Let $t := \|Au + Bv - c\|$. By using (3.8) and (3.9), we can see that $\frac{1}{2\beta}t^2 - \|\lambda^*\|t - [G_{\gamma\beta}(w) + \gamma D_{\mathcal{U}}^A] \leq 0$. Solving this quadratic inequation w.r.t. t and noting that $t \geq 0$, we obtain the second bound of (3.7). The last estimate of (3.7) is a direct consequence of (3.9) and the second one of (3.7). \square

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

DEFINITION 3.4. *Given an accuracy $\varepsilon > 0$, a primal-dual point $(\bar{x}^k, \bar{\lambda}^k) \in \mathcal{X} \times \mathbb{R}^n$ is said to be an ε -solution of (1.3)-(2.1) if $|f(\bar{x}^k) - f^*| \leq \varepsilon$ and $\|A\bar{u}^k + B\bar{v}^k - c\| \leq \varepsilon$.*

In practice, we usually meet the cases where the domain \mathcal{X} is simple (e.g., box, ball, cone, or simplex) so that $\bar{x}^k \in \mathcal{X}$ can be guaranteed via a closed form projection onto \mathcal{X} .

The goal is to generate a primal-dual sequence $\{\bar{w}^k\}$ and a parameter sequence $\{(\gamma_k, \beta_k)\}$ in Lemma 3.3 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 $|f(\bar{x}^k) - f^*|$ 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. Forward-backward splitting with the smoothed gap. We propose a new alternating direction method based on applying the forward-backward splitting in the smoothed gap function. We study the new algorithm, which is the natural analog of the AMA, in three steps: initialization, main steps, and parameter updates.

4.1. Computing an initial point. The first step of our AMA algorithm is to show that there exists a point $\bar{w}^0 := (\bar{u}^0, \bar{v}^0, \bar{\lambda}^0)$ such that $G_{\gamma_0\beta_0}(\bar{w}^0) \leq \eta_0 D_{\mathcal{X}}$ for given $\gamma_0, \beta_0, \eta_0 \in \mathbb{R}_{++}$. This point is constructed as follows:

$$(4.1) \quad \begin{cases} \bar{u}^0 & := \operatorname{argmin} \{g_{\gamma_0}(u) : u \in \mathcal{U}\}, \\ \bar{v}^0 & := \operatorname{argmin} \{h(v) + \frac{\eta_0}{2} \|A\bar{u}^0 + Bv - c\|^2 : v \in \mathcal{V}\}, \\ \bar{\lambda}^0 & := \eta_0(A\bar{u}^0 + B\bar{v}^0 - c), \end{cases}$$

where $\gamma_0 > 0$ is given parameter, η_0 is chosen accordingly to γ_0 , and $g_{\gamma_0}(\cdot) := g(\cdot) + \frac{\gamma_0}{2} \|A(\cdot - \bar{u}_c)\|^2$ defined by (3.1). The following lemma provides conditions for choosing γ_0, β_0 and η_0 , whose proof is in Appendix A.2.2.

LEMMA 4.1. *Let $\bar{w}^0 := (\bar{u}^0, \bar{v}^0, \bar{\lambda}^0)$ be the point computed by (4.1) for given $\gamma_0 > 0$ and $\eta_0 > 0$. Let $G_{\gamma\beta}$ be defined by (3.6). Then, for any $\beta_0 > 0$, \bar{w}^0 satisfies:*

$$(4.2) \quad G_{\gamma_0\beta_0}(\bar{w}^0) \leq \eta_0 D_{\mathcal{X}} + \frac{1}{2} \left[\frac{1}{\beta_0} - \frac{(3\gamma_0 - \eta_0)\eta_0}{\gamma_0} \right] \|A\bar{u}^0 + B\bar{v}^0 - c\|^2 - \frac{\gamma_0}{2} \|A(\bar{u}^0 - \bar{u}_c)\|^2.$$

Consequently, if $3\gamma_0 > \eta_0$ and $\beta_0 \geq \frac{\gamma_0}{(3\gamma_0 - \eta_0)\eta_0}$, then $G_{\gamma_0\beta_0}(\bar{w}^0) \leq \eta_0 D_{\mathcal{X}}$.

REMARK 4.2. *We note that we can choose an arbitrarily initial point $\bar{w}^0 := (\bar{u}^0, \bar{v}^0, \bar{\lambda}^0) \in \mathcal{X} \times \mathbb{R}^n$ for our algorithms below. However, the convergence guarantee bound of the algorithms depends on the value $G_{\gamma_0\beta_0}(\bar{w}^0)$ of the smoothed gap function.*

4.2. The main steps of the smooth alternating minimization. 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 step of our smooth alternating minimization (SAM) algorithm consists of one primal alternating step and one dual step as follows:

$$(4.3) \quad \begin{cases} \hat{u}^{k+1} & := \operatorname{argmin}_{u \in \mathcal{U}} \{g_{\gamma_{k+1}}(u) + \langle A^T \hat{\lambda}^k, u \rangle\}, \\ \hat{v}^{k+1} & := \operatorname{argmin}_{v \in \mathcal{V}} \{h(v) + \langle B^T \hat{\lambda}^k, v \rangle + \frac{\eta_k}{2} \|A\hat{u}^{k+1} + Bv - c\|^2\}, \\ \hat{\lambda}^{k+1} & := \hat{\lambda}^k + \eta_k(A\hat{u}^{k+1} + B\hat{v}^{k+1} - c), \end{cases}$$

where γ_{k+1} and η_k are referred to as the smoothness and the penalty parameter, respectively, and $g_\gamma(\cdot) := g(\cdot) + \frac{\gamma}{2}\|A(\cdot - \bar{u}_c)\|^2$ defined by (3.1). Clearly, when $\gamma_{k+1} = 0$ and $\hat{\lambda}^{k+1} = \bar{\lambda}^{k+1}$, (4.3) collapses to the standard AMA scheme (1.5).

In the SAM main step (4.3), we need to solve two convex subproblems in the first and the second lines. If $A = \mathbb{I}$, the identity matrix, or orthonormal, then computing \hat{u}^{k+1} reduces to computing the proximal operator of $g_{\mathcal{U}} := g + \delta_{\mathcal{U}}$, i.e.:

$$\hat{u}^{k+1} = \text{prox}_{\gamma_{k+1}^{-1}g_{\mathcal{U}}} \left(\bar{u}_c - \gamma_{k+1}^{-1}A^T\hat{\lambda}^k \right).$$

Similarly, if $B = \mathbb{I}$ or orthonormal, then computing \hat{v}^{k+1} reduces to computing the proximal operator of $h_{\mathcal{V}} := h + \delta_{\mathcal{V}}$, i.e.:

$$\hat{v}^{k+1} = \text{prox}_{\eta_k^{-1}h_{\mathcal{V}}} \left(B^T(c - A\hat{u}^{k+1}) - \eta_k^{-1}B^T\hat{\lambda}^k \right).$$

In addition to the SAM main step (4.3), our SAM algorithm also requires the following two steps:

$$(4.4) \quad \begin{cases} \hat{\lambda}_k := (1 - \tau_k)\bar{\lambda}^k + \tau_k\lambda_k^*, \\ [\bar{u}^{k+1}, \bar{v}^{k+1}] := (1 - \tau_k)[\bar{u}^k, \bar{v}^k] + \tau_k[\hat{u}^{k+1}, \hat{v}^{k+1}], \end{cases}$$

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

We first show that the point generated by (4.3) satisfies the following inequality, whose proof is postponed to Appendix A.2.1.

LEMMA 4.3. *Let $(\hat{u}^{k+1}, \hat{v}^{k+1}, \bar{\lambda}^{k+1})$ be the point generated by (4.3) and d_γ be defined by (3.6). Then, for any $\lambda \in \mathbb{R}^n$, we have*

$$(4.5) \quad d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \geq d_{\gamma_{k+1}}(\lambda) + \frac{1}{\eta_k} \langle \hat{\lambda}^k - \bar{\lambda}^{k+1}, \lambda - \hat{\lambda}^k \rangle + \frac{2\gamma_{k+1} - \eta_k}{2\gamma_{k+1}\eta_k} \|\bar{\lambda}^{k+1} - \hat{\lambda}^k\|^2.$$

Using Lemma 4.3, we can prove the following key result, whose proof can also be found in Appendix A.2.3.

LEMMA 4.4. *Let $\{\bar{w}^k\}_{k \geq 0}$ with $\bar{w}^k := (\bar{u}^k, \bar{v}^k, \bar{\lambda}^k)$ be the sequence generated by (4.3) and (4.4). If $\tau_k \in (0, 1)$ and $\gamma_k, \beta_k, \eta_k \in \mathbb{R}_{++}$ satisfy the following conditions:*

$$(4.6) \quad \eta_k = \gamma_{k+1} \geq \left(1 - \frac{\tau_k}{2}\right)\gamma_k, \quad \beta_{k+1} \geq (1 - \tau_k)\beta_k, \quad \text{and} \quad (1 - \tau_k^2)\gamma_{k+1}\beta_k \geq \tau_k^2,$$

then the following gap reduction condition holds:

$$(4.7) \quad G_{\gamma_{k+1}\beta_{k+1}}(\bar{w}^{k+1}) \leq (1 - \tau_k)G_{\gamma_k\beta_k}(\bar{w}^k) + \tau_k\eta_k D_{\mathcal{X}},$$

where $G_{\gamma_k\beta_k}$ is defined by (3.6) and $D_{\mathcal{X}}$ is defined by (3.2).

4.3. Updating the parameters. As indicated in Lemma 4.4, if the parameters $\tau_k, \gamma_k, \beta_k$, and η_k satisfy (4.6), then the gap reduction condition (4.7) holds. We outline below one way of updating these parameters.

LEMMA 4.5. *Given $\gamma_0 > 0$, the parameters $\tau_k, \gamma_k, \beta_k$, and η_k updated by:*

$$(4.8) \quad \tau_k = \frac{2}{k+3}, \quad \gamma_k = \frac{2\gamma_0}{k+2}, \quad \eta_k := \frac{2\gamma_0}{k+3}, \quad \text{and} \quad \beta_k = \frac{2(k+3)}{\gamma_0(k+1)(k+5)},$$

satisfy conditions (4.6). Moreover, the convergence rate of τ_k is optimal. In addition, $\beta_k < \frac{2}{\gamma_0(k+1)}$ and $\gamma_k\beta_k \leq \frac{4}{(k+1)^2}$.

Proof. The tightest update for γ_k and β_k is $\gamma_{k+1} := (1 - \tau_k/2)\gamma_k$ and $\beta_{k+1} := (1 - \tau_k)\beta_k$ due to (4.6). Using these updates in the last condition in (4.6) leads to:

$$\frac{(1 - \tau_{k+1}/2)(1 - \tau_{k+1})^2}{\tau_{k+1}^2} \geq \frac{1 + \tau_k}{\tau_k^2}.$$

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

Clearly, if we choose $\tau_k := \frac{2}{k+3}$, then $0 < \tau_k \leq \frac{2}{3} < 1$ for $k \geq 0$. Next, we choose $\gamma_{k+1} := (1 - \frac{\tau_k}{2})\gamma_k$. Substituting $\tau_k = \frac{2}{k+3}$ into this formula we have $\gamma_{k+1} = \left(\frac{k+2}{k+3}\right)\gamma_k$. By induction, we obtain $\gamma_k = \frac{2\gamma_0}{k+2}$. With $\tau_k = \frac{2}{k+3}$ and $\gamma_k = \frac{2\gamma_0}{k+2}$, we choose β_k from the last condition of (4.6) as:

$$\beta_k = \frac{\tau_k^2}{(1 - \tau_k^2)\gamma_{k+1}} = \frac{2(k+3)}{\gamma_0(k+1)(k+5)} < \frac{2}{\gamma_0(k+1)}.$$

We check the second condition $\beta_{k+1} \geq (1 - \tau_k)\beta_k$ in (4.6). Indeed, we have:

$$\beta_{k+1} = \frac{2(k+4)}{\gamma_0(k+2)(k+6)} \geq (1 - \tau_k)\beta_k = \left(1 - \frac{2}{k+3}\right) \frac{2(k+3)}{\gamma_0(k+1)(k+5)} \Leftrightarrow k+8 \geq 0.$$

Since $k+8 > 0$ for $k \geq 0$, the condition $\beta_{k+1} \geq (1 - \tau_k)\beta_k$ holds. From the update rule of β_k and γ_k , it is obvious to show that $\beta_k < \frac{2}{\gamma_0(k+1)}$ and $\gamma_k\beta_k \leq \frac{4}{(k+1)^2}$. \square

Uncertainty relation with the forward-backward splitting. Since $\tau_k \in (0, 1)$, $\gamma_{k+1} \geq (1 - \tau_k/2)\gamma_k$ and $\gamma_{k+1}\beta_k \geq \frac{\tau_k^2}{1 - \tau_k^2}$, the optimal choice of γ_k and β_k is summarized by the following uncertainty relation:

$$(4.9) \quad \gamma_k\beta_k = \frac{\tau_k^2}{(1 - \tau_k^2)(1 - 0.5\tau_k)} = \Omega(\tau_k^2).$$

As indicated in Lemma 4.5, the optimal convergence rate of $\{\tau_k\}$ is $\Omega(1/k)$. Consequently, by (4.9), the optimal convergence rate of $\{\gamma_k\beta_k\}$ is $\mathcal{O}(1/k^2)$. In addition, we can show that $\gamma_k = \mathcal{O}(\eta_k)$. Hence, by Lemma 3.3, we observe that γ_k and β_k trade off the convergence rate of $|f(\bar{x}^k) - f^*|$ and $\|A\bar{u}^k + B\bar{v}^k - c\|$, respectively, while the overall convergence rate of $\{\gamma_k\beta_k\}$ is $\mathcal{O}(1/k^2)$.

4.4. The new AMA-like algorithm. We now combine the initial point (4.1), the main step (4.3) and (4.4), and the update rule (4.8) to complete the SAM method as described in Algorithm 1.

Algorithm 1 (*Smooth Alternating Minimization (SAM) Algorithm*)

Initialization:

1. Fix $\bar{u}_c \in \mathcal{U}$ and choose $\gamma_0 > 0$. Set $\eta_0 := 2\gamma_0/3$.
 2. Compute \bar{u}^0, \bar{v}^0 and $\bar{\lambda}^0$ as in (4.1).
- for** $k := 0$ **to** k_{\max} **do**
3. Compute $\tau_k := \frac{2}{k+3}$, $\gamma_{k+1} := \frac{2\gamma_0}{k+3}$, and $\beta_k := \frac{2(k+3)}{\gamma_0(k+1)(k+5)}$.
 4. Compute $\lambda_k^* := \beta_k^{-1}(A\bar{u}^k + B\bar{v}^k - c)$.
 5. Update $\hat{\lambda}^k := (1 - \tau_k)\bar{\lambda}^k + \tau_k\lambda_k^*$.
 6. Compute the penalty parameter $\eta_k := \frac{2\gamma_0}{k+3}$.
 7. Update $(\hat{u}^{k+1}, \hat{v}^{k+1}, \bar{\lambda}^{k+1})$ as (4.3).
 8. 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**
-

Algorithm 1 requires two dual steps at Step 4 for λ_k^* and Step 7 for $\bar{\lambda}^{k+1}$. However, we can combine them in order to reduce the number of matrix-vector multiplications. More specifically, using Step 4, Step 7 and Step 8, we can derive:

$$(4.10) \quad \lambda_{k+1}^* := \beta_{k+1}^{-1}[(1 - \tau_k)\beta_k\lambda_k^* + \tau_k\eta_k^{-1}(\bar{\lambda}^{k+1} - \hat{\lambda}^k)].$$

Then, Algorithm 1 only requires one matrix-vector multiplication (Au, Bv) and one adjoint operation $(A^T\lambda, B^T\lambda)$ per iteration. Hence, the cost-per-iteration of (4.3) and the standard AMA (1.5) are essentially the same. We will discuss the stopping criterion of Algorithm 1 in the next section.

4.5. Convergence analysis. We prove the convergence and the worst-case analytical complexity of Algorithm in Theorem 4.6.

THEOREM 4.6. *Let $\{\bar{w}^k\}$ be the sequence generated by the SAM method (Algorithm 1). Then, for any $\gamma_0 > 0$, the following estimates hold:*

$$(4.11) \quad \begin{cases} -D_{\Lambda^*} \|A\bar{u}^k + B\bar{v}^k - c\| \leq & f(\bar{x}^k) - f^* & \leq \frac{2\gamma_0(D_{\mathcal{U}}^A + 2D_{\mathcal{X}})}{k+2}, \\ & \|A\bar{u}^k + B\bar{v}^k - c\| & \leq \frac{4D_{\Lambda^*}}{\gamma_0(k+1)} + \frac{2\sqrt{2(D_{\mathcal{U}}^A + 2D_{\mathcal{X}})}}{k+1}, \\ & d^* - d(\bar{\lambda}^k) & \leq \frac{4D_{\Lambda^*}^2}{\gamma_0(k+1)} + \frac{2D_{\Lambda^*}\sqrt{2(D_{\mathcal{U}}^A + 2D_{\mathcal{X}})}}{k+1}, \end{cases}$$

where D_{Λ^*} , $D_{\mathcal{U}}^A$ and $D_{\mathcal{X}}$ are defined by (3.2). As a consequence, if we choose $\gamma_0 := 1$, then the worst-case analytical complexity of Algorithm 1 to achieve an ε -primal solution \bar{x}^k of (1.3) in the sense of Definition 3.4 is $\mathcal{O}(\varepsilon^{-1})$.

Proof. First, we check the conditions of Lemma 4.1. From (4.8), we see that $\eta_0 = \frac{2}{3}\gamma_0$ and $\beta_0 = \frac{6}{5\gamma_0}$. Hence, $3\gamma_0 = \frac{9}{2}\eta_0 > \eta_0$, which is the first condition of Lemma 4.1. Moreover, $\frac{\gamma_0}{(3\gamma_0 - \eta_0)\eta_0} = \frac{9}{14\gamma_0} < \frac{6}{5\gamma_0} = \beta_0$, which is the second condition of Lemma 4.1. Hence $G_0(\bar{w}^0) \leq \eta_0 D_{\mathcal{X}} = (2/3)\gamma_0 D_{\mathcal{X}} < 2\gamma_0 D_{\mathcal{X}}$.

Next, we estimate the term $\tau_k \eta_k$ in (4.7) as follows:

$$\tau_k \eta_k = \frac{4\gamma_0}{(k+3)^2} < \frac{4\gamma_0}{k+3} \left(1 - \frac{k+1}{k+2}\right) = \frac{4\gamma_0}{k+3} - \left(1 - \frac{2}{k+3}\right) \frac{4\gamma_0}{k+2} = \frac{4\gamma_0}{k+3} - (1 - \tau_k) \frac{4\gamma_0}{k+2}.$$

Combing this estimate and (4.7), we get $G_{k+1}(\bar{w}^{k+1}) - \frac{4\gamma_0 D_{\mathcal{X}}}{k+3} \leq (1 - \tau_k) \left[G_k(\bar{w}^k) - \frac{4\gamma_0 D_{\mathcal{X}}}{k+2}\right]$. By induction, we have $G_k(\bar{w}^k) - \frac{4\gamma_0 D_{\mathcal{X}}}{k+2} \leq \omega_k [G_0(\bar{w}^0) - 2\gamma_0 D_{\mathcal{X}}] \leq 0$ whenever $G_0(\bar{w}^0) \leq 2\gamma_0 D_{\mathcal{X}}$, where $\omega_k := \prod_{i=0}^{k-1} (1 - \tau_i)$. Hence, we finally get:

$$(4.12) \quad G_k(\bar{w}^k) \leq \frac{4\gamma_0 D_{\mathcal{X}}}{k+2}.$$

We also note that $\gamma_k \beta_k = \frac{4(k+3)}{(k+1)(k+2)(k+5)} \leq \frac{4}{(k+1)^2}$. Using this estimate and (4.12) into Lemma 3.3, we obtain:

$$(4.13) \quad \begin{cases} -D_{\Lambda^*} \|A\bar{u}^k + B\bar{v}^k - c\| \leq & f(\bar{x}^k) - f^* & \leq \frac{2\gamma_0(D_{\mathcal{U}}^A + 2D_{\mathcal{X}})}{k+2}, \\ & \|A\bar{u}^k + B\bar{v}^k - c\| & \leq \frac{4D_{\Lambda^*}}{\gamma_0(k+1)} + \frac{2\sqrt{2(D_{\mathcal{U}}^A + 2D_{\mathcal{X}})}}{k+1}, \\ & d^* - d(\bar{\lambda}^k) & \leq \frac{4D_{\Lambda^*}^2}{\gamma_0(k+1)} + \frac{2D_{\Lambda^*}\sqrt{2(D_{\mathcal{U}}^A + 2D_{\mathcal{X}})}}{k+1}, \end{cases}$$

which is (4.11). Finally, if we choose $\gamma_0 := 1$ then, we obtain the worst-case complexity of Algorithm 1 is $\mathcal{O}(\varepsilon^{-1})$. \square

From Theorem 4.6, we see that if we choose $\gamma_0 := 1$, then (4.11) leads to:

$$\begin{cases} |f(\bar{x}^k) - f^*| & \leq \frac{2 \max \{ [D_{\mathcal{U}}^A + 2D_{\mathcal{X}}], [2D_{\Lambda^*}^2 + D_{\Lambda^*}\sqrt{2(D_{\mathcal{U}}^A + 2D_{\mathcal{X}})}] \}}{k+1}, \\ \|A\bar{u}^k + B\bar{v}^k - c\| & \leq \frac{4D_{\Lambda^*} + 2\sqrt{2D_{\mathcal{U}}^A + 4D_{\mathcal{X}}}}{k+1}. \end{cases}$$

This convergence rate is optimal under Assumption A.1 in the sense of first-order black-box models [24, 25].

4.6. Special cases. We now consider two special cases of the constrained problem (1.3): the full-column rank of A and the strong convexity of g .

4.6.1. Case 1: A is full column rank. If $\sigma_A^2 := \lambda_{\min}(A^T A) > 0$, where $\lambda_{\min}(A^T A)$ is the smallest eigenvalue of $A^T A$, then we can replace d_γ^1 in (3.1) by:

$$(4.14) \quad d_\gamma^1(\lambda) := \min_{u \in \mathcal{U}} \left\{ g(u) + \langle A^T \hat{\lambda}^k, u \rangle + \frac{\gamma}{2} \|u - \bar{u}_c\|^2 \right\} = \text{prox}_{\gamma^{-1}g_{\mathcal{U}}} \left(\bar{u}_c - \gamma^{-1} A^T \hat{\lambda}^k \right),$$

where $g_{\mathcal{U}}(\cdot) := g(\cdot) + \delta_{\mathcal{U}}(\cdot)$.

In this case, the function d_γ^1 is concave and smooth. Its gradient $\nabla d_\gamma^1(\cdot) = Au_\gamma^*(\cdot)$ is Lipschitz continuous with the Lipschitz constant $L_{d_\gamma^1} := \gamma^{-1} \|A\|^2$. Let $\bar{c} := 1 + \frac{\|A\|^2}{\sigma_A^2} = 1 + \text{cond}(A^T A) \geq 2$. We can modify the proof of Lemma 4.4 to obtain the gap reduction condition (4.7) under the following conditions:

$$(4.15) \quad \begin{aligned} \eta_k &= \frac{\gamma_{k+1}}{\|A\|^2}, & \gamma_{k+1} &\geq \left(1 - \frac{\tau_k}{\bar{c}}\right) \gamma_k, \\ \beta_{k+1} &\geq (1 - \tau_k) \beta_k, & \text{and} & \quad (1 - \tau_k^2) \gamma_{k+1} \beta_k \geq \|A\|^2 \tau_k^2, \end{aligned}$$

By analyzing directly these conditions as in Appendix A.3.1, we can obtain the following update rule:

$$(4.16) \quad \begin{aligned} \tau_k &:= \frac{\bar{c}}{k + \bar{c} + 1} \in (0, 1), & \eta_k &:= \frac{\bar{c} \gamma_0}{\|A\|^2 (k + \bar{c} + 1)}, \\ \gamma_k &:= \frac{\bar{c} \gamma_0}{k + \bar{c}}, & \text{and} & \quad \beta_k := \frac{\|A\|^2 \bar{c} (k + \bar{c} + 1)}{\gamma_0 (k + 1) (k + 2\bar{c} + 1)}. \end{aligned}$$

With this update, we see that $\eta_0 := \frac{\gamma_0 \bar{c}}{(\bar{c} + 1) \|A\|^2}$. Moreover, we substitute the initial point \bar{u}^0 in (4.1) by:

$$(4.17) \quad \bar{u}^0 := \text{argmin} \left\{ g(u) + \frac{\gamma_0}{2} \|u - \bar{u}_c\|^2 : u \in \mathcal{U} \right\} = \text{prox}_{\gamma_0^{-1}g_{\mathcal{U}}}(\bar{u}_c).$$

Using (4.17) and the update rule (4.16) in Algorithm 1, we obtain a new variant of Algorithm 1. The following corollary shows the convergence of this variant, whose proof is in Appendix A.3.1.

COROLLARY 4.7. *Let $\{\bar{w}^k\}$ be the sequence generated by Algorithm 1 using (4.17) and the update rule (4.16). Then, for any $\gamma_0 > 0$, the following estimates hold:*

$$(4.18) \quad \begin{cases} -D_{\Lambda^*} \|A\bar{u}^k + B\bar{v}^k - c\| \leq f(\bar{x}^k) - f^* \leq \frac{\bar{c} \gamma_0}{k + \bar{c}} \left[D_{\mathcal{U}} + \frac{D_{\mathcal{X}}}{\|A\|^2} \right], \\ \|A\bar{u}^k + B\bar{v}^k - c\| \leq \frac{2\|A\|^2 D_{\Lambda^*}}{\gamma_0 (k + 1)} + \frac{\bar{c} \sqrt{2(\|A\|^2 D_{\mathcal{U}} + D_{\mathcal{X}})}}{\sqrt{(k + 1)(k + \bar{c})}}, \end{cases}$$

where $\bar{c} := 1 + \frac{\|A\|^2}{\sigma_A^2} \geq 2$, D_{Λ^*} and $D_{\mathcal{X}}$ are defined by (3.2), and $D_{\mathcal{U}} := (1/2) \max \{ \|u - \bar{u}_c\|^2 : u \in \mathcal{U} \}$. As a consequence, if we choose $\gamma_0 := \frac{\|A\|}{\sigma_A}$, then the worst-case analytical complexity of Algorithm 1 to achieve an ε -primal solution \bar{x}^k of (1.3) in the sense of Definition 3.4 is $\mathcal{O}(\varepsilon^{-1})$.

4.6.2. Case 2: g is strongly convex. If g is strongly convex with the convexity parameter $\mu_g > 0$, then we can modify Algorithm 1 so that we obtain the convergence rate $\mathcal{O}(\frac{1}{k^2})$ in terms of the dual objective function d as shown in [17]. However, the convergence rate in terms of the primal objective residual $|f(\bar{x}^k) - f^*|$ and the primal feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$ we prove below unfortunately remains $\mathcal{O}(\frac{1}{k})$. This may be an artifact of our proof technique.

Let us consider again the dual function d_0^1 defined by (2.3). Since g is strongly convex with the strong convexity parameter $\mu_g > 0$, ∇d_0^1 is Lipschitz continuous with the Lipschitz constant $L_{d_0^1} := \frac{\|A\|^2}{\mu_g}$. 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 4.4, we can show in Appendix A.3.2 that if the following conditions hold:

$$(4.19) \quad \beta_{k+1} \geq (1 - \tau_k)\beta_k \quad \text{and} \quad \eta_k \left(1 + \frac{\tau_k}{2} - \frac{\|A\|^2 \eta_k}{2\mu_g}\right) \geq \frac{\tau_k^2}{2(1 - \tau_k)\beta_k},$$

then:

$$(4.20) \quad G_{\beta_{k+1}}(\bar{w}^{k+1}) \leq (1 - \tau_k)G_{\beta_k}(\bar{w}^k) + \tau_k \eta_k D_{\mathcal{X}},$$

where $G_{\beta_k}(\bar{w}^k) := f_{\beta_k}(\bar{x}^k) - d(\bar{\lambda}^k)$. The conditions (4.19) lead to the following update rule (see the analysis in Appendix A.3.2):

$$(4.21) \quad \tau_k := \frac{2}{k+3} \in (0, 1), \quad \eta_k := \frac{\mu_g \tau_k}{\|A\|^2}, \quad \text{and} \quad \beta_k := \frac{\|A\|^2 \tau_k}{2\mu_g(1 - \tau_k)} = \frac{\|A\|^2}{\mu_g(k+1)}.$$

The starting point \bar{u}^0 can be computed as:

$$(4.22) \quad \bar{u}^0 := \operatorname{argmin} \{g(u) : u \in \mathcal{U}\}.$$

Using (4.22) and the update rule (4.21) 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 Appendix A.3.2.

COROLLARY 4.8. *Let $\{\bar{w}^k\}$ be the sequence generated by Algorithm 1 using (4.22) and the update rule (4.21). Then, for any $\gamma_0 > 0$, the following estimates hold:*

$$(4.23) \quad \begin{cases} -D_{\Lambda^*} \|A\bar{u}^k + B\bar{v}^k - c\| \leq f(\bar{x}^k) - f^* & \leq \frac{4\mu_g D_{\mathcal{X}}}{\|A\|^2(k+2)}, \\ \|A\bar{u}^k + B\bar{v}^k - c\| & \leq \frac{2\|A\|^2 D_{\Lambda^*}}{\mu_g(k+1)} + \frac{2\sqrt{D_{\mathcal{X}}}}{\sqrt{(k+1)(k+2)}}, \end{cases}$$

where D_{Λ^*} and $D_{\mathcal{X}}$ are defined by (3.2). As a consequence, the worst-case analytical complexity of Algorithm 1 to achieve an ε -primal solution \bar{x}^k of (1.3) in the sense of Definition 3.4 is $\mathcal{O}(\varepsilon^{-1})$.

We emphasize that the authors in [17] proved the $\mathcal{O}(\frac{1}{k^2})$ -convergence rate in terms of the dual objective residual $d^* - d(\lambda^k)$ for strongly convex g . However, in Corollary 4.8, we show the $\mathcal{O}(\frac{1}{k})$ -convergence rate in terms of the primal objective residual $|f(\bar{x}^k) - f^*|$ and the primal feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$, which may not be optimal (in the ergodic sense) in this special case.

5. Douglas-Rachford splitting with the smoothed gap. We now present a new alternating direction method of multipliers (ADMM) algorithm for solving (1.3) by applying Douglas-Rachford splitting to the smoothed dual. Our new algorithm, named the smooth ADMM (S-ADMM), has an optimal convergence rate on the primal objective residual $|f(\bar{x}^k) - f^*|$ and the primal feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$ and sets the penalty parameters in a heuristic free fashion.

5.1. The main step of Smooth ADMM. The main step of our S-ADMM algorithm is as follows. Given $\hat{\lambda}^k \in \mathbb{R}^n$, $\hat{v}^k \in \mathcal{U}$ 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:

$$(5.1) \quad \begin{cases} \hat{u}^{k+1} := \operatorname{argmin}_{u \in \mathcal{U}} \left\{ g_{\gamma_{k+1}}(u) + \langle A^T \hat{\lambda}^k, u \rangle + \frac{\rho_k}{2} \|Au + B\hat{v}^k - c\|^2 \right\}, \\ \hat{v}^{k+1} := \operatorname{argmin}_{v \in \mathcal{V}} \left\{ h(v) + \langle B^T \hat{\lambda}^k, v \rangle + \frac{\eta_k}{2} \|A\hat{u}^{k+1} + Bv - c\|^2 \right\}, \\ \bar{\lambda}^{k+1} := \hat{\lambda}^k + \eta_k (A\hat{u}^{k+1} + B\hat{v}^{k+1} - c), \end{cases}$$

where g_{γ} is defined by (3.1). This scheme is different from the standard ADMM scheme (1.6) at two points. First, \hat{u}^{k+1} is computed from g_{γ} instead of g . Second, we use different penalty parameters ρ_k and η_k compared to (1.6).

In addition to the main step (5.1), our S-ADMM algorithm also requires additional steps as follows:

$$(5.2) \quad \begin{cases} \hat{\lambda}_k := (1 - \tau_k)\bar{\lambda}^k + \tau_k\lambda_k^*, \\ [\hat{u}^{k+1}, \hat{v}^{k+1}] := (1 - \tau_k)[\bar{u}^k, \bar{v}^k] + \tau_k[\hat{u}^{k+1}, \hat{v}^{k+1}], \end{cases}$$

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

The following inequality is a key to analyze the convergence of our S-ADMM scheme 5.1, whose proof can be found in Appendix A.4.1.

LEMMA 5.1. *Let d_γ be defined by (3.6) and $(\hat{\lambda}^k, \bar{\lambda}^{k+1})$ be generated by (5.1). Then, d_γ satisfies the following inequality:*

$$(5.3) \quad d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \geq d_{\gamma_{k+1}}(\lambda) + \frac{1}{\eta_k} \langle \hat{\lambda}^k - \bar{\lambda}^{k+1}, \lambda - \hat{\lambda}^k \rangle + \frac{1}{\eta_k} \|\hat{\lambda}^k - \bar{\lambda}^{k+1}\|^2 - \frac{1}{2\gamma_{k+1}} \|\bar{\lambda}^k - \bar{\lambda}^{k+1}\|^2,$$

for any $\lambda \in \mathbb{R}^n$, where $\tilde{\lambda}^k := \hat{\lambda}^k + \rho_k(A\hat{u}^{k+1} + B\hat{v}^k - c)$.

Next, we prove the following lemma in Appendix A.4.2, which provides conditions on the parameters to guarantee the gap reduction property.

LEMMA 5.2. *Let $\{\bar{w}^k\}_{k \geq 0}$ with $\bar{w}^k := (\bar{u}^k, \bar{v}^k, \bar{\lambda}^k)$ be the sequence generated by (5.1) and (5.2). If $\tau_k \in (0, 1)$ and $\gamma_k, \beta_k, \rho_k, \eta_k \in \mathbb{R}_{++}$ satisfy the following conditions:*

$$(5.4) \quad \eta_k \beta_k \geq \frac{\tau_k^2}{1 - \tau_k^2}, \quad (3 - \tau_k)\gamma_{k+1} \geq (3 - 2\tau_k)\gamma_k, \quad \beta_{k+1} \geq (1 - \tau_k)\beta_k, \quad \text{and} \quad \gamma_{k+1} \geq \eta_k + \frac{\rho_k}{\tau_k},$$

then the following gap reduction condition holds:

$$(5.5) \quad G_{\gamma_{k+1}\beta_{k+1}}(\bar{w}^{k+1}) \leq (1 - \tau_k)G_{\gamma_k\beta_k}(\bar{w}^k) + \tau_k(\rho_k + \eta_k)D_{\mathcal{X}},$$

where $G_{\gamma_k\beta_k}$ defined by (3.6) and $D_{\mathcal{X}}$ is defined by (3.2).

5.2. Updating parameters. The second step of our algorithmic design is to derive the update rule for the parameters.

LEMMA 5.3. *Given $\gamma_0 > 0$, the parameters $\tau_k, \gamma_k, \beta_k, \rho_k$ and η_k updated by:*

$$(5.6) \quad \tau_k := \frac{3}{k+4}, \quad \gamma_k := \frac{2\gamma_0}{k+2}, \quad \beta_k := \frac{9(k+3)}{\gamma_0(k+1)(k+7)}, \quad \rho_k := \frac{3\gamma_0}{(k+3)(k+4)}, \quad \eta_k := \frac{\gamma_0}{k+3},$$

satisfy the conditions (5.4). Moreover, the choice of τ_k is optimal. In addition, $\beta_k \leq \frac{9}{\gamma_0(k+2)}$ and $\gamma_k\beta_k \leq \frac{18}{(k+2)^2}$.

Proof. Similarly to the proof of Lemma 4.5, we can show that the optimal rate of $\{\tau_k\}$ is $\mathcal{O}(1/k)$. From the conditions (5.4), 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 (5.4). 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 $\gamma_k = \frac{2\gamma_0}{k+2}$. Now, we choose $\eta_k := \frac{\gamma_{k+1}}{2} = \frac{\gamma_0}{k+3}$. Then, from the last condition of (5.4), we choose $\rho_k := \frac{\tau_k\gamma_{k+1}}{2} = \frac{3\gamma_0}{(k+3)(k+4)}$.

From the third condition of (5.4), we can derive:

$$\beta_k = \frac{\tau_k^2}{(1 - \tau_k^2)\eta_k} = \frac{2\tau_k^2}{(1 - \tau_k^2)\gamma_{k+1}} = \frac{9(k+3)}{\gamma_0(k+1)(k+7)} < \frac{9}{\gamma_0(k+2)}.$$

We need to check the second condition $\beta_{k+1} \geq (1 - \tau_k)\beta_k$ in (5.4). Indeed, we have:

$$\beta_{k+1} = \frac{9(k+4)}{\gamma_0(k+2)(k+8)} \geq (1 - \tau_k)\beta_k = \left(1 - \frac{3}{k+4}\right) \frac{9(k+3)}{\gamma_0(k+1)(k+7)} \Leftrightarrow 2k^2 + 26k + 64 \geq 0.$$

Since $2k^2 + 26k + 64 > 0$ for $k \geq 0$. Hence, the second condition of (5.4) holds. The two last estimates of β_k and $\gamma_k\beta_k$ are trivial due to the update rule of β_k and γ_k . \square

We note that we have freedom to choose γ_0 in order to trade off the primal objective residual $|f(\bar{x}^k) - f^*|$ and the primal feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$ as in Algorithm 1.

Uncertainty relation with Douglas-Rachford splitting. We note that Lemma 5.3 only provides *one* possibility to update the parameters. From the conditions (5.4) of Lemma 5.2, we can see that the tightest rules for updating the parameters satisfy:

$$\eta_k = \mathcal{O}(\gamma_k), \quad \rho_k = \mathcal{O}(\gamma_k \tau_k), \quad \text{and} \quad \beta_k \gamma_k = \Omega(\tau_k^2).$$

However, since the optimal rate of τ_k is $\Omega(1/k)$ due to Lemma 5.2, the optimal rate of the product $\beta_k \gamma_k$ is at most $\mathcal{O}(1/k^2)$. Hence, by Lemma 3.3, the rate on separated parameters β_k and γ_k trades off the convergence rate of $|f(\bar{x}^k) - f^*|$ and $\|A\bar{u}^k + B\bar{v}^k - c\|$, respectively, while the overall rate of both quantities does not exceed $\mathcal{O}(1/k^2)$.

5.3. The ADMM-like algorithm. We also use the point $\bar{w}^0 = (\bar{u}^0, \bar{v}^0, \bar{\lambda}^0)$ computed by (4.1) as an initial point. By putting (4.1), (5.6), (5.1) and (5.2) together, we obtain the complete S-ADMM algorithm as presented in Algorithm 2.

Algorithm 2 (*Smooth Alternating Direction Method of Multipliers (S-ADMM)*)

Initialization:

1. Fix $\bar{u}_c \in \mathcal{U}$ and choose $\gamma_0 > 0$. Set $\eta_0 := \gamma_0/3$.
2. Compute u^0 and v^0 as in (4.1). Set $\hat{v}^0 := \bar{v}^0$.

for $k = 0$ **to** k_{\max} **do**

3. Compute $\tau_k := \frac{3}{k+4}$, $\gamma_{k+1} := \frac{2\gamma_0}{k+3}$, and $\beta_k := \frac{9(k+3)}{\gamma_0(k+1)(k+7)}$.
4. Compute $\lambda_k^* := \beta_k^{-1}(A\bar{u}^k + B\bar{v}^k - c)$.
5. Update $\hat{\lambda}^k := (1 - \tau_k)\bar{\lambda}^k + \tau_k \lambda_k^*$.
6. Compute the penalty parameters $\rho_k := \frac{3\gamma_0}{(k+3)(k+4)}$ and $\eta_k := \frac{\gamma_0}{k+3}$.
7. Update $(\hat{u}^{k+1}, \hat{v}^{k+1}, \bar{\lambda}^{k+1})$ as (5.1).
8. 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

Similarly to Algorithm 1, we can also combine two dual steps at Step 4 and Step 7 of Algorithm 2 by using (4.10). In this case, the cost-per-iteration of Algorithm 2 is essentially the same as in the standard ADMM scheme (1.6). We will discuss the stopping criterion of Algorithm 2 in the next section.

5.4. Convergence analysis. The following theorem shows the convergence and the worst-case analytical complexity of Algorithm 2.

THEOREM 5.4. *Let $\{(\bar{u}^k, \bar{v}^k, \bar{\lambda}^k)\}_{k \geq 0}$ be the sequence generated by Algorithm 2. Then the following estimates hold:*

$$(5.7) \quad \left\{ \begin{array}{l} -D_{\Lambda^*} \|A\bar{u}^k + B\bar{v}^k - c\| \leq f(\bar{x}^k) - f^* \leq \frac{2\gamma_0 D_{\mathcal{U}}^A}{k+2} + \frac{3\gamma_0 D_{\mathcal{X}}}{2(k+3)} \left(1 + \frac{6}{k+2}\right), \\ \|A\bar{u}^k + B\bar{v}^k - c\| \leq \frac{18D_{\Lambda^*}}{\gamma_0(k+2)} + \frac{6}{k+2} \sqrt{D_{\mathcal{U}}^A + \frac{3(k+8)}{2(k+3)} D_{\mathcal{X}}}, \\ d^* - d(\bar{\lambda}^k) \leq \frac{18D_{\Lambda^*}^2}{\gamma_0(k+2)} + \frac{6D_{\Lambda^*}}{k+2} \sqrt{D_{\mathcal{U}}^A + \frac{3(k+8)}{2(k+3)} D_{\mathcal{X}}}, \end{array} \right.$$

where D_{Λ^*} , $D_{\mathcal{U}}^A$ and $D_{\mathcal{X}}$ are defined by (3.2). If $\gamma_0 := 3$, then the worst-case analytical complexity of Algorithm 2 to achieve an ε -primal solution \bar{x}^k of (1.3) in the sense of Definition 3.4 is $\mathcal{O}(\varepsilon^{-1})$.

Proof. First, we check the conditions of Lemma 4.1. From (5.6), we have $\eta_0 = \frac{\gamma_0}{3}$ and $\beta_0 = \frac{27}{\gamma_0}$. Hence, $3\gamma_0 - \eta_0 = 8\eta_0 > 0$, which satisfies the first condition of Lemma 4.1. Now, $\frac{\gamma_0}{(3\gamma_0 - \eta_0)\eta_0} = \frac{9}{8\gamma_0} < \frac{27}{\gamma_0} = \beta_0$. Hence, the second condition of Lemma 4.1 holds.

Next, since $\tau_k = \frac{3}{k+4}$, $\rho_k = \frac{3\gamma_0}{(k+3)(k+4)}$ and $\eta_k = \frac{\gamma_0}{k+3}$, we can derive:

$$\begin{aligned}\tau_k(\eta_k + \rho_k) &= \frac{3\gamma_0}{(k+3)(k+4)} + \frac{9\gamma_0}{(k+3)(k+4)^2} \\ &\leq \frac{3\gamma_0}{2(k+4)} \left(1 + \frac{6}{k+3}\right) - \left(1 - \frac{3}{k+4}\right) \frac{3\gamma_0}{2(k+3)} \left(1 + \frac{6}{k+2}\right).\end{aligned}$$

Substituting this inequality into (5.5) and rearrange the result we obtain:

$$\left[G_{k+1}(\bar{w}^{k+1}) - \frac{3\gamma_0 D_{\mathcal{X}}}{2(k+4)} \left(1 + \frac{6}{k+3}\right)\right] \leq (1 - \tau_k) \left[G_k(\bar{w}^k) - \frac{3\gamma_0 D_{\mathcal{X}}}{2(k+3)} \left(1 + \frac{6}{k+2}\right)\right].$$

By induction, we obtain $G_k(\bar{w}^k) - \frac{3\gamma_0 D_{\mathcal{X}}}{2(k+3)} \left(1 + \frac{6}{k+2}\right) \leq \omega_k \left[G_0(\bar{w}^0) - 2\gamma_0 D_{\mathcal{X}}\right] \leq 0$ as long as $G_0(\bar{w}^0) \leq 2\gamma_0 D_{\mathcal{X}}$. Now using Lemma 4.1, we have $G_0(\bar{w}^0) \leq \eta_0 D_{\mathcal{X}} \leq \frac{\gamma_0}{3} D_{\mathcal{X}} \leq 2\gamma_0 D_{\mathcal{X}}$. Hence, $G_k(\bar{w}^k) \leq \frac{3\gamma_0 D_{\mathcal{X}}}{2(k+3)} \left(1 + \frac{6}{k+2}\right)$.

Finally, by using Lemma 3.3 with $\beta_k := \frac{9(k+3)}{\gamma_0(k+1)(k+7)}$ and $\gamma_k \beta_k \leq \frac{18}{(k+2)^2}$, we obtain the bounds in (5.7). If we choose $\gamma_0 := 3$ then, we obtain the worst-case analytical complexity of Algorithm 2 is $\mathcal{O}(\varepsilon^{-1})$. \square

If we choose $\gamma_0 := 3$, then (5.7) can be simplified as:

$$\begin{cases} |f(\bar{x}^k) - f^*| &\leq \frac{6 \max\{[D_{\mathcal{U}}^A + 3D_{\mathcal{X}}], [D_{\Lambda^*}^2 + D_{\Lambda^*} \sqrt{D_{\mathcal{U}}^A + 4D_{\mathcal{X}}}] \}}{k+2}, \\ \|A\bar{u}^k + B\bar{v}^k - c\| &\leq \frac{6[D_{\Lambda^*} + \sqrt{D_{\mathcal{U}}^A + 4D_{\mathcal{X}}}] }{k+2}. \end{cases}$$

As can be seen from Theorem 5.4, the term $\frac{6}{k+2} \sqrt{D_{\mathcal{U}}^A + \frac{3(k+8)}{2(k+3)} D_{\mathcal{X}}}$ in (5.7) does not depend on the choice of γ_0 . If we decrease γ_0 , the the upper bound of $|f(\bar{x}^k) - f^*|$ is decreasing, while the upper bound of $\|A\bar{u}^k + B\bar{v}^k - c\|$ is increasing, and vice versa. Hence, γ_0 can be chosen such that it trades off these upper bounds.

5.5. Special cases.

5.5.1. Case 1: A has full column rank. Similarly to Algorithm 1, we consider the case A is full-column rank, i.e., $\underline{\sigma}_A^2 := \lambda_{\min}(A^T A) > 0$. Then, we can use d_{γ}^1 defined by (4.14) instead of the one in (3.1). In this case, we can modify Algorithm 2 to take into account this structure. The details of this variant is very similar to the AMA variant in Subsection 4.6.1, which we omit the details here.

5.5.2. Case 2: g is strongly convex . If g is μ_g -strongly convex, then instead of using d_{γ}^1 defined by (3.1), we can use d_0^1 defined by (2.3) in Algorithm 2. With the same argument as in Subsection 4.6.2, we can derive a new variant of Algorithm 2 whose Step 7 is exactly the standard ADMM scheme (1.6). In this case, we can also maintain the $\mathcal{O}(\frac{1}{k})$ -convergence rate both on $|f(\bar{x}^k) - f^*|$ and on $\|A\bar{u}^k + B\bar{v}^k - c\|$.

6. Implementation remarks and extensions. This section provides some remarks on the implementation of Algorithms 1 and 2, and their variants.

6.1. Stopping criterion. Practically, we can not run Algorithms 1 and 2 to achieve the worst-case bounds as indicated in Theorems 4.6 and 5.4. We often terminate them at a given desired accuracy level $\varepsilon > 0$. Clearly, the feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$ can be computed at each iteration without additional cost. However, to evaluate the dual objective value $d(\bar{\lambda}^k)$, one requires to solve two convex subproblems, which can substantially increase the cost-per-iteration. In implementation, we can measure the change of the primal objective values in the s successive iterations, i.e., $|f(\bar{x}^{k+j}) - f(\bar{x}^k)|$ for $j = 1, \dots, s$. More precisely, if the relative objective change does not improve in the s successive iterations, i.e., $|f(\bar{x}^{k+j}) - f(\bar{x}^k)| / \max\{1, |f(\bar{x}^k)|\} \leq \varepsilon$ for $j = 1, \dots, s$, and the relative feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\| / \max\{1, \|c\|, \|A\bar{u}^k\|, \|B\bar{v}^k\|\} \leq \varepsilon$, then we can terminate the algorithms at suboptimal point \bar{x}^k .

6.2. Preconditioned AMA and ADMM. When $g_{\mathcal{U}}$ and $h_{\mathcal{V}}$ possess a “tractable” proximity operator $\text{prox}_{g_{\mathcal{U}}}$ and $\text{prox}_{h_{\mathcal{V}}}$, respectively, one can linearize the quadratic terms in (4.3) (or (5.1)) in order to obtain a preconditioned AMA (or a preconditioned ADMM) variant that has a closed form solution for \hat{u}^{k+1} and \hat{v}^{k+1} . While the preconditioned ADMM can be found in [9], we briefly present the preconditioned AMA variant here. When g is not strongly convex, we linearize the term $(\gamma_{k+1}/2)\|A(u - \bar{u}_c)\|^2$ in (3.1) and the quadratic term in the computation of \hat{v}^{k+1} to obtain:

$$(6.1) \quad \begin{cases} \hat{u}^{k+1} := \underset{u \in \mathcal{U}}{\text{argmin}} \left\{ g(u) + \langle A^T \hat{\lambda}^k, u \rangle + \frac{\gamma_{k+1} L_A}{2} \|u - \tilde{u}^k\|^2 \right\} \\ \quad = \text{prox}_{(\gamma_{k+1} L_A)^{-1} g_{\mathcal{U}}} \left(\tilde{u}^k - (\gamma_{k+1} L_A)^{-1} A^T \hat{\lambda}^k \right), \\ \hat{v}^{k+1} := \underset{v \in \mathcal{V}}{\text{argmin}} \left\{ h(v) + \langle B^T \hat{\lambda}^k, v \rangle + \frac{\eta_k L_B}{2} \|v - \tilde{v}^k\|^2 \right\} \\ \quad = \text{prox}_{(\eta_k L_B)^{-1} h_{\mathcal{V}}} \left(\tilde{v}^k - (\eta_k L_B)^{-1} B^T \hat{\lambda}^k \right), \end{cases}$$

where $L_A := \|A\|^2$ and $L_B := \|B\|^2$ are the Lipschitz constants of the quadratic terms, $\tilde{u}^k := \hat{u}^k - \frac{1}{L_A} A^T A(\hat{u}^k - \bar{u}_c)$ and $\tilde{v}^k := \hat{v}^k - \frac{1}{L_B} B^T (A\hat{u}^{k+1} + B\hat{v}^k - c)$. Using these two steps in Algorithm 1, we obtain a preconditioned AMA variant of this algorithm. However, the convergence rate guarantee for this variant is still not known yet. Since we linearize the quadratic terms, one can use exact line-search to determine the local Lipschitz constants instead of the global ones L_A and L_B , see, e.g., [37].

6.3. Trading off the primal objective residual and the feasibility gap. We have seen that γ_k and β_k trade off the primal objective residual $|f(\bar{x}^k) - f^*|$ and the feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$. However, in many applications (e.g., in control), we often attempt pushing the feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$ so that it quickly decreases, while we pay less attention to the objective function. We can fix γ_k at certain value or even slightly increase it. In Algorithm 2, the corresponding penalty parameters ρ_k and η_k are also fixed or are increasing. Hence, we can update β_k and τ_k based on the conditions (5.4) rather than the explicit rule (5.6). This heuristic strategy often performs better than the theoretical algorithms. However, we do not have convergence rate guarantee for this variant. The details of this strategy can be found in [32, 33].

7. Numerical experiments. We first verify the $\mathcal{O}(1/k)$ -convergence rate of Algorithms 1 and 2. Then, we provide two numerical examples in imaging sciences for Algorithm 2.

7.1. Empirical convergence rate vs. theoretical bounds. We first illustrate the $\mathcal{O}(1/k)$ convergence rate of Algorithms 1 and 2. Our test is based on the following square-root LASSO problem:

$$(7.1) \quad \min_{v \in \mathcal{V}} \{ \|B(v) - c\|_2 + \kappa \|v\|_1 \},$$

where \mathcal{V} is the domain of the signals which is assumed to be bounded, c is the observed measurement vector, B is a linear operator from $\mathbb{R}^{p_2} \rightarrow \mathbb{R}^n$, and $\kappa > 0$ is a regularization parameter. As indicated in [4], under given assumptions on the data (B, c) , one can determine the optimal parameter κ such that we obtain the exact recovery. By introducing $u := B(v) - c$, we have $\mathcal{U} := \{u : u = B(v) - c, v \in \mathcal{V}\}$, which is also bounded in \mathbb{R}^n . Moreover, we can transform (7.1) into (1.3), i.e.:

$$(7.2) \quad f^* := \min_{u \in \mathcal{U}, v \in \mathcal{V}} \{ f(x) := \|u\|_2 + \kappa \|v\|_1 : B(v) - u = c \}.$$

When applying Algorithms 1 and 2 to solve (7.2), the first convex subproblem in (4.3) or (5.1) can be solved in a closed form, while we need to solve the second convex subproblem in these schemes with FISTA [2]. By using a restart and warm-start strategy, we can solve

the last subproblem within a few inner iterations to get a high accuracy approximation for \hat{v}^{k+1} .

In order to verify our theoretical guarantee, we solve (7.2) by CVX [18] with the best precision to obtain high accuracy solution of (7.2). We use this result as a baseline for our comparison. We compute the theoretical bounds given in Theorems 4.6 and 5.4 using the CVX solution. We test both algorithms on synthetic data generated randomly using standard iid Gaussian distribution: $c := B(v^{\natural}) + 10^{-3}\epsilon$, where ϵ is an iid Gaussian noise vector, and v^{\natural} is a given s -sparse vector. The domain \mathcal{V} of the signal is a box generated by the lower and upper bounds of v^{\natural} .

Figures 7.1 and 7.2 show the empirical performance vs. the theoretical bounds of Algorithms 1 and 2, respectively on the test instance of size $p_2 = 1000$, $n = 350$ and $s = 100$ up to 10^4 iterations.

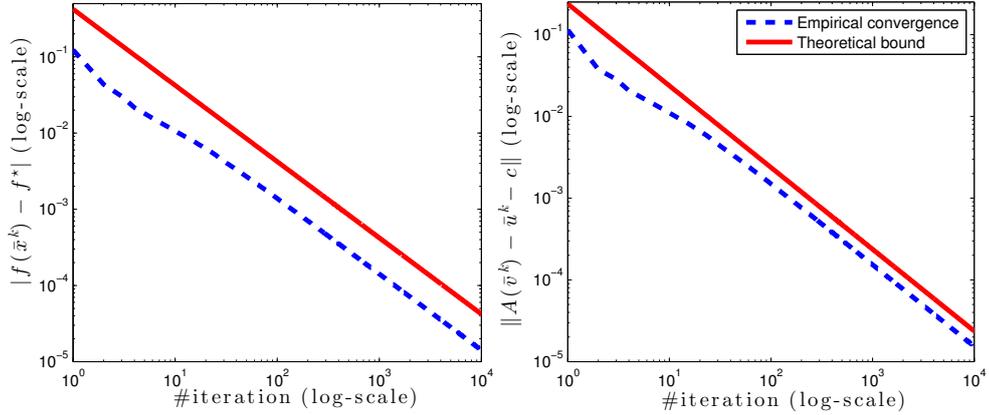


FIG. 7.1. Empirical convergence vs. Theoretical bounds in Algorithm 1.

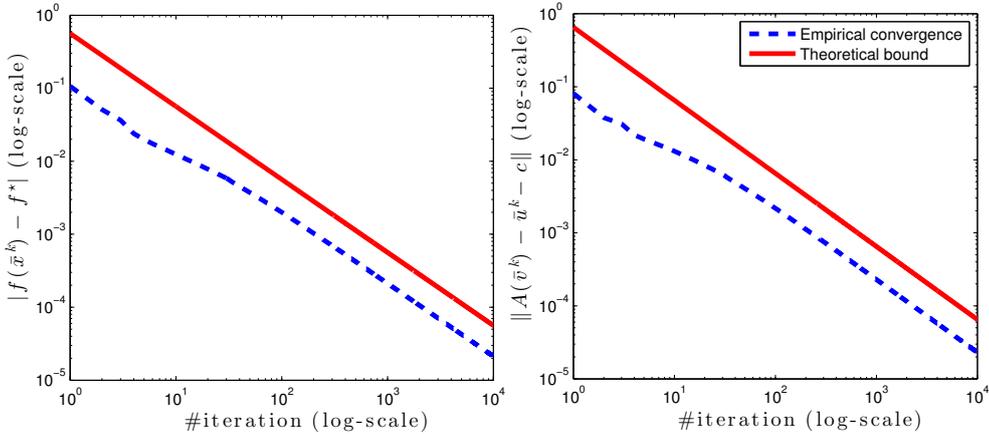


FIG. 7.2. Empirical convergence vs. Theoretical bounds in Algorithm 2.

We observe from Figures 7.1 and 7.2 that the convergence behavior of both algorithms nearly follows the $\mathcal{O}(1/k)$ -theoretical bounds up to a given constant in the absolute values. This behavior is often observed in optimal first-order primal-dual methods.

7.2. Image processing application. We demonstrate the performance robustness of Algorithm 2 by applying it to the following image deconvolution problem:

$$(7.3) \quad F^* := \min_{0 \leq v \leq 255} (1/2) \{ F(v) := \|\mathcal{B}(v) - c\|_2^2 + \kappa \|v\|_{\text{TV}} \},$$

where c is a given blurry image with a known blur kernel \mathcal{B} , and $\|\cdot\|_{\text{TV}}$ is the isotropic total variation norm and $\kappa > 0$ is a regularization parameter.

As opposed to directly using the TV-norm proximal operator, we simply use the linear mapping \mathbf{D} of its ℓ_1 -norm $\|v\|_{\text{TV}} = \|\mathbf{D}v\|_1$ and introduce a slack variable $u = \mathbf{D}v$ to split (7.3) into v and u variables with additional linear coupling constraint $u - \mathbf{D}v = 0$. Hence, we can reformulate (7.3) into (1.3), where $u \in \mathcal{U} := \{\hat{u} : \hat{u} = \mathbf{D}v, 0 \leq v \leq 255\}$ is also bounded.

We apply the enhanced variants of Algorithm 2 to solve the resulting problem and compare it with the ADMM solver implemented in [10] since both algorithms have similar complexity per iteration. In the first enhanced variant, we fix ρ_k and update accordingly other parameters based the conditions (5.4). In the Enhanced-Algorithm 2, we also increase ρ by $\rho_{k+1} := 1.5\rho_k$ at each iteration. We choose the initial regularization parameters ρ_0 the same as the recent *exact* ADMM solver suggested in [10].

Fortunately, if we assume periodic boundary conditions for the TV-norm, then ADMM can efficiently obtain accurate solutions to the subproblems in computing \hat{u}^{k+1} and \hat{v}^{k+1} in (5.1). The key idea is that the operator $\mathbf{D}^T\mathbf{D} + \mathcal{B}^T\mathcal{B}$ is diagonalizable by the Fourier transform. Hence, the complexity-per-iteration in the exact ADMM scheme (1.6) and Algorithm 2 is approximately the same.

We now illustrates the performance of the two variants of Algorithm 2 and the ADMM code [10] for test images. Our test is based on the `camera_man` (256×256) and the `epfl-art` (612×816) images, with the regularization $\kappa := 55$, which we find the best one. The suggested value for ρ_0 is $\rho_0 := 2$ in [10].

Figure 7.3 shows the convergence of three algorithms after 100 iterations for two images. We can see that ADMM quickly decreases the objectives at early iterations and is saturated at a certain value, while Enhanced-Algorithm 2 continues to descend on the objective function.

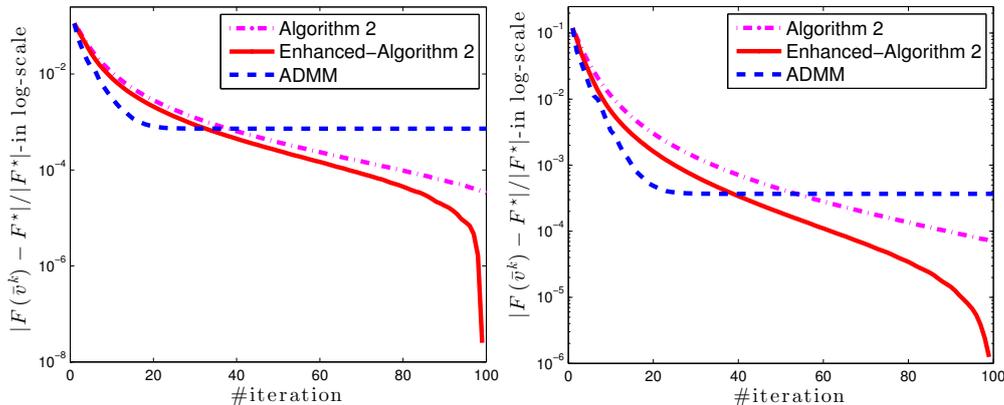


FIG. 7.3. The convergence behavior of three algorithms: Left: *cameraman*, Right: *epfl-art*.

We note that, in this test, the ADMM solver is sensitive to the choice of ρ . For several different values of ρ , if we run the ADMM solver further (up to 120 iterations), then it starts oscillating and diverging as shown in Figure 7.4, although the algorithm must converge theoretically. Our algorithms are relatively numerically robust since the parameters are updated accordingly to the conditions (5.4).

Figure 7.5 reveals the original `epfl-art` image, the noisy image with $\mathcal{N}(0, 0.01)$ -Gaussian noise, and two recovered images of two algorithms: Enhanced-Algorithm 2 and the ADMM solver [10]. We observe that both algorithms produce similar results (in terms of PSNR) while Enhanced-Algorithm 2 gives lower objective value $F(\bar{v}^k) = 2'103'983.92$ compared to $F(v^k) = 2'104'629.31$ of the ADMM solver.

7.3. Poisson noise image reconstruction. In this test, we study the empirical impact of inexact proximal operator calculations to the performance of Algorithm 2. Again, we choose the enhancement variant, which has better performance ability. For

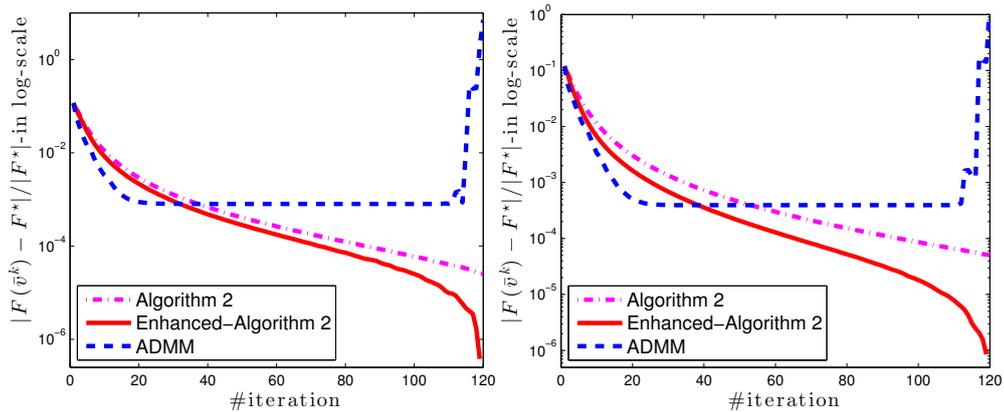


FIG. 7.4. The divergence of the ADMM solver: Left: *cameraman*, Right: *epfl-art*.

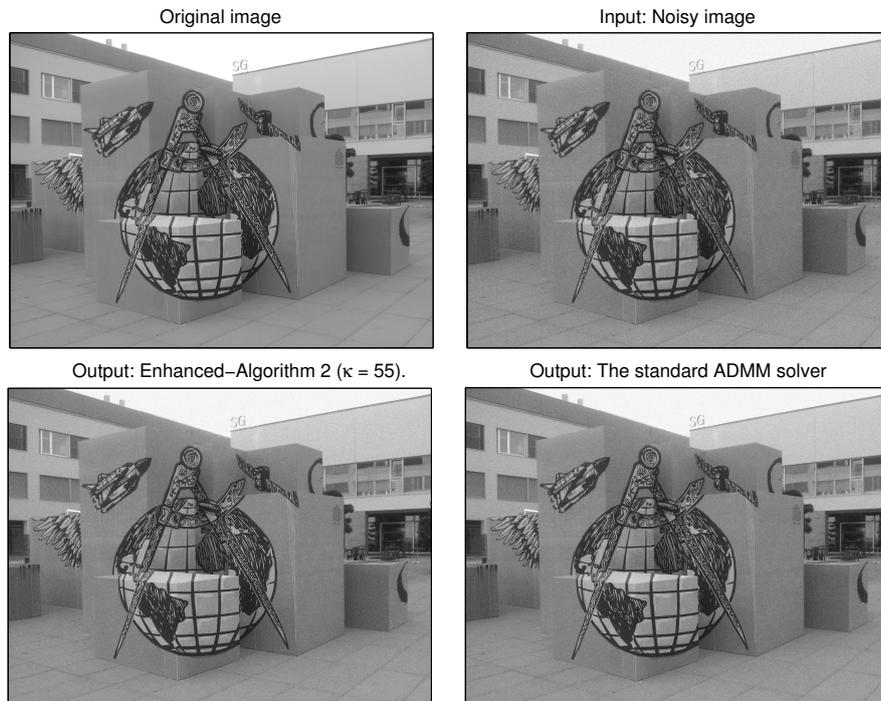


FIG. 7.5. The result of the two algorithms: $PSNR = 28.74$ for *Enhanced-Algorithm 2* and $PSNR = 27.69$ for *ADMM*.

this, we use a Schatten norm based regularizer on a Poisson log-likelihood data model:

$$(7.4) \quad F^* := \min_{v \in \mathcal{V}} \left\{ F(v) := (\mathcal{B}(v))^T \mathbf{1} - \sum_{i=1}^m \mathbf{y}_i \log((\mathcal{B}(v))_i + b) + \kappa \|v\|_S \right\},$$

where $\mathcal{V} := [0, 255]^{p^2}$, \mathbf{y} is a given photon count vector in \mathbb{Z}^m , b is the background intensity, $\kappa > 0$ is a chosen regularization parameter, and \mathcal{B} is a blur kernel. This log-likelihood model is quite common in imaging sciences (see [20] and the references quoted therein).

The work in [20] proposed a norm based on exploiting self-similarities within the images via $\|v\|_S := \|\text{mat}(\mathcal{H}(v))\|_*$, which is the Schatten-norm of a matrix $\text{mat}(\mathcal{H}(v))$ for a suitably chosen linear operator \mathcal{H} . Since the proximal operator regarding the second term $h(v) := \kappa \|v\|_S + \delta_{\mathcal{V}}(v)$, where $\delta_{\mathcal{V}}$ is the indicator of \mathcal{V} , does not have a closed form.

The resulting inexact computation affects the performance of optimization algorithms. Here, we compare our enhanced variant of Algorithm 2 (called *Enhanced-*

Algorithm 2) with PADMM and PADMM based on our tuning strategy as well as the exact ADMM solver provided by [20]. The ADMM solver exploits boundary conditions and Fourier transform to invert $\mathbb{I} + \mathcal{B}^T \mathcal{B}$ for solving its subproblems. When b is zero (i.e., there is no background), then the logarithmic term pose computational problems since its gradient is no longer Lipschitz continuous. Fortunately, the proximal operator of the log function can be efficiently calculated. We test these algorithms on the Clown im-

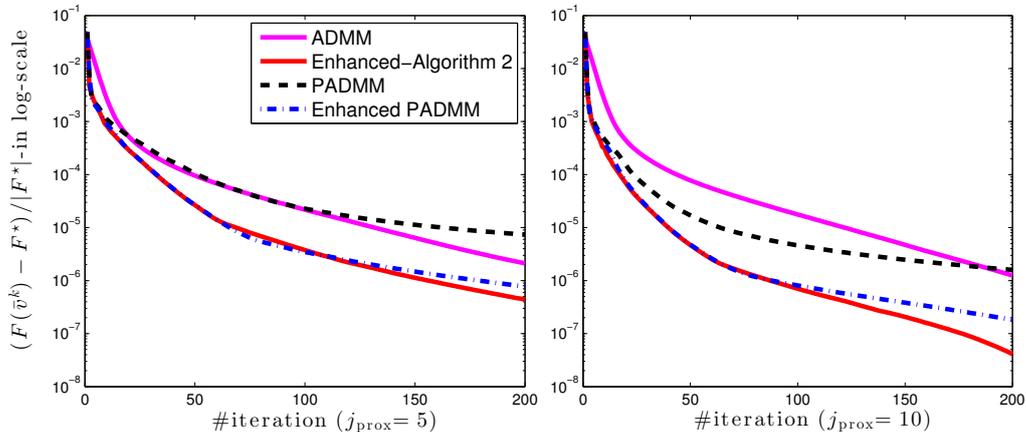


FIG. 7.6. The performance of four algorithms on the Clown image [20].

age and the confocal microscope image [20], where we take the regularization parameter $\kappa = 0.055$ suggested in [20]. We use the `Denoise` solver in [20] to approximately compute the prox-operator of h with inner iterations $j_{\text{prox}} = 5, 10$, where we can warm start each iteration using the current estimate. The exact ADMM solver is already implemented with penalty parameter updates.

Figure 7.6 illustrates that our Enhanced-Algorithm 2 solver and PADMM are quite robust to the inexact proximal operator calculations and outperform exact ADMM for a range of j_{prox} iterations. Against intuition, we observe that PADMM exhibits numerical instability when j_{prox} is getting high. Overall, our algorithm provides the best time to reach an ε -solution since doubling j_{prox} roughly doubles the overall time. For instance, $j_{\text{prox}} = 5$ and 200 iterations roughly takes the same time as $j_{\text{prox}} = 10$ and 100 iterations, where our algorithm provides the best accuracy.

If we choose the confocal microscope image [20], then Enhanced-Algorithm 2 outperforms the rest as can be seen in Figure 7.7 for three choices of j_{prox} : 5, 10 and 50.

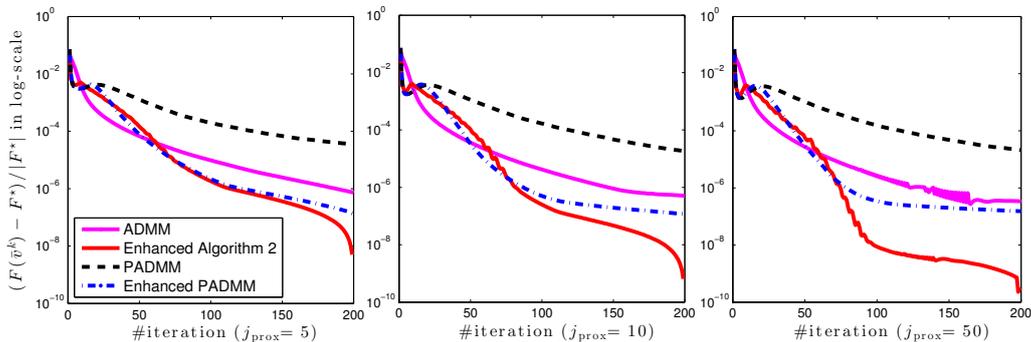


FIG. 7.7. The performance of four algorithms on the confocal microscope image [20].

8. Conclusion. 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 [32], and unifies three main ideas: smoothing, alternating direction, and gap reduction. By splitting the gap, we have

developed two new smooth alternating optimization algorithms: SAM and S-ADMM with rigorous convergence guarantees. We have shown that our algorithms have optimal convergence rate in the sense of first-order black-box models [24, 25]. 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. We have also considered special cases of our SAM and S-ADMM algorithms, discussed some heuristic enhancements, and provided three numerical examples to verify the theoretical and practical aspects.

Acknowledgments. This work was supported in part by the European Commission under Grant MIRG-268398, ERC Future Proof and SNF 200021-132548, SNF 200021-146750 and SNF CRSII2-147633.

Appendix. The proof of technical results. This appendix provides the full proof of technical results presented in the main text.

A.1. Auxiliary results. We require the following technical lemmas. The first one is standard in convex optimization, which we omit the proof here.

LEMMA A.1. *Let ψ be a proper, closed and convex function from \mathbb{R}^p to $\mathbb{R} \cup \{+\infty\}$, \mathcal{U} be a nonempty, closed and convex set in \mathbb{R}^p , $\bar{u}_c \in \mathcal{U}$, and $A \in \mathbb{R}^{n \times p}$. For $\mu > 0$, let:*

$$\psi_\mu^* := \min_{u \in \mathcal{U}} \{ \psi_\mu(u) := \psi(u) + (\mu/2) \|A(u - \bar{u}_c)\|^2 \},$$

and u_μ^* be the solution of this minimization problem. Then, we have:

$$(A.1) \quad \psi_\mu^* + (\mu/2) \|A(u - u_\mu^*)\|^2 \leq \psi_\mu(u), \quad \forall u \in \mathcal{U}.$$

Next, we recall the following important results from [25, Theorem 2.1.5].

THEOREM A.2. *Assume that ψ is a smooth concave function with the L_ψ -Lipschitz gradient $\nabla\psi$ on \mathbb{R}^p , i.e., $\|\nabla\psi(u) - \nabla\psi(\hat{u})\| \leq L_\psi \|u - \hat{u}\|$ for any $u, \hat{u} \in \mathbb{R}^p$. Then:*

a) *For any $u, \hat{u} \in \mathbb{R}^p$, we have:*

$$(A.2) \quad 0 \leq \psi(u) + \langle \nabla\psi(u), \hat{u} - u \rangle - \psi(\hat{u}) \leq \frac{L_\psi}{2} \|\hat{u} - u\|^2.$$

b) *For any $u, \hat{u} \in \mathbb{R}^p$ and $\tau \in [0, 1]$, we have:*

$$(A.3) \quad \psi((1 - \tau)u + \tau\hat{u}) \geq (1 - \tau)\psi(u) + \tau\psi(\hat{u}) + \frac{\tau(1 - \tau)}{2L_\psi} \|\nabla\psi(u) - \nabla\psi(\hat{u})\|^2.$$

A.2. Convergence analysis of Algorithm 1. We provide the full proof of Lemma 4.3, Lemma 4.1 and Lemma 4.4.

A.2.1. The proof of Lemma 4.3: The AMA quadratic surrogate of d_γ . Let us denote by $\bar{g}_\gamma := g_\gamma + \delta_{\mathcal{U}}$, and $\bar{h} := h + \delta_{\mathcal{V}}$. Then by the definition (3.1), we can write:

$$d_\gamma^1(\lambda) := -\bar{g}_\gamma^*(-A^T \lambda) \quad \text{and} \quad d_0^2(\lambda) := -\bar{h}^*(-B^T \lambda),$$

where \bar{g}_γ^* and \bar{h}^* are the Fenchel conjugate of \bar{g}_γ and \bar{h} , respectively.

Now, we show that $\bar{\lambda}^{k+1}$ computed by (4.3) is by applying the proximal-gradient method (ISTA) [2] to the dual problem (2.1). Indeed, we can write the optimality condition for the two subproblems in (4.3) and using the third line of (4.3) to get:

$$(A.4) \quad \begin{cases} 0 \in \partial \bar{g}_{\gamma_{k+1}}(\hat{u}^{k+1}) + A^T \hat{\lambda}^k \\ 0 \in \partial \bar{h}(\hat{v}^{k+1}) + B^T \hat{\lambda}^k + \eta_k B^T (A \hat{u}^{k+1} + B \hat{v}^{k+1} - c) \equiv \partial \bar{h}(\hat{v}^{k+1}) + B^T \bar{\lambda}^{k+1}. \end{cases}$$

The first line of (A.4) can be rewritten as $-A^T \hat{\lambda}^k \in \partial \bar{g}_{\gamma_{k+1}}(\hat{u}^{k+1})$, which is equivalent to $\hat{u}^{k+1} \in \partial \bar{g}_{\gamma_{k+1}}^*(-A^T \hat{\lambda}^k)$. Multiplying this inclusion by A and noting that $d_{\gamma_{k+1}}^1(\hat{\lambda}^k) = -\bar{g}_{\gamma_{k+1}}^*(-A^T \hat{\lambda}^k)$, we have:

$$(A.5) \quad A \hat{u}^{k+1} \in A \partial \bar{g}_{\gamma_{k+1}}^*(-A^T \hat{\lambda}^k) = \nabla d_{\gamma_{k+1}}^1(\hat{\lambda}^k),$$

due to the differentiability of $d_{\gamma_{k+1}}^1$. Similarly, from the second line of (A.4) we have:

$$(A.6) \quad B\hat{v}^{k+1} \in B\partial\bar{h}^*(-B^T\bar{\lambda}^{k+1}) = \partial d_0^2(\bar{\lambda}^{k+1}).$$

Summing up (A.5) and (A.6), and using the third line of (4.3), we obtain $\eta_k^{-1}(\bar{\lambda}^{k+1} - \hat{\lambda}^k) = A\hat{u}^{k+1} + B\hat{v}^{k+1} - c \in \nabla d_{\gamma_{k+1}}^1(\hat{\lambda}^k) + \partial d_0^2(\bar{\lambda}^{k+1}) - c$. This expression is equivalent to:

$$(A.7) \quad \hat{\lambda}^k + \eta_k(\nabla d_{\gamma_{k+1}}^1(\hat{\lambda}^k) - c) \in \bar{\lambda}^{k+1} - \eta_k \partial d_0^2(\bar{\lambda}^{k+1}).$$

Since d_0^2 is concave, using the prox notion of $-d_0^2$, we can write (A.7) in the following proximal-gradient scheme:

$$(A.8) \quad \bar{\lambda}^{k+1} = \text{prox}_{-\eta_k d_0^2} \left(\hat{\lambda}^k + \eta_k(\nabla d_{\gamma_{k+1}}^1(\hat{\lambda}^k) - c) \right).$$

It remains applying Lemma 2.3. in [2] to obtain (4.5). \square

A.2.2. The proof of Lemma 4.1: Computing initial points. It is obvious that (4.1) has the same form as (4.3) with $k = 0$ and $\hat{\lambda}^k = \mathbf{0}^n$. By using Lemma 4.3 with $k = 0$, $\hat{\lambda}^k = \mathbf{0}^n$ and $\lambda = \mathbf{0}^n$, we obtain:

$$(A.9) \quad d_{\gamma_0}(\bar{\lambda}^0) \geq d_{\gamma_0}(\mathbf{0}^n) + \frac{2\gamma_0 - \eta_0}{2\gamma_0\eta_0} \|\bar{\lambda}^0\|^2.$$

Since \bar{v}^0 is the solution of the second problem in (4.1) and $v^*(\mathbf{0}^n) \in \mathcal{V}$, we have $h(v^*(\mathbf{0}^n)) + \frac{\eta_0}{2} \|A\bar{u}^0 + Bv^*(\mathbf{0}^n) - c\|^2 \geq h(\bar{v}^0) + \frac{\eta_0}{2} \|A\bar{u}^0 + B\bar{v}^0 - c\|^2$. With $D_{\mathcal{X}}$ defined by (3.2), this inequality implies:

$$(A.10) \quad h(v^*(\mathbf{0}^n)) + \eta_0 D_{\mathcal{X}} \geq h(\bar{v}^0) + \frac{\eta_0}{2} \|A\bar{u}^0 + B\bar{v}^0 - c\|^2.$$

Using the definition of d_{γ_0} , we further estimate (A.9) using (A.10) as follows:

$$\begin{aligned} d_{\gamma_0}(\bar{\lambda}^0) &\geq d_{\gamma_0}^1(\mathbf{0}^n) + d_0^2(\mathbf{0}^n) \stackrel{(3.1)}{=} g(\bar{u}^0) + \frac{\gamma_0}{2} \|A(\bar{u}^0 - \bar{u}_c)\|^2 + h(v^*(\mathbf{0}^n)) \\ &\stackrel{(A.10)}{\geq} g(\bar{u}^0) + h(\bar{v}^0) + \frac{\gamma_0}{2} \|A(\bar{u}^0 - \bar{u}_c)\|^2 + \frac{\eta_0}{2} \|A\bar{u}^0 + B\bar{v}^0 - c\|^2 + \frac{2\gamma_0 - \eta_0}{2\gamma_0\eta_0} \|\bar{\lambda}^0\|^2 - \eta_0 D_{\mathcal{X}} \\ &= f_{\beta_0}(\bar{x}^0) - \frac{1}{2} \left[\frac{1}{\beta_0} - \frac{(3\gamma_0 - \eta_0)\eta_0}{\gamma_0} \right] \|A\bar{u}^0 + B\bar{v}^0 - c\|^2 + \frac{\gamma_0}{2} \|A(\bar{u}^0 - \bar{u}_c)\|^2 - \eta_0 D_{\mathcal{X}}. \end{aligned}$$

Since $G_{\gamma_0\beta_0}(\bar{w}^0) = f_{\beta_0}(\bar{x}^0) - d_{\gamma_0}(\bar{\lambda}^0)$, we obtain (4.2) from the last inequality. If $\beta_0 \geq \frac{\gamma_0}{\eta_0(3\gamma_0 - \eta_0)}$, then (4.2) leads to $G_{\gamma_0\beta_0}(\bar{w}^0) \leq \eta_0 D_{\mathcal{X}}$. \square

A.2.3. The proof of Lemma 4.4: Maintaining the gap reduction condition.

For notational simplicity, let us abbreviate $Mx - c := Au + Bv - c$ for any u and v . The proof of this lemma is divided into several steps.

Step 1: Key bound on $d_{\gamma_{k+1}}^1(\bar{\lambda}^{k+1})$. We note that $d_{\gamma_{k+1}}^1$ is concave and its gradient ∇d_{γ}^1 is Lipschitz continuous with $L_{d_{\gamma_{k+1}}^1} := \gamma_{k+1}^{-1}$ due to Lemma 3.2, for any $\tau_k \in [0, 1]$, $\bar{\lambda}^k, \hat{\lambda}^k \in \mathbb{R}^m$ and $\lambda^k := (1 - \tau_k)\bar{\lambda}^k + \tau_k\hat{\lambda}^k$, if we define:

$$(A.11) \quad \tilde{r}_k := \frac{\tau_k(1 - \tau_k)}{2} \|\nabla d_{\gamma_{k+1}}^1(\bar{\lambda}^k) - \nabla d_{\gamma_{k+1}}^1(\hat{\lambda}^k)\|^2,$$

then, by (A.3) in Theorem A.2, we have:

$$(A.12) \quad d_{\gamma_{k+1}}^1(\lambda^k) \geq (1 - \tau_k)d_{\gamma_{k+1}}^1(\bar{\lambda}^k) + \tau_k d_{\gamma_{k+1}}^1(\hat{\lambda}^k) + \gamma_{k+1}\tilde{r}_k.$$

Moreover, since $d_0^2(\cdot)$ is concave, we also have $d_0^2(\lambda^k) \geq (1 - \tau_k)d_0^2(\bar{\lambda}^k) + \tau_k d_0^2(\hat{\lambda}^k)$. Summing up this inequality and (A.12) and then using the definition of $d_\gamma(\cdot) = d_\gamma^1(\cdot) + d_0^2(\cdot) - \langle c, \cdot \rangle$ we obtain:

$$(A.13) \quad d_{\gamma_{k+1}}(\lambda^k) \geq (1 - \tau_k)d_{\gamma_{k+1}}(\bar{\lambda}^k) + \tau_k d_{\gamma_{k+1}}(\hat{\lambda}^k) + \gamma_{k+1}\tilde{r}_k.$$

Now, using Lemma 4.3 and (4.3) with $\lambda = \lambda^k$ and $\bar{\lambda}^{k+1} - \hat{\lambda}^k = \eta_k(A\hat{u}^{k+1} + B\hat{v}^{k+1} - c) \equiv \eta_k(M\hat{x}^{k+1} - c)$, we can derive:

$$(A.14) \quad d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \geq d_{\gamma_{k+1}}(\lambda^k) + \langle M\hat{x}^{k+1} - c, \hat{\lambda}^k - \lambda^k \rangle + \eta_k \left(1 - \frac{\eta_k}{2\gamma_{k+1}}\right) \|M\hat{x}^{k+1} - c\|^2.$$

Substituting (A.13) into (A.14), and using $\lambda^k = (1 - \tau_k)\bar{\lambda}^k + \tau_k\hat{\lambda}^k$ and $\hat{\lambda}^k = (1 - \tau_k)\bar{\lambda}^k + \tau_k\lambda_k^*$ with $\lambda_k^* := \frac{1}{\beta_k}(A\bar{u}^k + B\bar{v}^k - c) \equiv \frac{1}{\beta_k}(M\bar{x}^k - c)$ we get:

$$(A.15) \quad \begin{aligned} d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) &\stackrel{(A.13)}{\geq} (1 - \tau_k)d_{\gamma_{k+1}}(\bar{\lambda}^k) + \tau_k d_{\gamma_{k+1}}(\hat{\lambda}^k) + \tau_k \langle M\hat{x}^{k+1} - c, \lambda_k^* - \hat{\lambda}^k \rangle \\ &+ \eta_k \left(1 - \frac{\eta_k}{2\gamma_{k+1}}\right) \|M\hat{x}^{k+1} - c\|^2 + \gamma_{k+1}\tilde{r}_k. \end{aligned}$$

Step 2: Bound on $(1 - \tau_k)G_k(\bar{w}^k)$. By (3.6), we have $f_{\beta_k}(\bar{x}^k) = f(\bar{x}^k) + \frac{1}{2\beta_k}\|M\bar{x}^k - c\|^2$ and $G_k(\bar{w}^k) = f_{\beta_k}(\bar{x}^k) - d_{\gamma_k}(\bar{\lambda}^k)$. Using these expressions we have:

$$(A.16) \quad (1 - \tau_k)d_{\gamma_k}(\bar{\lambda}^k) = (1 - \tau_k)f(\bar{x}^k) + \frac{(1 - \tau_k)}{2\beta_k}\|M\bar{x}^k - c\|^2 - (1 - \tau_k)G_k(\bar{w}^k).$$

Next, let $\bar{u}_{k+1}^* := u_{\gamma_{k+1}}^*(\bar{\lambda}^k)$ be the solution of the first subproblem (4.3). Then, by (3.4), for $\gamma_{k+1}, \gamma_k > 0$, we have:

$$(A.17) \quad d_{\gamma_{k+1}}^1(\bar{\lambda}^k) \geq d_{\gamma_k}^1(\bar{\lambda}^k) - \frac{1}{2}(\gamma_k - \gamma_{k+1})\|A(\bar{u}_{k+1}^* - \bar{u}_c)\|^2 := d_{\gamma_k}^1(\bar{\lambda}^k) - \bar{r}_k,$$

where $\bar{r}_k := \frac{1}{2}(\gamma_k - \gamma_{k+1})\|A(\bar{u}_{k+1}^* - \bar{u}_c)\|^2 \geq 0$. Combing (A.17) and (A.16), we get:

$$(1 - \tau_k)d_{\gamma_{k+1}}(\bar{\lambda}^k) \geq (1 - \tau_k)f(\bar{x}^k) + \frac{(1 - \tau_k)}{2\beta_k}\|M\bar{x}^k - c\|^2 - (1 - \tau_k)G_k(\bar{w}^k) - (1 - \tau_k)\bar{r}_k.$$

Now, substituting this inequality into (A.15) and exchanging $(1 - \tau_k)G_k(\bar{w}^k)$ and $d_{\gamma_{k+1}}(\bar{\lambda}^{k+1})$, we further obtain:

$$(A.18) \quad \begin{aligned} (1 - \tau_k)G_k(\bar{w}^k) &\geq -d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + (1 - \tau_k)f(\bar{x}^k) + \frac{(1 - \tau_k)}{2\beta_k}\|M\bar{x}^k - c\|^2 + \tau_k d_{\gamma_{k+1}}(\hat{\lambda}^k) \\ &+ \tau_k \langle M\hat{x}^{k+1} - c, \lambda_k^* - \hat{\lambda}^k \rangle + \eta_k \left(1 - \frac{\eta_k}{2\gamma_{k+1}}\right) \|M\hat{x}^{k+1} - c\|^2 + \gamma_{k+1}\tilde{r}_k - (1 - \tau_k)\bar{r}_k. \end{aligned}$$

Step 3: Bound on $d_{\gamma_{k+1}}(\hat{\lambda}^k)$. With d_0^2 defined by (2.3), we denote by $\hat{v}_k^* := v^*(\hat{\lambda}^k)$ the solution of the second problem in (2.3). Now, we consider the following functions:

$$(A.19) \quad \begin{aligned} \tilde{d}_{k+1}^2(\hat{\lambda}^k) &:= \min_{v \in \mathcal{V}} \{h(v) + \langle B^T \hat{\lambda}^k, v \rangle + \frac{\eta}{2}\|A\hat{u}^{k+1} + Bv - c\|^2\}, \\ \tilde{d}_{k+1}^1(\hat{\lambda}^k) &:= d_\gamma^1(\hat{\lambda}^k) + \tilde{d}_{k+1}^2(\hat{\lambda}^k) - \langle c, \hat{\lambda}^k \rangle. \end{aligned}$$

We first estimate the bounds between $d_0^2(\cdot)$ and $\tilde{d}_{k+1}^2(\hat{\lambda}^k)$ as follows:

$$(A.20) \quad \begin{aligned} \tilde{d}_{k+1}^2(\hat{\lambda}^k) &\leq h(\hat{v}_k^*) + \langle B^T \hat{\lambda}^k, \hat{v}_k^* \rangle + \frac{\eta_k}{2}\|A\hat{u}^{k+1} + B\hat{v}_k^* - c\|^2 - \frac{\eta_k}{2}\|B(\hat{v}_k^* - \hat{v}^{k+1})\|^2 \\ &= d_0^2(\hat{\lambda}^k) + \frac{\eta_k}{2}\|A\hat{u}^{k+1} + B\hat{v}_k^* - c\|^2 - \frac{\eta_k}{2}\|B(\hat{v}_k^* - \hat{v}^{k+1})\|^2. \end{aligned}$$

By the definition of $\tilde{d}_{k+1}(\hat{\lambda}^k)$ in (A.19), it follows from (A.20), the definition of d_γ in (3.1), $\gamma = \gamma_{k+1}$, $\eta = \eta_k$, $\lambda = \hat{\lambda}^k$, and $u_c = \hat{u}^{k+1}$ that:

$$d_{\gamma_{k+1}}(\hat{\lambda}^k) \geq \tilde{d}_{k+1}(\hat{\lambda}^k) - \frac{\eta_k}{2} \|A\hat{u}^{k+1} + B\hat{v}_k^* - c\|^2 + \frac{\eta_k}{2} \|B(\hat{v}_k^* - \hat{v}^{k+1})\|^2.$$

Let $r_k := \frac{1}{2} \|A\hat{u}^{k+1} + B\hat{v}_k^* - c\|^2$, by the definition of \tilde{d}_{k+1} , the last inequality leads to:

$$d_{\gamma_{k+1}}(\hat{\lambda}^k) \geq f(\hat{x}^{k+1}) + \langle \hat{\lambda}^k, M\hat{x}^{k+1} - c \rangle + \frac{\eta_k}{2} \|M\hat{x}^{k+1} - c\|^2 + \frac{\gamma_{k+1}}{2} \|A(\hat{u}^{k+1} - \bar{u}_c)\|^2 - \eta_k r_k.$$

Using this inequality into (A.18) with $\hat{r}_k := \frac{1}{2} \|A(\hat{u}^{k+1} - \bar{u}_c)\|^2$, we can further estimate (A.18) as:

$$\begin{aligned} (1 - \tau_k)G_k(\bar{w}^k) &\geq -d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + (1 - \tau_k)f(\bar{x}^k) + \frac{(1 - \tau_k)}{2\beta_k} \|M\bar{x}^k - c\|^2 \\ &\quad + \tau_k \tilde{d}_{k+1}(\hat{\lambda}^k) + \tau_k \langle M\hat{x}^{k+1} - c, \lambda_k^* - \hat{\lambda}^k \rangle + \eta_k \left(1 - \frac{\eta_k}{2\gamma_{k+1}}\right) \|M\hat{x}^{k+1} - c\|^2 \\ &\quad - \tau_k \eta_k r_k + \gamma_{k+1} \tilde{r}_k - (1 - \tau_k) \bar{r}_k \\ &\geq -d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + (1 - \tau_k)f(\bar{x}^k) + \tau_k f(\hat{x}^{k+1}) + \frac{(1 - \tau_k)}{2\beta_k} \|M\bar{x}^k - c\|^2 \\ &\quad + \tau_k \langle M\hat{x}^{k+1} - c, \lambda_k^* \rangle + \eta_k \left(1 + \frac{\tau_k}{2} - \frac{\eta_k}{2\gamma_{k+1}}\right) \|M\hat{x}^{k+1} - c\|^2 \\ (A.21) \quad &\quad + \gamma_{k+1} \tilde{r}_k - (1 - \tau_k) \bar{r}_k + \tau_k \gamma_{k+1} \hat{r}_k - \tau_k \eta_k r_k. \end{aligned}$$

Step 4: Conditions on the parameters. Now, from (A.21), we assume that the parameters τ_k , γ_k , β_k and η_k are chosen such that:

$$(A.22) \quad \beta_{k+1} \geq (1 - \tau_k)\beta_k \quad \text{and} \quad \eta_k \left(1 + \frac{\tau_k}{2} - \frac{\eta_k}{2\gamma_{k+1}}\right) \geq \frac{\tau_k^2}{2(1 - \tau_k)\beta_k}.$$

These conditions are the two last conditions in (4.6).

Step 5: Refining the bound on $(1 - \tau_k)G_k(\bar{w}^k)$. For \tilde{r}_k , \bar{r}_k and \hat{r}_k defined by (A.11), (A.13) and (A.21), respectively, we define:

$$(A.23) \quad R_k := \gamma_{k+1} \tilde{r}_k - (1 - \tau_k) \bar{r}_k + \tau_k \gamma_{k+1} \hat{r}_k.$$

Since $\bar{x}^k = [\bar{u}^k, \bar{v}^k]$ and $\hat{x}^{k+1} = [\hat{u}^{k+1}, \hat{v}^{k+1}]$, by (4.4), we have $\bar{x}^{k+1} := (1 - \tau_k)\bar{x}^k + \tau_k \hat{x}^{k+1}$. Using the convexity of f we get:

$$f(\bar{x}^{k+1}) = f((1 - \tau_k)\bar{x}^k + \tau_k \hat{x}^{k+1}) \leq (1 - \tau_k)f(\bar{x}^k) + \tau_k f(\hat{x}^{k+1}).$$

Using this relation, $\lambda_k^* := \frac{1}{\beta_k}(M\bar{x}^k - c)$, and the conditions (A.22), we can further estimate (A.21) as:

$$\begin{aligned} (1 - \tau_k)G_k(\bar{w}^k) &\geq f(\bar{x}^{k+1}) - d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + \frac{1}{2(1 - \tau_k)\beta_k} \left[(1 - \tau_k)^2 \|M\bar{x}^k - c\|^2 \right. \\ &\quad \left. + \tau_k^2 \|M\hat{x}^{k+1} - c\|^2 + 2\tau_k(1 - \tau_k) \langle M\hat{x}^{k+1} - c, M\bar{x}^k - c \rangle \right] + R_k - \tau_k \eta_k r_k \\ &= f(\bar{x}^{k+1}) - d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + \frac{1}{2\beta_k(1 - \tau_k)} \|M((1 - \tau_k)\bar{x}^k + \tau_k \hat{x}^{k+1}) - c\|^2 + R_k - \tau_k \eta_k r_k \\ (A.22) \quad &\geq f(\bar{x}^{k+1}) + \frac{1}{2\beta_{k+1}} \|M\bar{x}^{k+1} - c\|^2 - d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + R_k - \tau_k \eta_k r_k \\ &= f_{\beta_{k+1}}(\bar{x}^{k+1}) - d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + R_k - \tau_k \eta_k r_k \\ (A.24) \quad &\stackrel{(3.6)}{=} G_{k+1}(\bar{w}^{k+1}) + R_k - \tau_k \eta_k r_k, \end{aligned}$$

Step 6: Estimating R_k . Next step, we need to further estimate R_k . Let $\bar{a}_k := A(u_{\gamma_{k+1}}^*(\bar{\lambda}^k) - \bar{u}_c)$, $\hat{a}_k := A(u_{\gamma_{k+1}}^*(\hat{\lambda}^k) - \bar{u}_c)$. By the definition (3.1) of d_γ^1 , we have $\nabla d_\gamma^1(\lambda) := Au_\gamma^*(\lambda)$. Therefore, using the definition of \tilde{r}_k , \bar{r}_k and \hat{r}_k , we can write R_k explicitly as:

$$\begin{aligned} 2\gamma_{k+1}^{-1}R_k &= \tau_k(1 - \tau_k)\|\bar{a}_k - \hat{a}_k\|^2 - (1 - \tau_k)(\gamma_{k+1}^{-1}\gamma_k - 1)\|\bar{a}_k\|^2 + \tau_k\|\hat{a}_k\|^2 \\ &= \tau_k(2 - \tau_k)\|\hat{a}_k\|^2 - 2(1 - \tau_k)\tau_k\langle \hat{a}_k, \bar{a}_k \rangle + (1 - \tau_k)(\tau_k - \gamma_{k+1}^{-1}\gamma_k + 1)\|\bar{a}_k\|^2. \end{aligned}$$

Hence, if $(2 - \tau_k)(1 + \tau_k - \gamma_{k+1}^{-1}\gamma_k) \geq \tau_k(1 - \tau_k)$, then $R_k \geq 0$. The last condition leads to $\gamma_{k+1} \geq (1 - \frac{\tau_k}{2})\gamma_k$, which is the first condition of (4.6).

Step 7: Obtaining (4.7). Finally, since $r_k := \frac{1}{2}\|A\hat{u}^{k+1} + B\hat{v}_k^* - c\|^2 \leq D_{\mathcal{X}}$ due to (3.2) and $\bar{R}_k \geq 0$, it follows from (A.24) that:

$$G_{k+1}(\bar{w}^{k+1}) \leq (1 - \tau_k)G_k(\bar{w}^k) + \tau_k\eta_k r_k \leq (1 - \tau_k)G_k(\bar{w}^k) + \tau_k\eta_k D_{\mathcal{X}},$$

which is indeed (4.7). \square

A.3. Special cases. We prove two corollaries: Corollary 4.7 and Corollary 4.8.

A.3.1. The proof of Corollary 4.7: The full-column rank of A . First, we show that if the conditions (4.15) hold, then (4.7) holds. Indeed, since the function d_γ^1 defined by (4.14) is concave and smooth, and its gradient is Lipschitz continuous with the Lipschitz constant $L_{d_\gamma^1} := \gamma^{-1}\|A\|^2$, with the same augment as in the proof of Lemma 4.4, we can show from (A.21) that:

$$\begin{aligned} (1 - \tau_k)G_k(\bar{w}^k) &\geq -d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + (1 - \tau_k)f(\bar{x}^k) + \tau_k f(\hat{x}^{k+1}) + \frac{(1 - \tau_k)}{2\beta_k}\|M\bar{x}^k - c\|^2 \\ &\quad + \tau_k\langle M\hat{x}^{k+1} - c, \lambda_k^* \rangle + \eta_k\left(1 + \frac{\tau_k}{2} - \frac{\|A\|^2\eta_k}{2\gamma_{k+1}}\right)\|M\hat{x}^{k+1} - c\|^2 \\ (A.25) \quad &\quad + \gamma_{k+1}\tilde{r}_k - (1 - \tau_k)\bar{r}_k + \tau_k\gamma_{k+1}\hat{r}_k - \tau_k\eta_k r_k, \end{aligned}$$

where $\tilde{r}_k := \frac{\tau_k(1 - \tau_k)}{2\|A\|^2}\|\nabla d_{\gamma_{k+1}}^1(\hat{\lambda}^k) - \nabla d_{\gamma_{k+1}}^1(\bar{\lambda}^k)\|^2$, $\bar{r}_k := \frac{1}{2}(\gamma_k - \gamma_{k+1})\|u_{\gamma_{k+1}}^*(\bar{\lambda}^k) - \bar{u}_c\|^2$, $\hat{r}_k := \frac{1}{2}\|\hat{u}^{k+1} - \bar{u}_c\|^2$, and $r_k := \frac{1}{2}\|A\hat{u}^{k+1} + B\hat{v}_k^* - c\|^2 \leq D_{\mathcal{X}}$.

Now, if the following two conditions hold:

$$(A.26) \quad \beta_{k+1} \geq (1 - \tau_k)\beta_k \quad \text{and} \quad \eta_k\left(1 + \frac{\tau_k}{2} - \frac{\|A\|^2\eta_k}{2\gamma_{k+1}}\right) \geq \frac{\tau_k^2}{2(1 - \tau_k)\beta_k},$$

then, with the same argument as the proof of (A.24), we can show that:

$$(A.27) \quad (1 - \tau_k)G_k(\bar{w}^k) \geq G_{k+1}(\bar{w}^{k+1}) + \bar{R}_k - \tau_k\eta_k r_k,$$

where $\bar{R}_k := \gamma_{k+1}\tilde{r}_k - (1 - \tau_k)\bar{r}_k + \tau_k\gamma_{k+1}\hat{r}_k$.

Next, we estimate \bar{R}_k . Let $\bar{a}_k := u_{\gamma_{k+1}}^*(\bar{\lambda}^k) - \bar{u}_c$, $\hat{a}_k := u_{\gamma_{k+1}}^*(\hat{\lambda}^k) - \bar{u}_c$. By the definition (3.1) of d_γ^1 , we have $\nabla d_\gamma^1(\lambda) := Au_\gamma^*(\lambda)$. Hence, due to Assumption: $\underline{\sigma}_A^2 := \lambda_{\min}(A^T A) > 0$, we have:

$$\|\nabla d_{\gamma_{k+1}}^1(\hat{\lambda}^k) - \nabla d_{\gamma_{k+1}}^1(\bar{\lambda}^k)\|^2 \geq \underline{\sigma}_A^2\|u_{\gamma_{k+1}}^*(\hat{\lambda}^k) - u_{\gamma_{k+1}}^*(\bar{\lambda}^k)\|^2 = \underline{\sigma}_A^2\|\hat{a}_k - \bar{a}_k\|^2.$$

Using $\kappa := \frac{\underline{\sigma}_A^2}{\|A\|^2}$, the definition of \tilde{r}_k , \bar{r}_k and \hat{r}_k , and this estimate, we can write \bar{R}_k explicitly as:

$$\begin{aligned} 2\gamma_{k+1}^{-1}\bar{R}_k &= \kappa\tau_k(1 - \tau_k)\|\bar{a}_k - \hat{a}_k\|^2 - (1 - \tau_k)(\gamma_{k+1}^{-1}\gamma_k - 1)\|\bar{a}_k\|^2 + \tau_k\|\hat{a}_k\|^2 \\ &= \tau_k(1 + \kappa - \kappa\tau_k)\|\hat{a}_k\|^2 - 2\kappa(1 - \tau_k)\tau_k\langle \hat{a}_k, \bar{a}_k \rangle \\ (A.28) \quad &\quad + (1 - \tau_k)(\kappa\tau_k - \gamma_{k+1}^{-1}\gamma_k + 1)\|\bar{a}_k\|^2. \end{aligned}$$

Hence, if $\gamma_{k+1} \geq \left(1 - \frac{\tau_k}{1+1/\kappa}\right) \gamma_k$, then $\bar{R}_k \geq 0$. This condition, $\eta_k := \frac{\gamma_{k+1}}{\|A\|^2}$ and (A.26) are indeed the conditions of (4.15).

Finally, since $r_k := \frac{1}{2}\|A\hat{u}^{k+1} + B\hat{v}_k^* - c\|^2 \leq D_{\mathcal{X}}$ due to (3.2) and $\bar{R}_k \geq 0$, it follows from (A.27) that:

$$G_{k+1}(\bar{w}^{k+1}) \leq (1 - \tau_k)G_k(\bar{w}^k) + \tau_k\eta_k r_k \leq (1 - \tau_k)G_k(\bar{w}^k) + \tau_k\eta_k D_{\mathcal{X}},$$

which is indeed (4.7).

The update rule (4.16) is in fact derived from (4.15). We can easily check that $\gamma_k\beta_k < \frac{\bar{c}^2\|A\|^2}{(k+1)^2}$, where $\bar{c} := 1 + \frac{1}{\kappa} = 1 + \frac{\|A\|^2}{\sigma_A^2}$. We finally prove the bounds (4.18). First, we consider the product $\tau_k\eta_k$. By (4.16) we have:

$$\tau_k\eta_k = \frac{\gamma_0\bar{c}^2}{\|A\|^2} \frac{1}{(k + \bar{c} + 1)^2} \leq \frac{\gamma_0\bar{c}^2}{(\bar{c} - 1)\|A\|^2} \left[\frac{1}{k + \bar{c} + 1} - (1 - \tau_k)\frac{1}{k + \bar{c}} \right].$$

By induction, it follows from (4.7) and this last expression that:

$$(A.29) \quad G_k(\bar{w}^k) - \frac{\gamma_0\bar{c}}{\|A\|^2} \frac{1}{k + \bar{c}} \leq \bar{c}_k \left(G_0(\bar{w}^0) - \frac{\gamma_0\bar{c}}{\|A\|^2} \right) \leq 0,$$

whenever $G_0(\bar{w}^0) \leq \frac{\gamma_0\bar{c}}{\|A\|^2}$. Since \bar{w}^0 is given by (4.17), with the same argument as the proof of Lemma 4.1, we can show that if $\beta_0 \geq \frac{\gamma_0}{(3\gamma_0 - \eta_0\|A\|^2)\eta_0}$, then $G_0(\bar{w}^0) \leq \eta_0 D_{\mathcal{X}}$. However, from (4.16), we can see that $\eta_0 = \frac{\bar{c}\gamma_0}{(\bar{c}+1)\|A\|^2}$ and $\beta_0 = \frac{\|A\|^2\bar{c}(\bar{c}+1)}{\gamma_0(2\bar{c}+1)}$. Using these quantities, we can easily show that $\beta_0 \geq \frac{\gamma_0}{(3\gamma_0 - \eta_0\|A\|^2)\eta_0}$. Moreover, $G_0(\bar{w}^0) \leq \eta_0 D_{\mathcal{X}} = \frac{\bar{c}\gamma_0}{(\bar{c}+1)\|A\|^2} < \frac{\gamma_0\bar{c}}{\|A\|^2}$. Hence, (A.29) holds. Finally, it remains to use Lemma 3.3 and $\gamma_k\beta_k < \frac{\bar{c}^2\|A\|^2}{(k+1)^2}$ to obtain (4.18). \square

A.3.2. The proof of Corollary 4.8: The strong convexity of g . First, we show that if the conditions (4.19) hold, then (4.20) holds. Since ∇d_0^1 defined by (2.3) is Lipschitz continuous with the Lipschitz constant $L_{d_0^1} := \mu_g^{-1}\|A\|^2$, similarly to the proof of Lemma 4.3, we have:

$$(A.30) \quad d(\bar{\lambda}^{k+1}) \geq d(\lambda) + \frac{1}{\eta_k} \langle \hat{\lambda}^k - \bar{\lambda}^{k+1}, \lambda - \hat{\lambda}^k \rangle + \left(\frac{1}{\eta_k} - \frac{\|A\|^2}{2\mu_g} \right) \|\hat{\lambda}^k - \bar{\lambda}^{k+1}\|^2.$$

With the same argument as in the proof of Lemma 4.4, we can show from (A.21) that:

$$(A.31) \quad \begin{aligned} (1 - \tau_k)G_{\beta_k}(\bar{w}^k) &\geq -d(\bar{\lambda}^{k+1}) + (1 - \tau_k)f(\bar{x}^k) + \tau_k f(\hat{x}^{k+1}) + \frac{(1 - \tau_k)}{2\beta_k} \|M\bar{x}^k - c\|^2 \\ &+ \tau_k \langle M\hat{x}^{k+1} - c, \lambda_k^* \rangle + \eta_k \left(1 + \frac{\tau_k}{2} - \frac{\|A\|^2\eta_k}{2\mu_g} \right) \|M\hat{x}^{k+1} - c\|^2 - \tau_k\eta_k r_k, \end{aligned}$$

where $G_{\beta_k}(\bar{w}^k) := f_{\beta_k}(\bar{x}^k) - d(\bar{\lambda}^k)$ and $r_k := \frac{1}{2}\|A\hat{u}^{k+1} + B\hat{v}_k^* - c\|^2 \leq D_{\mathcal{X}}$. Now, if the following two conditions hold:

$$(A.32) \quad \beta_{k+1} \geq (1 - \tau_k)\beta_k \quad \text{and} \quad \eta_k \left(1 + \frac{\tau_k}{2} - \frac{\|A\|^2\eta_k}{2\mu_g} \right) \geq \frac{\tau_k^2}{2(1 - \tau_k)\beta_k},$$

then, with the same argument as the proof of (A.24), we have:

$$(A.33) \quad (1 - \tau_k)G_{\beta_k}(\bar{w}^k) \geq G_{\beta_{k+1}}(\bar{w}^{k+1}) - \tau_k\eta_k r_k.$$

Finally, since $r_k := \frac{1}{2}\|A\hat{u}^{k+1} + B\hat{v}_k^* - c\|^2 \leq D_{\mathcal{X}}$ due to (3.2), (A.33) leads to:

$$G_{\beta_{k+1}}(\bar{w}^{k+1}) \leq (1 - \tau_k)G_{\beta_k}(\bar{w}^k) + \tau_k\eta_k r_k \leq (1 - \tau_k)G_{\beta_k}(\bar{w}^k) + \tau_k\eta_k D_{\mathcal{X}},$$

which is indeed (4.20).

The update rule (4.21) is in fact derived from (4.19). We finally prove the bounds (4.23). First, we consider the product $\tau_k \eta_k$. By (4.21) we have:

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

By induction, it follows from (4.20) and this last expression that:

$$(A.34) \quad G_{\beta_k}(\bar{w}^k) - \frac{4\mu_g D\mathcal{X}}{\|A\|^2(k+2)} \leq \omega_k \left(G_{\beta_0}(\bar{w}^0) - \frac{4\mu_g D\mathcal{X}}{\|A\|^2} \right) \leq 0,$$

whenever $G_{\beta_0}(\bar{w}^0) \leq \frac{4\mu_g D\mathcal{X}}{\|A\|^2}$. Since \bar{w}^0 is given by (4.22), with the same argument as the proof of Lemma 4.1, we can show that if $\frac{1}{\beta_0} \leq 3\eta_0 - \frac{\|A\|^2 \eta_0^2}{\mu_g}$, then $G_{\beta_0}(\bar{w}^0) \leq \mu_0 D\mathcal{X}$. However, from the update rule (4.21), we can see that $\eta_0 = \frac{2\mu_g}{3\|A\|^2}$ and $\beta_0 = \frac{\|A\|^2}{\mu_g}$. Using these quantities, we can clearly show that $\frac{1}{\beta_0} \leq 3\eta_0 - \frac{\|A\|^2 \eta_0^2}{\mu_g}$. Moreover, $G_{\beta_0}(\bar{w}^0) \leq \eta_0 D\mathcal{X} < \frac{4\mu_g}{\|A\|^2} D\mathcal{X}$. Hence, (A.34) holds. Finally, it remains to use Lemma 3.3 to obtain (4.23). \square

A.4. Convergence analysis of Algorithm 2. This appendix provides the full proof of Lemma 5.1 and Lemma 5.2.

A.4.1. The proof of Lemma 5.1: ADMM quadratic surrogate of d_γ . Similarly to the proof of Lemma 4.3, if we denote by $\bar{g}_\gamma(\cdot) := g_\gamma(\cdot) + \delta_{\mathcal{U}}(\cdot)$, and $\bar{h}(\cdot) := h(\cdot) + \delta_{\mathcal{V}}(\cdot)$, then, by the definition (3.1), we can write $d_\gamma^1(\lambda) := -\bar{g}_\gamma^*(-A^T \lambda)$ and $d_0^2(\lambda) := -\bar{h}^*(-B^T \lambda)$.

Next, we write the optimality condition for two subproblems in (5.1) and using the third line of (5.1) and the definition $\bar{\lambda}^k := \hat{\lambda}^k + \rho_k(A\hat{u}^{k+1} + B\hat{v}^k - c)$ to get:

$$(A.35) \quad \begin{cases} 0 \in \partial \bar{g}_{\gamma_{k+1}}(\hat{u}^{k+1}) + A^T \hat{\lambda}^k + \rho_k A^T (A\hat{u}^{k+1} + B\hat{v}^k - c) \equiv \partial \bar{g}_{\gamma_{k+1}}(\hat{u}^{k+1}) + A^T \bar{\lambda}^k, \\ 0 \in \partial \bar{h}(\hat{v}^{k+1}) + B^T \hat{\lambda}^k + \eta_k B^T (A\hat{u}^{k+1} + B\hat{v}^{k+1} - c) \equiv \partial \bar{h}(\hat{v}^{k+1}) + B^T \bar{\lambda}^{k+1}. \end{cases}$$

The first condition of (A.35) can be rewritten as $-A^T \bar{\lambda}^k \in \partial \bar{g}_{\gamma_{k+1}}(\hat{u}^{k+1})$, which is equivalent to $\hat{u}^{k+1} \in \partial \bar{g}_{\gamma_{k+1}}^*(-A^T \bar{\lambda}^k)$. Multiplying this inclusion by A and noting that $d_{\gamma_{k+1}}^1(\lambda) = -\bar{g}_{\gamma_{k+1}}^*(-A^T \lambda)$, we have $A\hat{u}^{k+1} \in A\partial \bar{g}_{\gamma_{k+1}}^*(-A^T \bar{\lambda}^k) = \nabla d_{\gamma_{k+1}}^1(\bar{\lambda}^k)$ due to the differentiability of $d_{\gamma_{k+1}}^1$. Similarly, from the second line of (A.4) we have $B\hat{v}^{k+1} \in B\partial \bar{h}^*(-B^T \bar{\lambda}^{k+1}) = \partial d_0^2(\bar{\lambda}^{k+1})$. Summing up these inclusions an using the third line of (5.1), we obtain $\eta_k^{-1}(\bar{\lambda}^{k+1} - \hat{\lambda}^k) = A\hat{u}^{k+1} + B\hat{v}^{k+1} - c \in \nabla d_{\gamma_{k+1}}^1(\bar{\lambda}^k) + \partial d_0^2(\hat{\lambda}^k) - c$. This expression is equivalent to the following:

$$(A.36) \quad c + \eta_k^{-1}(\bar{\lambda}^{k+1} - \hat{\lambda}^k) - \nabla d_{\gamma_{k+1}}^1(\bar{\lambda}^k) \in \partial d_0^2(\hat{\lambda}^k).$$

Since $\nabla d_\gamma^1(\cdot) = Au_\gamma^*(\cdot)$ is Lipschitz continuous with a Lipschitz constant $L_{d_\gamma^1} := \gamma^{-1}$ due to Lemma 3.2, by (A.2) of Theorem A.2, for any $\lambda \in \mathbb{R}^m$, we have:

$$(A.37) \quad \begin{aligned} d_{\gamma_{k+1}}^1(\lambda) &\leq d_{\gamma_{k+1}}^1(\bar{\lambda}^k) + \langle \nabla d_{\gamma_{k+1}}^1(\bar{\lambda}^k), \lambda - \bar{\lambda}^k \rangle, \\ d_{\gamma_{k+1}}^1(\bar{\lambda}^k) + \langle \nabla d_{\gamma_{k+1}}^1(\bar{\lambda}^k), \bar{\lambda}^{k+1} - \bar{\lambda}^k \rangle &\leq d_{\gamma_{k+1}}^1(\bar{\lambda}^{k+1}) + \frac{1}{2\gamma_{k+1}} \|\bar{\lambda}^{k+1} - \bar{\lambda}^k\|^2. \end{aligned}$$

On the other hand, since d_0^2 is concave, we have $d_0^2(\lambda) \leq d_0^2(\bar{\lambda}^{k+1}) + \langle s_k, \lambda - \bar{\lambda}^{k+1} \rangle$ for any $s_k \in \partial d_0^2(\bar{\lambda}^{k+1})$. Taking $s_k = \eta_k^{-1}(\bar{\lambda}^{k+1} - \hat{\lambda}^k) + c - \nabla d_{\gamma_{k+1}}^1(\bar{\lambda}^k) \in \partial d_0^2(\bar{\lambda}^{k+1})$ from (A.7), we have:

$$d_0^2(\lambda) \leq d_0^2(\bar{\lambda}^{k+1}) + \langle \frac{1}{\eta_k}(\bar{\lambda}^{k+1} - \hat{\lambda}^k) + c - \nabla d_{\gamma_{k+1}}^1(\bar{\lambda}^k), \lambda - \bar{\lambda}^{k+1} \rangle.$$

Summing up this inequality and the first inequality of (A.37), and then using the second inequality of (A.37) and $d_{\gamma_{k+1}}(\cdot) = d_{\gamma_{k+1}}^1(\cdot) + d_0^2(\cdot) - \langle c, \cdot \rangle$, we can derive:

$$\begin{aligned}
d_{\gamma_{k+1}}(\lambda) &\leq d_{\gamma_{k+1}}^1(\tilde{\lambda}^k) + \langle \nabla d_{\gamma_{k+1}}^1(\tilde{\lambda}^k), \bar{\lambda}^{k+1} - \tilde{\lambda}^k \rangle \\
&\quad + \eta_k^{-1} \langle \bar{\lambda}^{k+1} - \hat{\lambda}^k, \lambda - \bar{\lambda}^{k+1} \rangle + d_0^2(\bar{\lambda}^{k+1}) - \langle c, \bar{\lambda}^{k+1} \rangle \\
&\leq d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + \frac{1}{\eta_k} \langle \bar{\lambda}^{k+1} - \hat{\lambda}^k, \lambda - \bar{\lambda}^{k+1} \rangle + \frac{1}{2\gamma_{k+1}} \|\bar{\lambda}^{k+1} - \tilde{\lambda}^k\|^2 \\
&= d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) - \frac{1}{\eta_k} \langle \hat{\lambda}^k - \bar{\lambda}^{k+1}, \lambda - \hat{\lambda}^k \rangle - \frac{1}{\eta_k} \|\bar{\lambda}^{k+1} - \hat{\lambda}^k\|^2 + \frac{1}{2\gamma_{k+1}} \|\bar{\lambda}^{k+1} - \tilde{\lambda}^k\|^2,
\end{aligned}$$

which is exactly the inequality (5.3). \square

A.4.2. The proof of Lemma 5.2: Maintaining the gap reduction condition.

The proof of this lemma is also divided into several steps.

Step 1: Key bound on $d_{\gamma_{k+1}}$. With \tilde{r}_k defined by (A.11) and $Mx \equiv Au + Bv$, similarly to the proof of Lemma 4.4, for any $\tau_k \in [0, 1]$, $\bar{\lambda}^k, \hat{\lambda}^k \in \mathbb{R}^n$ and $\lambda^k := (1 - \tau_k)\bar{\lambda}^k + \tau_k\hat{\lambda}^k$, we have:

$$(A.38) \quad d_{\gamma_{k+1}}(\lambda^k) \geq (1 - \tau_k)d_{\gamma_{k+1}}(\bar{\lambda}^k) + \tau_k d_{\gamma_{k+1}}(\hat{\lambda}^k) + \gamma_{k+1}\tilde{r}_k.$$

Next, using Lemma 5.1 and (5.1) with $\lambda = \lambda^k$ and $\bar{\lambda}^{k+1} - \hat{\lambda}^k = \eta_k(M\hat{x}^{k+1} - c)$, we can estimate:

$$d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \geq d_{\gamma_{k+1}}(\lambda^k) + \langle M\hat{x}^{k+1} - c, \hat{\lambda}^k - \lambda^k \rangle + \eta_k \|M\hat{x}^{k+1} - c\|^2 - \frac{1}{2\gamma_{k+1}} \|\bar{\lambda}^{k+1} - \tilde{\lambda}^k\|^2.$$

Substituting (A.38) into this inequality, and using $\lambda^k = (1 - \tau_k)\bar{\lambda}^k + \tau_k\hat{\lambda}^k$ and $\hat{\lambda}^k = (1 - \tau_k)\bar{\lambda}^k + \tau_k\lambda_k^*$ with $\lambda_k^* := \frac{1}{\beta_k}(M\bar{x}^k - c)$ in (5.2), we get:

$$\begin{aligned}
d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) &\geq (1 - \tau_k)d_{\gamma_{k+1}}(\bar{\lambda}^k) + \tau_k d_{\gamma_{k+1}}(\hat{\lambda}^k) + \langle M\hat{x}^{k+1} - c, \hat{\lambda}^k - \lambda^k \rangle \\
&\quad + \eta_k \|M\hat{x}^{k+1} - c\|^2 + \gamma_{k+1}\tilde{r}_k - \frac{1}{2\gamma_{k+1}} \|\bar{\lambda}^{k+1} - \tilde{\lambda}^k\|^2 \\
&= (1 - \tau_k)d_{\gamma_{k+1}}(\bar{\lambda}^k) + \tau_k d_{\gamma_{k+1}}(\hat{\lambda}^k) + \tau_k \langle M\hat{x}^{k+1} - c, \lambda_k^* - \hat{\lambda}^k \rangle \\
(A.39) \quad &\quad + \eta_k \|M\hat{x}^{k+1} - c\|^2 + \gamma_{k+1}\tilde{r}_k - \frac{1}{2\gamma_{k+1}} \|\bar{\lambda}^{k+1} - \tilde{\lambda}^k\|^2.
\end{aligned}$$

Step 2: Bound on $(1 - \tau_k)G_k(\bar{w}^k)$. Let $\bar{u}_{k+1}^* := u_{\gamma_{k+1}}^*(\bar{\lambda}^k)$ be the solution of the first subproblem (5.1). Then, by (3.4), for $\gamma_{k+1}, \gamma_k > 0$, we have:

$$(A.40) \quad d_{\gamma_{k+1}}^1(\bar{\lambda}^k) \geq d_{\gamma_k}^1(\bar{\lambda}^k) - \frac{1}{2}(\gamma_k - \gamma_{k+1})\|A(\bar{u}_{k+1}^* - \bar{u}_c)\|^2 := d_{\gamma_k}^1(\bar{\lambda}^k) - \bar{r}_k,$$

where $\bar{r}_k := \frac{1}{2}(\gamma_k - \gamma_{k+1})\|A(\bar{u}_{k+1}^* - \bar{u}_c)\|^2 \geq 0$. Moreover, since $f_{\beta_k}(\bar{x}^k) = f(\bar{x}^k) + \frac{1}{2\beta_k}\|M\bar{x}^k - c\|^2$ and $G_k(\bar{w}^k) := f_{\beta_k}(\bar{x}^k) - d_{\gamma_k}(\bar{\lambda}^k)$ by (3.6), we have $d_{\gamma_k}(\bar{\lambda}^k) = f_{\beta_k}(\bar{x}^k) - G_k(\bar{w}^k)$. Combining the last expression and (A.40), we have:

$$d_{\gamma_{k+1}}(\bar{\lambda}^k) \geq f_{\beta_k}(\bar{x}^k) - G_k(\bar{w}^k) + \bar{r}_k.$$

Substituting this into (A.40) to obtain:

$$(1 - \tau_k)d_{\gamma_{k+1}}(\bar{\lambda}^k) \geq (1 - \tau_k)f(\bar{x}^k) - (1 - \tau_k)G_k(\bar{w}^k) + \frac{(1 - \tau_k)}{2\beta_k}\|M\bar{x}^k - c\|^2 - (1 - \tau_k)\bar{r}_k.$$

Using this estimate into (A.39) and exchanging $(1 - \tau_k)G_k(\bar{w}^k)$ and $d_{\gamma_{k+1}}(\bar{\lambda}^{k+1})$, we eventually obtain:

$$\begin{aligned}
(1 - \tau_k)G_k(\bar{w}^k) &\geq -d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + (1 - \tau_k)f(\bar{x}^k) + \frac{(1 - \tau_k)}{2\beta_k} \|M\bar{x}^k - c\|^2 \\
&\quad + \tau_k d_{\gamma_{k+1}}(\hat{\lambda}^k) + \tau_k \langle M\hat{x}^{k+1} - c, \lambda_k^* - \hat{\lambda}^k \rangle + \gamma_{k+1} \tilde{r}_k \\
(A.41) \quad &\quad + \eta_k \|M\hat{x}^{k+1} - c\|^2 - \frac{1}{2\gamma_{k+1}} \|\bar{\lambda}^{k+1} - \tilde{\lambda}^k\|^2 - (1 - \tau_k) \tilde{r}_k.
\end{aligned}$$

Step 3: Estimate of $d_{\gamma_{k+1}}(\hat{\lambda}^k)$. Now, we define the following functions:

$$(A.42) \quad \begin{cases} \tilde{d}_k^1(\hat{\lambda}^k) := \min_{u \in \mathcal{U}} \{g_{\gamma_{k+1}}(u) + \langle \hat{\lambda}^k, Au \rangle + \frac{\rho_k}{2} \|Au + B\hat{v}^k - c\|^2\}, \\ \tilde{d}_k^2(\hat{\lambda}^k) := \min_{v \in \mathcal{V}} \{h(v) + \langle \hat{\lambda}^k, Bv \rangle + \frac{\eta_k}{2} \|A\hat{u}^{k+1} + Bv - c\|^2\}, \\ \tilde{d}_k(\hat{\lambda}^k) := \tilde{d}_k^1(\hat{\lambda}^k) + \tilde{d}_k^2(\hat{\lambda}^k) - \langle c, \hat{\lambda}^k \rangle, \end{cases}$$

and three quantities:

$$(A.43) \quad r_k^1 := \frac{1}{2} \|A\hat{u}_{k+1}^* + B\hat{v}^k - c\|^2, \quad r_k^2 := \frac{1}{2} \|A\hat{u}^{k+1} + B\hat{v}_k^* - c\|^2 \text{ and } r_k := \rho_k r_k^1 + \eta_k r_k^2.$$

With d_γ^1 defined by (3.1) and d_0^2 defined by (2.3), we use again $\hat{u}_{k+1}^* := u_{\gamma_{k+1}}^*(\hat{\lambda}^k)$ and $\hat{v}_k^* := v^*(\hat{\lambda}^k)$ the solution of the two convex subproblems in (3.1) and (2.3), respectively. Next, using the result in Lemma A.1 with $\psi(\cdot) = g(\cdot) + \langle A^T \hat{\lambda}^k, \cdot \rangle$, $\mu = \gamma_{k+1} + \rho_k$ and $u = \hat{u}_{k+1}^*$, we first estimate the bounds between $d_{\gamma_{k+1}}^1(\hat{\lambda}^k)$ and $\tilde{d}_k^1(\hat{\lambda}^k)$ as follows:

$$\begin{aligned}
\tilde{d}_k^1(\hat{\lambda}^k) &\leq \left[g(\hat{u}_{k+1}^*) + \langle \hat{\lambda}^k, A\hat{u}_{k+1}^* \rangle + \frac{\gamma_{k+1}}{2} \|A\hat{u}_{k+1}^* - \bar{u}_c\|^2 \right]_{d_{\gamma_{k+1}}^1(\hat{\lambda}^k)} \\
&\quad + \frac{\rho_k}{2} \|A\hat{u}_{k+1}^* + B\hat{v}^k - c\|^2 - \frac{\gamma_{k+1} + \rho_k}{2} \|A\hat{u}_{k+1}^* - \hat{u}^{k+1}\|^2 \\
&= d_{\gamma_{k+1}}^1(\hat{\lambda}^k) + \frac{\rho_k}{2} \|A\hat{u}_{k+1}^* + B\hat{v}^k - c\|^2 - \frac{\gamma_{k+1} + \rho_k}{2} \|A(\hat{u}_{k+1}^* - \hat{u}^{k+1})\|^2.
\end{aligned}$$

Using r_k^1 in (A.43), the last estimate leads to:

$$(A.44) \quad d_{\gamma_{k+1}}^1(\hat{\lambda}^k) \geq \tilde{d}_k^1(\hat{\lambda}^k) - \rho_k r_k^1 + \frac{\gamma_{k+1} + \rho_k}{2} \|A\hat{u}_{k+1}^* - \hat{u}^{k+1}\|^2.$$

Similarly, by using Lemma A.1 with $\psi(\cdot) = h(\cdot) + \langle B^T \hat{\lambda}^k, \cdot \rangle$, $\mu = \eta_k$ and $u = \hat{v}_k^*$, we then estimate the bound between $d_0^2(\hat{\lambda}^k)$ and $\tilde{d}_k^2(\hat{\lambda}^k)$ as follows:

$$\begin{aligned}
\tilde{d}_k^2(\hat{\lambda}^k) &\leq h(\hat{v}_k^*) + \langle \hat{\lambda}^k, B\hat{v}_k^* \rangle + \frac{\eta_k}{2} \|A\hat{u}^{k+1} + B\hat{v}_k^* - c\|^2 - \frac{\eta_k}{2} \|B(\hat{v}_k^* - \hat{v}^{k+1})\|^2 \\
&= d_0^2(\hat{\lambda}^k) + \frac{\eta_k}{2} \|A\hat{u}^{k+1} + B\hat{v}_k^* - c\|^2 - \frac{\eta_k}{2} \|B(\hat{v}_k^* - \hat{v}^{k+1})\|^2.
\end{aligned}$$

Using r_k^2 in (A.43), this inequality also leads to:

$$(A.45) \quad d_0^2(\hat{\lambda}^k) \geq \tilde{d}_k^2(\hat{\lambda}^k) - \eta_k r_k^2.$$

Combining (A.44) and (A.45), and using the fact that $d_\gamma(\cdot) = d_\gamma^1(\cdot) + d_0^2(\cdot) - \langle c, \cdot \rangle$, we obtain:

$$\begin{aligned}
d_{\gamma_{k+1}}(\hat{\lambda}^k) &\geq \tilde{d}_k^1(\hat{\lambda}^k) + \tilde{d}_k^2(\hat{\lambda}^k) - \langle c, \hat{\lambda}^k \rangle + \frac{\rho_k + \gamma_{k+1}}{2} \|A(\hat{u}_{k+1}^* - \hat{u}^{k+1})\|^2 - (\rho_k r_k^1 + \eta_k r_k^2) \\
&= f(\hat{x}^{k+1}) + \langle M\hat{x}^{k+1} - c, \hat{\lambda}^k \rangle + \frac{\rho_k}{2} \|M\hat{x}^{k+1} - c\|^2 + \frac{\eta_k}{2} \|M\hat{x}^{k+1} - c\|^2 \\
(A.46) \quad &\quad + \frac{\gamma_{k+1} + \rho_k}{2} \|A(\hat{u}_{k+1}^* - \hat{u}^{k+1})\|^2 + \frac{\gamma_{k+1}}{2} \|A(\hat{u}^{k+1} - \bar{u}_c)\|^2 - r_k.
\end{aligned}$$

Step 4: Refining the bound on $(1 - \tau_k)G_k(\bar{w}^k)$. Let $\hat{r}_k := \|A(\hat{u}_{k+1}^* - \bar{u}_c)\|^2$. By using an elementary inequality $p\|a\|^2 + q\|b\|^2 \geq \frac{pq}{p+q}\|a - b\|^2$ with $a = A(\hat{u}_{k+1}^* - \hat{u}^{k+1})$, $b = A(\hat{u}^{k+1} - \bar{u}_c)$, $p = \gamma_{k+1} + \rho_k$ and $q = \gamma_{k+1}$, we can show that:

$$(A.47) \quad \begin{aligned} (\gamma_{k+1} + \rho_k)\|A(\hat{u}_{k+1}^* - \hat{u}^{k+1})\|^2 + \gamma_{k+1}\|A(\hat{u}^{k+1} - \bar{u}_c)\|^2 &\geq \frac{\gamma_{k+1}(\gamma_{k+1} + \rho_k)}{2\gamma_{k+1} + \rho_k}\|A(\hat{u}_{k+1}^* - \bar{u}_c)\|^2 \\ &= \frac{\gamma_{k+1}(\gamma_{k+1} + \rho_k)}{2\gamma_{k+1} + \rho_k}\hat{r}_k \geq \frac{\gamma_{k+1}}{2}\hat{r}_k. \end{aligned}$$

By using (A.47) and $\tilde{\lambda}^k := \hat{\lambda}^k + \rho_k(A\hat{u}^{k+1} + B\hat{v}^k - c)$ from Lemma 5.1, we obtain from (A.46) that:

$$d_{\gamma_{k+1}}(\hat{\lambda}^k) \geq f(\hat{x}^{k+1}) + \langle M\hat{x}^{k+1} - c, \hat{\lambda}^k \rangle + \frac{\eta_k}{2}\|M\hat{x}^{k+1} - c\|^2 + \frac{1}{2\rho_k}\|\tilde{\lambda}^k - \hat{\lambda}^k\|^2 + \frac{\gamma_{k+1}}{2}\hat{r}_k - r_k.$$

Substituting this inequality into (A.41), we can further derive:

$$(A.48) \quad \begin{aligned} (1 - \tau_k)G_k(\bar{w}^k) &\geq (1 - \tau_k)f(\bar{x}^k) + \tau_k f(\hat{x}^{k+1}) - d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + \frac{(1 - \tau_k)}{2\beta_k}\|M\bar{x}^k - c\|^2 \\ &\quad + \tau_k \langle M\hat{x}^{k+1} - c, \lambda_k^* \rangle + \frac{(1 + \tau_k)\eta_k}{2}\|M\hat{x}^{k+1} - c\|^2 + \frac{\eta_k}{2}\|M\hat{x}^{k+1} - c\|^2 \\ &\quad + \frac{\tau_k}{2\rho_k}\|\tilde{\lambda}^k - \hat{\lambda}^k\|^2 - \frac{1}{2\gamma_{k+1}}\|\tilde{\lambda}^k - \hat{\lambda}^k\|^2 + \gamma_{k+1}\tilde{r}_k + \frac{\gamma_{k+1}\tau_k}{2}\hat{r}_k - (1 - \tau_k)\bar{r}_k - \tau_k r_k. \end{aligned}$$

Let us denote by:

$$(A.49) \quad \hat{R}_k := \gamma_{k+1}\tilde{r}_k + \frac{\gamma_{k+1}\tau_k}{2}\hat{r}_k - (1 - \tau_k)\bar{r}_k,$$

where \tilde{r}_k , \hat{r}_k and \bar{r}_k have been defined previously. Since $\bar{x}^{k+1} := (1 - \tau_k)\bar{x}^k + \tau_k\hat{x}^{k+1}$ due to (5.2), and f is convex, we have $f(\bar{x}^{k+1}) = f((1 - \tau_k)\bar{x}^k + \tau_k\hat{x}^{k+1}) \leq (1 - \tau_k)f(\bar{x}^k) + \tau_k f(\hat{x}^{k+1})$. Moreover, $\bar{\lambda}^{k+1} - \hat{\lambda}^k = \eta_k(M\hat{x}^{k+1} - c)$ and $\lambda_k^* := \beta_k^{-1}(M\bar{x}^k - c)$ by (5.1) and (5.2). Using these relations, we can refine the estimate (A.48) as:

$$(A.50) \quad \begin{aligned} (1 - \tau_k)G_k(\bar{w}^k) &\geq f(\bar{w}^{k+1}) - d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + \frac{1}{2(1 - \tau_k)\beta_k}\left[(1 - \tau_k)^2\|M\bar{x}^k - c\|^2\right. \\ &\quad \left.+ 2\tau_k(1 - \tau_k)\langle M\bar{x}^k - c, M\hat{x}^{k+1} - c \rangle + \tau_k^2\|M\hat{x}^{k+1} - c\|^2\right] \\ &\quad + \left[\frac{(1 + \tau_k)\eta_k}{2} - \frac{\tau_k^2}{2(1 - \tau_k)\beta_k}\right]\|M\hat{x}^{k+1} - c\|^2 \\ &\quad + \frac{1}{2\eta_k}\|\bar{\lambda}^{k+1} - \hat{\lambda}^k\|^2 + \frac{\tau_k}{2\rho_k}\|\tilde{\lambda}^k - \hat{\lambda}^k\|^2 - \frac{1}{2\gamma_{k+1}}\|\bar{\lambda}^{k+1} - \tilde{\lambda}^k\|^2 + \hat{R}_k - \tau_k r_k. \end{aligned}$$

Using again the elementary inequality $p\|a\|^2 + q\|b\|^2 \geq \frac{pq}{p+q}\|a - b\|^2$ with $a = \bar{\lambda}^{k+1} - \hat{\lambda}^k$, $b = \tilde{\lambda}^k - \hat{\lambda}^k$, $p = \eta_k^{-1}$ and $q = \rho_k^{-1}\tau_k$, we can see that if $\gamma_{k+1} \geq \eta_k + \frac{\rho_k}{\tau_k}$, then:

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

Substituting this inequality into (A.50) and noting that $\bar{x}^{k+1} = (1 - \tau_k)\bar{x}^k + \tau_k\hat{x}^{k+1}$, we have:

$$(A.51) \quad \begin{aligned} (1 - \tau_k)G_k(\bar{w}^k) &\geq f(\bar{w}^{k+1}) - d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + \frac{1}{2(1 - \tau_k)\beta_k}\|M\bar{x}^{k+1} - c\|^2 \\ &\quad + \left[\frac{(1 + \tau_k)\eta_k}{2} - \frac{\tau_k^2}{2(1 - \tau_k)\beta_k}\right]\|M\hat{x}^{k+1} - c\|^2 + \hat{R}_k - \tau_k r_k. \end{aligned}$$

Step 5: Conditions on the parameters. From (A.51), we assume that:

$$\beta_{k+1} \geq (1 - \tau_k)\beta_k, \text{ and } (1 - \tau_k^2)\eta_k\beta_k \geq \tau_k^2.$$

These are indeed the first and the third conditions of (5.4). Moreover, we can further estimate (A.51) as:

$$(1 - \tau_k)G_k(\bar{w}^k) \geq f(\bar{w}^{k+1}) - d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) + \frac{1}{2\beta_{k+1}}\|M\bar{x}^{k+1} - c\|^2 + \hat{R}_k - \tau_k r_k$$

$$(A.52) \quad \stackrel{(3.6)}{=} G_{k+1}(\bar{w}^{k+1}) + \hat{R}_k - \tau_k r_k.$$

Step 6: Bound on \hat{R}_k . Let us now estimate \hat{R}_k defined by (A.49). Let $\bar{a}_k := A(u_{\gamma_{k+1}}^*(\bar{\lambda}^k) - \bar{u}_c)$, $\hat{a}_k := A(u_{\gamma_{k+1}}^*(\hat{\lambda}^k) - \bar{u}_c)$. By the definition (3.1) of d_{γ}^1 , we have $\nabla d_{\gamma_{k+1}}^1(\lambda) := Au_{\gamma_{k+1}}^*(\lambda)$ due to Lemma 3.2. Therefore, using the definition of \tilde{r}_k , \bar{r}_k and \hat{r}_k , we can write \hat{R}_k explicitly as:

$$\frac{2\hat{R}_k}{\gamma_{k+1}} = \tau_k(1 - \tau_k)\|\bar{a}_k - \hat{a}_k\|^2 + \frac{\tau_k}{2}\|\hat{a}_k\|^2 - (1 - \tau_k)\left(\frac{\gamma_k}{\gamma_{k+1}} - 1\right)\|\bar{a}_k\|^2$$

$$= \tau_k(3/2 - \tau_k)\|\hat{a}_k\|^2 - 2(1 - \tau_k)\tau_k\langle \hat{a}_k, \bar{a}^k \rangle + (1 - \tau_k)(\tau_k - \gamma_{k+1}^{-1}\gamma_k + 1)\|\bar{a}^k\|^2.$$

Hence, if $(3 - \tau_k)\gamma_{k+1} \geq (3 - 2\tau_k)\gamma_k$, then $\hat{R}_k \geq 0$. The last condition and $\gamma_{k+1} \geq \eta_k + \frac{\rho_k}{\tau_k}$ are exactly the second and the fourth condition of (5.4).

Step 6: Obtaining (5.5). Under the conditions (5.4), we have from (A.52) and the definition of r_k that:

$$G_{k+1}(\bar{w}^{k+1}) \leq (1 - \tau_k)G_k(\bar{w}^k) + \tau_k(\rho_k r_k^1 + \eta_k r_k^2).$$

However, from the definition (3.2) of $D_{\mathcal{X}}$, we can easily see that $r_k^1 \leq D_{\mathcal{X}}$ and $r_k^2 \leq D_{\mathcal{X}}$. Therefore, we obtain from the last inequality that $G_{k+1}(\bar{w}^{k+1}) \leq (1 - \tau_k)G_k(\bar{w}^k) + \tau_k(\rho_k + \eta_k)D_{\mathcal{X}}$, which is exactly (5.5). \square

REFERENCES

- [1] H.H. BAUSCHKE AND P. COMBETTES, *Convex analysis and monotone operators theory in Hilbert spaces*, Springer-Verlag, 2011.
- [2] A. BECK AND M. TEOULLE, *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*, SIAM J. Imaging Sciences, 2 (2009), pp. 183–202.
- [3] A. BECK AND M. TEOULLE, *A fast dual proximal gradient algorithm for convex minimization and applications*, Oper. Res. Letter, 42 (2014), pp. 1–6.
- [4] A. BELLONI, V. CHERNOZHUKOV, AND L. WANG, *Square-root LASSO: Pivotal recovery of sparse signals via conic programming*, Biometrika, 94 (2011), pp. 791–806.
- [5] DIMITRI P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series)*, Athena Scientific, 1996.
- [6] S. BOYD, N. PARIKH, E. CHU, B. PELEATO, AND J. ECKSTEIN, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning, 3 (2011), pp. 1–122.
- [7] S. BOYD AND L. VANDENBERGHE, *Convex Optimization*, University Press, Cambridge, 2004.
- [8] LUIS M BRICENO-ARIAS AND PATRICK L COMBETTES, *A monotone + skew splitting model for composite monotone inclusions in duality*, SIAM Journal on Optimization, 21 (2011), pp. 1230–1250.
- [9] A. CHAMBOLLE AND T. POCK, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of Mathematical Imaging and Vision, 40 (2011), pp. 120–145.
- [10] S. H. CHAN, R. KHOSHABEH, K.B. GIBSON, P. E. GILL, AND T.Q. NGUYEN, *An Augmented Lagrangian Method for Total Variation Video Restoration*, IEEE Trans. Image Processing, 20 (2011), pp. 3097–3111.
- [11] P. COMBETTES, *Solving monotone inclusions via compositions of nonexpansive averaged operators*, Optimization, 53 (2004), pp. 475–504.
- [12] D. DAVIS, *Convergence rate analysis of the forward-Douglas-Rachford splitting scheme*, UCLA CAM report 14-73, (2014).

- [13] D. DAVIS AND W. YIN, *Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions*, UCLA CAM report 14-58, (2014).
- [14] J. ECKSTEIN AND D. BERTSEKAS, *On the Douglas - Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Program., 55 (1992), pp. 293–318.
- [15] F. FACCHINEI AND J.-S. PANG, *Finite-dimensional variational inequalities and complementarity problems*, vol. 1-2, Springer-Verlag, 2003.
- [16] D. GABAY AND B. MERCIER, *A dual algorithm for the solution of nonlinear variational problems via finite element approximation*, Computers & Mathematics with Applications, 2 (1976), pp. 17 – 40.
- [17] T. GOLDSTEIN, B. ODOGHOUE, AND S. SETZER, *Fast Alternating Direction Optimization Methods*, SIAM J. Imaging Sci., 7 (2012), pp. 1588–1623.
- [18] M. GRANT, S. BOYD, AND Y. YE, *Disciplined convex programming*, in Global Optimization: From Theory to Implementation, L. Liberti and N. Maculan, eds., Nonconvex Optimization and its Applications, Springer, 2006, pp. 155–210.
- [19] B.S. HE AND X.M. YUAN, *On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method*, SIAM J. Numer. Anal., 50 (2012), pp. 700–709.
- [20] S. LEFKIMMIATIS AND M. UNSER, *Poisson Image Reconstruction with Hessian Schatten-Norm Regularization*, IEEE Trans. Image Processing, 22 (2013), pp. 4314–4327.
- [21] I. NECOARA AND J.A.K. SUYKENS, *Applications of a smoothing technique to decomposition in convex optimization*, IEEE Trans. Automatic control, 53 (2008), pp. 2674–2679.
- [22] V. NEDELICU, I. NECOARA, AND Q. TRAN-DINH, *Computational Complexity of Inexact Gradient Augmented Lagrangian Methods: Application to Constrained MPC*, SIAM J. Optim. Control, 52 (2014), pp. 3109–3134.
- [23] 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. Op, 15 (2004), pp. 229–251.
- [24] A. NEMIROVSKII AND D. YUDIN, *Problem Complexity and Method Efficiency in Optimization*, Wiley Interscience, 1983.
- [25] Y. NESTEROV, *Introductory lectures on convex optimization: a basic course*, vol. 87 of Applied Optimization, Kluwer Academic Publishers, 2004.
- [26] ———, *Excessive gap technique in nonsmooth convex minimization*, SIAM J. Optimization, 16 (2005), pp. 235–249.
- [27] ———, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.
- [28] Y. OUYANG, Y. CHEN, G. LANG. LAN., AND E. JR. PASILIAO, *An accelerated linearized alternating direction method of multiplier*, Tech, (2014).
- [29] N. PARIKH AND S. BOYD, *Proximal algorithms*, Foundations and Trends in Optimization, 1 (2013), pp. 123–231.
- [30] R. T. ROCKAFELLAR, *Convex Analysis*, vol. 28 of Princeton Mathematics Series, Princeton University Press, 1970.
- [31] R. SHEFI AND M. TEBoulLE, *Rate of Convergence Analysis of Decomposition Methods Based on the Proximal Method of Multipliers for Convex Minimization*, SIAM J. Optim., 24 (2014), pp. 269–297.
- [32] Q. TRAN-DINH AND V. CEVHER, *Constrained convex minimization via model-based excessive gap*, in Proc. the Neural Information Processing Systems Foundation conference (NIPS2014), Montreal, Canada, December 2014, pp. 1–9.
- [33] ———, *A primal-dual algorithmic framework for constrained convex minimization*, Tech. Report., LIONS, (2014), pp. 1–54.
- [34] P. TSENG, *Alternating projection-proximal methods for convex programming and variational inequalities*, SIAM J. Optimization, 7 (1997), pp. 951–965.
- [35] P. TSENG AND D.P. BERTSEKAS, *Relaxation methods for problems with strictly convex cost and linear constraints*, Math. Oper. Research, 16 (1991), pp. 462–481.
- [36] H. WANG AND A. BANERJEE, *Bregman Alternating Direction Method of Multipliers*, in Proc. the Neural Information Processing Systems Foundation conference (NIPS2014), Montreal, Canada, December 2014, pp. 1–9.
- [37] J. YANG AND Y. ZHANG, *Alternating direction algorithms for ℓ_1 -problems in compressive sensing*, SIAM J. Scientific Computing, 33 (2011), pp. 250–278.
- [38] A. YURTSEVER, Q. TRAN-DINH, AND V. CEVHER, *Universal primal-dual proximal-gradient methods*, Tech. Report. (LIONS, EPFL), Available at: <http://arxiv.org/pdf/1502.03123.pdf>. (2015).