A SMOOTH PRIMAL-DUAL OPTIMIZATION FRAMEWORK FOR NONSMOOTH COMPOSITE CONVEX MINIMIZATION[*]

QUOC TRAN-DINH[†], OLIVIER FERCOQ[‡], AND VOLKAN CEVHER[§]

Abstract. We propose a new and low per-iteration complexity first-order primal-dual optimization framework for a convex optimization template with broad applications. Our analysis relies on a novel combination of three classic ideas applied to the primal-dual gap function: smoothing, acceleration, and homotopy. The algorithms due to the new approach achieve the best-known convergence rate results, in particular when the template consists of only nonsmooth functions. We also outline a restart strategy for the acceleration to significantly enhance the practical performance. We demonstrate relations with the augmented Lagrangian method and show how to exploit the strongly convex objectives with rigorous convergence rate guarantees. We provide representative examples to illustrate that the new methods can outperform the state of the art, including Chambolle–Pock, and the alternating direction method-of-multipliers algorithms. We also compare our algorithms with the well-known Nesterov smoothing method.

Key words. gap reduction technique, first-order primal-dual methods, augmented Lagrangian, smoothing techniques, homotopy, separable convex minimization, parallel and distributed computation

AMS subject classifications. 90C25, 90C06, 90-08

DOI. 10.1137/16M1093094

1. Introduction. We introduce a new analysis framework for designing primal-dual optimization algorithms to obtain numerical solutions to the following convex optimization template described in the primal space:

$$(1) \qquad P^\star := \min_{x \in \mathbb{R}^n} \{P(x) := f(x) + g(Ax)\},$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ are proper, closed and convex functions and $A \in \mathbb{R}^{m \times n}$ is given. For generality, we do not impose any smoothness assumption on f and g. In particular, we refer to (1) as a nonsmooth composite minimization problem.

Associated with the primal problem (1), we define the following dual formulation:

$$(2) \qquad D^\star := \max_{y \in \mathbb{R}^m} \left\{ D(y) := -f^*(-A^\top y) - g^*(y) \right\},$$

where f^* and g^* are the Fenchel conjugate of f and g, respectively. Clearly, (2) has the same form as (1) in the dual space.

The templates (1)–(2) provide a unified formulation for a broad set of applications in various disciplines, see, e.g., [8, 12, 14, 16, 44, 58, 73]. While problem (1)

[†]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill (UNC), Chapel Hill, NC 27599 (quoctd@email.unc.edu).

[‡]LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France (olivier.fercoq@telecom-paristech.fr).

[§]Laboratory for Information and Inference Systems (LIONS), École Polytechnique Fédérale de Lausanne (EPFL), CH1015, Lausanne, Switzerland (volkan.cevher@epfl.ch).

is presented in the unconstrained form, it automatically covers constrained settings by means of indicator functions. For example, (1) covers the following prototypical optimization template via $g(z) := \delta_{\{c\}}(z)$ (i.e., the indicator function of the convex set $\{c\}$):

$$(3) \qquad f^\star := \min_{x \in \mathbb{R}^n} \left\{ f(x) + \delta_{\{c\}}(Ax) \right\} \equiv \min_{x \in \mathbb{R}^n} \left\{ f(x) \mid Ax = c \right\},$$

where f is a proper, closed and convex function as in (1). Note that (3) is sufficiently general to cover standard convex optimization subclasses, such as conic programming, monotropic programming, and geometric programming, as specific instances [7, 9, 11].

Among classical convex optimization methods, the primal-dual approach is perhaps one of the best candidates to solve the primal-dual pair (1)–(2). Theory and methods along this approach have been developed for several decades and have led to a diverse set of algorithms; see, e.g., [2, 11, 15, 17, 18, 21, 23, 24, 25, 26, 28, 32, 34, 35, 38, 39, 42, 46, 47, 48, 55, 60, 61, 64, 71], and the references quoted therein. A more thorough comparison between existing primal-dual methods and our approach in this paper is postponed to section 7. There are several reasons for our emphasis on first-order primal-dual methods for (1)–(2), with the most obvious one being their scalability. Coupled with recent demand for low- to medium-accuracy solutions in applications, these methods indeed provide important trade-offs between the per-iteration complexity and the iteration-convergence rate along with the ability to distribute and decentralize the computation.

Unfortunately, the newfound popularity of primal-dual optimization has led to an explosion in the number of different algorithmic variants, each of which requires different set of assumptions on problem settings or methods, such as strong convexity, error bound conditions, metric regularity, Lipschitz gradient, Kurdyka–Łojasiewicz conditions, or penalty parameter tuning [13, 41, 40]. As a result, the optimal choice of the algorithm for a given application is often unclear, as it is not guided by theoretical principles but rather trial-and-error procedures, which can incur unpredictable computational costs. A vast list of key references can be found, e.g., in [15, 64].

To this end, we address the following key question: Can we construct heuristic-free, accelerated first-order primal-dual methods for nonsmooth composite minimization that have the best-known convergence rate guarantees? To our best knowledge, this question has never been addressed fully in a unified fashion in this generality. Intriguingly, our theory is still applicable to the smooth cases of f without requiring either Lipschitz gradient or strongly convex-type assumption. Such a model covers several important applications, such as graphical learning models and Poisson imaging reconstruction [70].

1.1. Our approach. Associated with the primal problem (1) and the dual one (2), we define

$$(4) \qquad G(w) := P(x) - D(y)$$

as a primal-dual gap function, where $w := (x, y)$ is the concatenated primal-dual variable. The gap function G in (4) is convex in terms of w. Under strong duality, we have $G(w^\star) = 0$ if and only if $w^\star := (x^\star, y^\star)$ is a primal-dual solution of (1) and (2).

The gap function (4) is widely used in convex optimization and variational inequalities; see, e.g., [29]. Several researchers have already used the gap function as a tool to characterize the convergence of optimization algorithms, e.g., within a variational inequality framework [15, 35, 61].

In stark contrast with the existing literature, our analysis relies on a novel combination of three ideas applied to the primal-dual gap function: *smoothing, acceleration,* and *homotopy*. While some combinations of these techniques have already been studied in the literature, their full combination is important for the desiderata and has not been studied yet.

Smoothing: We can obtain a smoothed estimate of the gap function within Nesterov's smoothing technique applied to f and g [4, 56]. In what follows, we denote the smoothed gap function by $G_{\gamma\beta}(w) := P_\beta(x) - D_\gamma(y)$ to approximate the primal-dual gap function $G(w)$, where P_β is a smoothed approximation to P depending on the smoothness parameter $\beta > 0$ and D_γ is a smoothed approximation to D depending on the smoothness parameter $\gamma > 0$. By smoothed approximation, we mean the same max-form approximation as [56]. However, it is still unclear how to properly update these smoothness parameters in primal-dual methods.

Acceleration: Using an accelerated scheme, we will design new primal-dual decomposition methods that satisfy the following smoothed gap reduction model:

$$(5) \qquad G_{\gamma_{k+1}\beta_{k+1}}(\bar{w}^{k+1}) \le (1 - \tau_k)G_{\gamma_k\beta_k}(\bar{w}^k) + \psi_k,$$

where $\{\bar{w}^k\}$ and the parameters are generated by the algorithms with $\tau_k \in [0,1)$ and $\{\max\{\psi_k, 0\}\}$ converges to zero. Similar ideas have been proposed before; for instance, Nesterov's excessive gap technique [55] is a special case of the gap reduction model (5) when $\psi_k \le 0$ (see [67]).

Homotopy: We will design algorithms to maintain (5) while simultaneously updating β_k, γ_k, and τ_k to zero to achieve the best-known convergence rate based on the assumptions imposed on the problem template. This strategy will also allow our theoretical guarantees not to depend on the diameter of the feasible set of (3). A similar technique is also proposed in [55], but only for symmetric primal-dual methods. It is also used in conjunction with Nesterov's smoothing technique in [10] for unconstrained problem but had only an $\mathcal{O}(\ln(k)/k)$ convergence rate.

Note that without homotopy, we can directly apply Nesterov's accelerated methods to minimize the smoothed gap function $G_{\gamma\beta}$ for given $\gamma > 0$ and $\beta > 0$. In this case, these smoothness parameters must be fixed a priori depending on the desired accuracy and the prox-diameter of both primal and dual problems, which may not be applicable to (3) due to the unboundedness of the dual feasible domain.

1.2. Our contributions. Our main contributions can be summarized as follows:

(a) (*Theory*) We propose to use differentiable smoothing prox function to smooth both primal and dual objective functions, which allows us to update the smoothness parameters in a heuristic-free manner. We introduce a new model-based gap reduction condition for constructing novel first-order primal-dual methods that can operate in a black-box fashion (in the sense of [54]). Our analysis technique unifies several classical concepts in convex optimization, from Auslander's gap function [1] and Nesterov's smoothing technique [4, 56] to the accelerated proximal gradient descent method, in a nontrivial manner. We also prove a fundamental bound on the primal objective residual and the feasibility violation for (3), which leads to the main results of our convergence guarantees.

(b) (*Algorithms and convergence theory*) We propose two novel primal-dual first-order algorithms for solving (1) and (3). The first algorithm requires performing only one primal step and one dual step without using any primal averaging scheme. The second algorithm needs one primal step and two dual

steps but using a weighted averaging scheme on the primal. We prove an $\mathcal{O}(1/k)$ convergence rate on the objective residual $P(\bar{x}^k) - P^\star$ of (1) for both algorithms, which is the best-known in the literature for the fully nonsmooth setting. For the constrained case (3), we also prove the convergence of both algorithms in terms of the primal objective residual and the feasibility violation; both achieve an $\mathcal{O}(1/k)$ convergence rate and are independent of the prox-diameters unlike existing smoothing techniques [4, 55, 56].

(c) (*Special cases*) We illustrate that the new techniques enable us to exploit additional structures, including the augmented Lagrangian smoothing scheme, and the strong convexity of the objectives. We show the flexibility of our framework by applying it to different constrained settings including conic programs.

Let us emphasize some key aspects of this work in detail. First, our characterization is radically different from existing results such as [5, 15, 27, 34, 35, 61, 64] thanks to the separation of the convergence rates for primal objective residual and the feasibility gap for (3). We believe that this is important since the separated constraint feasibility guarantee can be interpreted as a consensus rate in distributed optimization. Second, our assumptions cover a broader class of problems: we can trade-off the primal objective residual and the feasibility gap without any heuristic strategy on the algorithmic parameters while maintaining the best-known convergence rate for a class of fully nonsmooth convex problems in (3). Third, our augmented Lagrangian algorithm generates simultaneously both the primal-dual sequence compared to existing augmented Lagrangian algorithms, while it maintains its $\mathcal{O}\left(\frac{1}{k^2}\right)$-worst-case convergence rate both on the objective residual and on the feasibility gap. Fourth, we also describe how to adapt known structures on the objective and the constraint components, such as strong convexity to obtain new variants of our methods. Fifth, this work significantly expands on our earlier conference work [67] not only with new methods but also by demonstrating the impact of warm-start and restart. Finally, our forthcoming paper [69] also demonstrates how our analysis framework and gap reduction model extend to cover alternating direction optimization methods.

1.3. Paper organization. In sction 2, we propose a smoothing technique with proximity functions for (1)–(3) to estimate the primal-dual gap. We also investigate the properties of smoothed gap function and introduce the model-based gap reduction condition. Section 3 presents the first primal-dual algorithmic framework using accelerated (proximal-) gradient schemes for solving (1)–(3) and its convergence theory. Section 4 provides the second primal-dual algorithmic framework using averaging sequences for solving (1)–(3) and its convergence theory. Section 5 specifies different instances of our algorithmic framework for (1)–(3) under other common optimization structures and generalizes it to the cone constraint $Ax - c \in \mathcal{K}$. Numerical examples are presented in section 6. A comparison between our approach and existing methods is given in section 7. For clarity of exposition, technical proofs are moved to the appendix.

2. Smoothed gap function and optimality characterization. We propose to smooth the primal-dual gap function defined by (4) by proximity functions. Then we provide a key lemma to characterize the optimality condition for (1) and (2).

2.1. Basic notation. We use $\langle \cdot, \cdot \rangle$ for the standard inner product and $\|x\|_2$ for the Euclidean norm. Given a matrix $\mathbf{S}$, we define a seminorm of x as $\|x\|_\mathbf{S} := \sqrt{\langle \mathbf{S}x, \mathbf{S}x \rangle}$. When $\mathbf{S}$ is the identity matrix $\mathbb{I}$, we recover the standard Euclidean norm.

When $\mathbf{S}^\top \mathbf{S}$ is positive definite, the seminorm becomes a weighted norm. In this case, its dual norm exists and is defined by $\|u\|_{\mathcal{S},*} = \max\{\langle u, v\rangle \mid \|v\|_{\mathbf{S}} = 1\}$. When $\mathbf{S}^\top \mathbf{S}$ is not positive definite, we still consider the quantity $\|u\|_{\mathcal{S},*} = \max\{\langle u, v\rangle \mid \|v\|_{\mathbf{S}} = 1\}$, although $\|u\|_{\mathcal{S},*}$ is finite if and only if $u \in \mathrm{Ran}(\mathbf{S}^\top)$.

We also use $\|\cdot\|_{\mathcal{X}}$ (respectively, $\|\cdot\|_{\mathcal{Y}}$) and $\|\cdot\|_{\mathcal{X},*}$ (respectively, $\|\cdot\|_{\mathcal{Y},*}$) for the norm and the corresponding dual norm in the primal space $\mathcal{X}$ (respectively, the dual space $\mathcal{Y}$) induced by the above standard inner product in $\mathcal{X}$ (respectively, in $\mathcal{Y}$). Given a proper, closed and convex function f, we use $\mathrm{dom}\,(f)$ and $\partial f(x)$ to denote its domain and its subdifferential at x, respectively. If f is differentiable, then we use $\nabla f(x)$ for its gradient at x. For a given set $\mathcal{C}$, $\delta_{\mathcal{C}}(x) := 0$ if $x \in \mathcal{C}$, and $\delta_{\mathcal{C}}(x) := +\infty$ otherwise denotes the indicator function of $\mathcal{C}$. In addition, $\mathrm{ri}\,(\mathcal{C})$ denotes the relative interior of $\mathcal{C}$.

For a smooth function $f : \mathcal{Z} \to \mathbb{R}$, we say that f has the L_f-Lipschitz gradient with respect to the norm $\|\cdot\|_{\mathcal{Z}}$ if for any $z, \tilde{z} \in \mathrm{dom}\,(f)$, we have $\|\nabla f(z) - \nabla f(\tilde{z})\|_{\mathcal{Z},*} \leq L_f \|z - \tilde{z}\|_{\mathcal{Z}}$, where $L_f \equiv L(f) \in [0, \infty)$. We denote by $\mathcal{F}_{L_f}^{1,1}$ the class of all convex functions f with the L_f-Lipschitz gradient. We also use $\mu_f \equiv \mu(f)$ for the strong convexity parameter of a convex function f with respect to the seminorm $\|\cdot\|_{\mathcal{Z}}$, i.e., $f(\cdot) - (\mu_f/2)\|\cdot\|_{\mathcal{Z}}^2$ is convex. For a proper, closed and convex function f, we use prox_f to denote its proximal operator, which is defined as $\mathrm{prox}_f(z) := \mathrm{argmin}\{f(u) + (1/2)\|u - z\|_{\mathcal{Z}}^2 \mid u \in \mathrm{dom}\,(f)\}$.

2.2. Smooth proximity functions and Bregman distance. We use the following two mathematical tools in what follows.

2.2.1. Proximity functions. Given a nonempty, closed and convex set $\mathcal{Z}$ in the primal space or in the dual space, a continuous and μ_p-strongly convex function p is called a *proximity* function (or a prox-function) of $\mathcal{Z}$ if $\mathcal{Z} \subseteq \mathrm{dom}\,(p)$. We also denote

$$(6) \qquad \bar{z}^c := \mathrm{argmin}\{p(z) \mid z \in \mathrm{dom}\,(p)\} \quad \text{and} \quad D_{\mathcal{Z}} := \sup\{p(z) \mid z \in \mathcal{Z}\}$$

as the prox-center of p and the prox-diameter of $\mathcal{Z}$, respectively. Without loss of generality, we can assume that $\mu_p = 1$ and $p(\bar{z}^c) = 0$. Otherwise, we can shift and rescale the function p. Moreover, $D_{\mathcal{Z}} \geq 0$, and it is finite if $\mathcal{Z}$ is bounded.

In addition to the strong convexity, we also limit our class of prox-functions to the smooth ones, which have a Lipschitz gradient with the Lipschitz constant $L_p \geq 1$. We denote the class of prox-functions whose gradient has Lipschitz constant L by $\mathcal{S}_{L,1}^{1,1}$. For example, $p_{\mathcal{Z}}(z) := (1/2)\|z\|_{\mathbf{S}}^2$ is a simple prox-function in $\mathcal{Z} = \mathbb{R}^{n_z}$, i.e., $\frac{1}{2}\|\cdot\|_{\mathbf{S}}^2 \in \mathcal{S}_{\|\mathbf{S}\|^2,1}^{1,1}(\mathcal{Z})$.

2.2.2. Bregman distance. Instead of smoothing the primal and dual problems (1)–(2) by smooth proximity functions, we use a Bregman distance defined via $p_{\mathcal{Z}}$ as

$$(7) \qquad b_{\mathcal{Z}}(z, \dot{z}) := p_{\mathcal{Z}}(z) - p_{\mathcal{Z}}(\dot{z}) - \langle \nabla p_{\mathcal{Z}}(\dot{z}), z - \dot{z}\rangle \quad \forall z, \dot{z} \in \mathcal{Z},$$

where $p_{\mathcal{Z}} \in \mathcal{S}_{L,1}^{1,1}(\mathcal{Z})$. Clearly, if we fix $\dot{z} = \bar{z}^c$ at the center point of $p_{\mathcal{Z}}$, then $b_{\mathcal{Z}}(z, \bar{z}^c) = p_{\mathcal{Z}}(z)$. In addition, $\nabla_1 b_{\mathcal{Z}}(z, z) = 0$ for all $z \in \mathcal{Z}$. We use in what follows $\nabla b_{\mathcal{Z}}$ for $\nabla_1 b_{\mathcal{Z}}$.

2.3. Basic assumption. Our main assumption for problems (1)–(2) is to guarantee the strong duality, which essentially requires the following assumption (see [2, Proposition 15.22]).

ASSUMPTION A.1. *The solution set $\mathcal{X}^\star$ of the primal problem* (1) *(or* (3)*) is nonempty. In addition, the following assumption holds for either* (1) *or* (3)*:*

(a) *The condition $0 \in \mathrm{ri}\,(\mathrm{dom}\,(g) - A(\mathrm{dom}\,(f)))$ for* (1) *holds.*

(b) *The Slater condition* $\mathrm{ri}\,(\mathrm{dom}\,(f)) \cap \{x \in \mathbb{R}^n \mid Ax = c\} \neq \emptyset$ *for* (3) *holds.*

Now we define $\mathcal{X} := \overline{\mathrm{dom}\,(f)}$, $\mathcal{Y} := \overline{\mathrm{dom}\,(g^*)}$, and $\mathcal{W} := \mathcal{X} \times \mathcal{Y}$. Note that if the function g in (1) is Lipschitz continuous on $\mathcal{Y}$, then Assumption A.1 holds.

Under Assumption A.1, the strong duality for (1)–(2) holds; see, e.g., [2]. The solution set $\mathcal{Y}^\star$ of the dual problem (2) is nonempty, and

$$(8) \quad P^\star = f(x^\star) + g(Ax^\star) = D^\star = -f^*(-A^\top y^\star) - g^*(y^\star), \ \forall x^\star \in \mathcal{X}^\star, \ \forall y^\star \in \mathcal{Y}^\star.$$

Let $\mathcal{W}^\star := \mathcal{X}^\star \times \mathcal{Y}^\star$ be the primal-dual (or the saddle point) set of (1)–(2). Then (8) is equivalent to $f(x^\star) + g(Ax^\star) + f^*(-A^\top y^\star) + g^*(y^\star) = 0$ for all $(x^\star, y^\star) \in \mathcal{X}^\star \times \mathcal{Y}^\star$. In addition, we can write the optimality condition of (1)–(2) as follows:

$$(9) \quad -A^\top y^\star \in \partial f(x^\star) \ \text{and} \ y^\star \in \partial g(Ax^\star).$$

Note that this condition can be written as $0 \in \partial f(x^\star) + A^\top \partial g(Ax^\star)$ for the primal problem (1) and as $0 \in \partial g^*(y^\star) - A\partial f^*(-A^\top y^\star)$ for the dual problem (2).

2.4. Smoothed primal-dual gap function. The gap function G defined in (4) is *convex* but generally *nonsmooth*. This subsection introduces a smoothed primal-dual gap function that approximates G using smooth prox-functions.

2.4.1. The first smoothed approximation. Let $b_\mathcal{X}$ be a Bregman distance defined on $\mathcal{X}$ and $\dot{x} \in \mathcal{X}$ be given; we consider an approximation to the dual objective function $D(\cdot)$ as

$$(10) \quad D_\gamma(y; \dot{x}) := \min_{x \in \mathcal{X}} \{f(x) + \langle y, Ax \rangle + \gamma b_\mathcal{X}(x, \dot{x})\} - g^*(y) \equiv -f_\gamma^*(-A^\top y; \dot{x}) - g^*(y),$$

where $\gamma > 0$ is a dual smoothness parameter. The minimization subproblem in (10) always admits a solution, which is denoted by

$$(11) \quad x_\gamma^*(y; \dot{x}) := \underset{x \in \mathcal{X}}{\mathrm{argmin}} \{f(x) + \langle y, Ax \rangle + \gamma b_\mathcal{X}(x, \dot{x})\}.$$

We emphasize that our algorithms presented in the next sections support parallel and distributed computation for the *decomposable setting* of (1) or (3), where f is decomposed into N terms as $f(x) := \sum_{i=1}^{N} f_i(x_i)$ with the ith block being in $\mathbb{R}^{n_i}$ such that $\sum_{i=1}^{N} n_i = n$. In this case, we can choose a separable prox-function to generate a decomposable Bregman distance $b_\mathcal{X}(x, \dot{x}) := \sum_{i=1}^{N} b_i(x_i, \dot{x}_i)$ to approximate the dual function D defined in (2). By exploiting this decomposable structure, we can evaluate the smoothed dual function and its gradient in a parallel or distributed fashion. We will discuss the detail of this setting in section 5.

2.4.2. The second smoothed approximation. Let $b_\mathcal{Y}$ be a Bregman distance defined on $\mathcal{Y}$ the feasible set of the dual problem (2) and $\dot{y} \in \mathcal{Y}$. We consider an approximation to the objective $g(\cdot)$ in (1) as

$$(12) \quad g_\beta(u; \dot{y}) := \max_{y \in \mathcal{Y}} \{\langle u, y \rangle - g^*(y) - \beta b_\mathcal{Y}(y, \dot{y})\},$$

where $\beta > 0$ is a primal smoothness parameter. We also denote the solution of the maximization problem in (12) by $y_\beta^*(u; \dot{y})$, i.e.,

$$(13) \qquad y_\beta^*(u; \dot{y}) := \arg\max_{y \in \mathcal{Y}} \left\{ \langle u, y \rangle - g^*(y) - \beta b_{\mathcal{Y}}(y, \dot{y}) \right\}.$$

We consider an approximation to the primal objective function P as

$$(14) \qquad P_\beta(x; \dot{y}) := f(x) + g_\beta(Ax; \dot{y}).$$

This function is the second smoothed approximation for the primal problem. We note that if $g(\cdot) := \delta_{\{c\}}(\cdot)$ and $p_{\mathcal{Y}}(\cdot) := (1/2)\|\cdot\|_2^2$, then $y_\beta^*(u; \dot{y}) = \dot{y} + \beta^{-1}(u - c)$, which has a closed form.

2.5. Smoothed gap function and its properties. Given D_γ and P_β defined by (10) and (14), respectively, and the primal-dual variable $w := (x, y)$, the smoothed primal-dual gap (or the smoothed gap) $G_{\gamma\beta}$ is now defined as

$$(15) \qquad G_{\gamma\beta}(w; \dot{w}) := P_\beta(x; \dot{y}) - D_\gamma(y; \dot{x}),$$

where γ and β are two smoothness parameters and $\dot{w} := (\dot{x}, \dot{y})$.

The following lemma provides fundamental bounds of the objective residual $P(x) - P^\star$ for the unconstrained form (1) and the objective residual $f(x) - f^\star$ and the feasibility gap $\|Ax - c\|_{\mathcal{Y},*}$ for the constrained form (3). For clarity of exposition, we move its proof to Appendix A.2.

LEMMA 1. *Let $G_{\gamma\beta}$ be the smoothed gap function defined by (15) and $S_\beta(x; \dot{y}) := P_\beta(x; \dot{y}) - P^\star = f(x) + g_\beta(Ax; \dot{y}) - P^\star$ be the smoothed objective residual. Then we have*

$$(16) \qquad S_\beta(x; \dot{y}) \le G_{\gamma\beta}(w; \dot{w}) + \gamma b_{\mathcal{X}}(x^\star, \dot{x}) \quad and$$

$$\frac{1}{2}\|y_\beta^*(Ax; \dot{y}) - y^\star\|_{\mathcal{Y},*}^2 \le b_{\mathcal{Y}}(y^\star, \dot{y}) + \frac{1}{\beta}S_\beta(x; \dot{y}).$$

Suppose that $g(\cdot) := \delta_{\{c\}}(\cdot)$. Then, for any $y^\star \in \mathcal{Y}^\star$ and $x \in \mathcal{X}$, one has

$$(17) \qquad -\|y^\star\|_{\mathcal{Y}}\|Ax - c\|_{\mathcal{Y},*} \le f(x) - f^\star,$$

and the following primal objective residual and feasibility gap estimates hold for (3):

$$(18) \qquad \begin{cases} f(x) - f^\star & \le S_\beta(x; \dot{y}) - \langle y^\star, Ax - c \rangle + \beta b_{\mathcal{Y}}(y^\star, \dot{y}), \\ \|Ax - c\|_{\mathcal{Y},*} & \le \beta L_{b_{\mathcal{Y}}} \left[\|y^\star - \dot{y}\|_{\mathcal{Y}} + \left(\|y^\star - \dot{y}\|_{\mathcal{Y}}^2 + 2L_{b_{\mathcal{Y}}}^{-1} \beta^{-1} S_\beta(x; \dot{y}) \right)^{1/2} \right], \end{cases}$$

where the quantity in the square root is always nonnegative.

The estimates (17) and (18) are independent of optimization methods used to construct $\{\bar{w}^k\}$ for the primal-dual variable $w = (x, y)$. However, their convergence guarantee depends on the smoothness parameters γ_k and β_k. Hence, the convergence rate of the objective residual $f(\bar{x}^k) - f^\star$ and feasibility gap $\|A\bar{x}^k - c\|_{\mathcal{Y},*}$ depends on the rate of $\{(\gamma_k, \beta_k)\}$.

The first inequality in (16) is more precise than what [56] tells us. It holds even if $\mathcal{X}$ is unbounded and it shows that if $\dot{x}$ is close to $x^\star$, then the smoothed function is more accurate. The second inequality in (16) shows that the distance between $y_\beta^*(Ax; \dot{y})$ and $y^\star$ is controlled by quantities that will remain bounded. In practice, we observe that $y_\beta^*(Ax; \dot{y})$ cconverges to $y^\star$. Hence, restarting the algorithm with $\dot{y}' = y_\beta^*(Ax; \dot{y})$ gives us a chance to accelerate the actual performance of our algorithms while not hurting the convergence guarantee [30].

3. The accelerated primal-dual gap reduction algorithm. Our new scheme builds upon Nesterov's acceleration idea [53, 54]. At each iteration, we apply an accelerated proximal-gradient step to minimize $f + g_\beta$. Since $f + g_\beta$ is nonsmooth, we use the proximal operator of f to generate a proximal-gradient step. As a key feature, we must update the parameters τ_k and β_k simultaneously at each iteration with analytical updating formulas.

3.1. The method. Let $\bar{x}^k \in \mathcal{X}$ and $\tilde{x}^k \in \mathcal{X}$ be given. The *Accelerated Smoothed GAp ReDuction* (ASGARD) scheme generates a new point $(\bar{x}^{k+1}, \tilde{x}^{k+1})$ as

$$\text{(ASGARD)} \quad \begin{cases} \hat{x}^k & := (1 - \tau_k)\bar{x}^k + \tau_k\tilde{x}^k, \\ y^*_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) := \arg\max_{y \in \mathcal{Y}} \left\{ \langle A\hat{x}^k, y \rangle - g^*(y) - \beta_{k+1}b_{\mathcal{Y}}(y, \dot{y}) \right\}, \\ \bar{x}^{k+1} & := \text{prox}_{\beta_{k+1}\bar{L}_A^{-1}f} \left(\hat{x}^k - \beta_{k+1}\bar{L}_A^{-1}A^\top y^*_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) \right), \\ \tilde{x}^{k+1} & := \tilde{x}^k - \tau_k^{-1}(\hat{x}^k - \bar{x}^{k+1}), \end{cases}$$

where $\tau_k \in (0, 1]$ and $\beta_k > 0$ are parameters that will be defined in what follows. The constant $\bar{L}_A$ is defined as

$$\text{(19)} \qquad \bar{L}_A := \|A\|^2 = \max_{x \in \mathbb{R}^n} \left\{ \frac{\|Ax\|_{\mathcal{Y},*}^2}{\|x\|_{\mathcal{X}}^2} \right\}.$$

The ASGARD scheme requires a mirror step with the conjugate g^* of g to get $y^*_{\beta_{k+1}}(\cdot; \dot{y})$ in (13) and a proximal step of f in the third line.

The following lemma shows that $\bar{x}^{k+1}$ updated by ASGARD decreases the smoothed objective residual $P_{\beta_k}(\bar{x}^k; \dot{y}) - P^\star$, whose proof can be found in Appendix A.3.1.

LEMMA 2. *Let us choose $\tau_0 := 1$. If $\tau_k \in (0, 1)$ is the unique positive root of the cubic polynomial equation $p_3(\tau) := \tau^3/L_{b_{\mathcal{Y}}} + \tau^2 + \tau_{k-1}^2\tau - \tau_{k-1}^2 = 0$ for $k \geq 1$ and $\beta_k := \frac{\beta_{k-1}}{1 + \tau_{k-1}/L_{b_{\mathcal{Y}}}}$, then $\beta_k = \mathcal{O}\left(\frac{1}{k^{1/L_{b_{\mathcal{Y}}}}}\right)$ as $k \to \infty$ and*

$$\text{(20)} \qquad P_{\beta_{k+1}}(\bar{x}^{k+1}; \dot{y}) - P^\star + \frac{\tau_k^2}{\beta_{k+1}}\frac{\bar{L}_A}{2}\|\tilde{x}^{k+1} - x^\star\|_{\mathcal{X}}^2 \leq \frac{\tau_k^2}{\beta_{k+1}}\frac{\bar{L}_A}{2}\|\tilde{x}^0 - x^\star\|_{\mathcal{X}}^2.$$

Moreover, if $L_{b_{\mathcal{Y}}} = 1$, then $\frac{1}{k+1} \leq \tau_k \leq \frac{2}{k+2}$, $\frac{\tau_k^2}{\beta_{k+1}} \leq \frac{\tau_0^2}{\beta_1(k+1)} = \frac{1}{\beta_1(k+1)}$ and $\beta_k \leq \frac{2\beta_1}{k+1}$.

3.2. The primal-dual algorithmic template. Similar to the accelerated scheme [3, 53], we can eliminate $\tilde{x}^k$ in ASGARD by combining its first line and last line to obtain

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

Now we combine all the ingredients presented previously and this step to obtain a primal-dual algorithmic template for solving (1) as in Algorithm 1 below.

Per-iteration complexity of Algorithm 1. The computationally heavy steps of Algorithm 1 are Steps 4 and 5. The per-iteration complexity of Algorithm 1 consists of the following:

- One matrix-vector multiplication Ax and one mirror step of g^* at Step 4 to compute $y^*_{\beta_{k+1}}(A\hat{x}^k; \dot{y})$. If $g(\cdot) := \delta_{\{c\}}(\cdot)$ and $p_{\mathcal{Y}}(\cdot) := (1/2)\|\cdot\|_2^2$, then $y^*_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) = \dot{y} + \beta_{k+1}^{-1}(A\hat{x}^k - c)$, which only requires one matrix-vector multiplication Ax.

Algorithm 1. (*Accelerated Smoothed GAp ReDuction* (ASGARD) *algorithm*).

Initialization:

1: Choose $\beta_1 > 0$ (e.g., $\beta_1 := 0.5\sqrt{\bar{L}_A}$, where $\bar{L}_A$ is given in (19)), and set $\tau_0 := 1$.

2: Choose $\bar{x}^0 \in \mathcal{X}$ arbitrarily, and set $\hat{x}^0 := \bar{x}^0$.

For $k = 0$ to $k_{\max}$, perform:

3: Compute $\tau_{k+1} \in (0,1)$ the unique positive root of $\tau^3/L_{b_{\mathcal{Y}}} + \tau^2 + \tau_k^2\tau - \tau_k^2 = 0$.

4: Compute the dual step by solving

$$y^*_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) := \arg\max_{\hat{y} \in \mathcal{Y}} \left\{ \langle A\hat{x}^k, \hat{y} \rangle - g^*(\hat{y}) - \beta_{k+1}b_{\mathcal{Y}}(\hat{y}, \dot{y}) \right\}.$$

5: Compute the primal step $\bar{x}^{k+1}$ using the prox_f of f as

$$\bar{x}^{k+1} := \text{prox}_{\beta_{k+1}\bar{L}_A^{-1}f}\left(\hat{x}^k - \beta_{k+1}\bar{L}_A^{-1}A^\top y^*_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) \right).$$

6: Update $\hat{x}^{k+1} = \bar{x}^{k+1} + \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}(\bar{x}^{k+1} - \bar{x}^k)$ and $\beta_{k+2} := \frac{\beta_{k+1}}{1+L_{b_{\mathcal{Y}}}^{-1}\tau_{k+1}}$.

End for

- One adjoint matrix-vector multiplication $A^\top y$ and one proximal step of f at Step 5. If f is decomposable, evaluating prox_f can be implemented *in parallel*.

We note that if $p_{\mathcal{Y}}(\cdot) := (1/2)\|\cdot\|_2^2$, the the mirror step in g^* becomes a proximal step prox_{g^*}.

3.3. Convergence analysis. Our first main result is the following two theorems, which show an $\mathcal{O}(1/k)$ convergence rate of Algorithm 1 for both the unconstrained problem (1) and the constrained setting (3).

THEOREM 3. *Suppose that $g = \delta_{\{c\}}$. Let $\beta_1 > 0$ and $b_{\mathcal{Y}}$ be chosen such that $L_{b_{\mathcal{Y}}} = 1$. Let $\{\bar{x}^k\}$ be the primal sequence generated by Algorithm 1. Then the following bounds hold for (3):*

$$(21) \quad \begin{cases} f(\bar{x}^k) - f^\star & \geq -\|y^\star\|_{\mathcal{Y}}\|A\bar{x}^k - c\|_{\mathcal{Y},*}, \\ f(\bar{x}^k) - f^\star & \leq \frac{1}{k}\frac{\bar{L}_A}{2\beta_1}\|\bar{x}^0 - x^\star\|_{\mathcal{X}}^2 + \|y^\star\|_{\mathcal{Y}}\|A\bar{x}^k - c\|_{\mathcal{Y},*} + \frac{2\beta_1}{k+1}b_{\mathcal{Y}}(y^\star, \dot{y}), \\ \|A\bar{x}^k - c\|_{\mathcal{Y},*} & \leq \frac{\beta_1}{k+1}\left[\|y^\star - \dot{y}\|_{\mathcal{Y}} + \left(\|y^\star - \dot{y}\|_{\mathcal{Y}}^2 + \beta_1^{-2}\bar{L}_A\|\bar{x}^0 - x^\star\|_{\mathcal{X}}^2\right)^{1/2}\right]. \end{cases}$$

Proof. If $g = \delta_{\{c\}}$, then we apply Lemma 1 and use the bound on the smoothed optimality gap given by Lemma 2, noting that $\tilde{x}^0 = \bar{x}^0 = \hat{x}^0$, to get the bounds as in (18). Moreover, since $\beta_k \leq \frac{2\beta_1}{k+1}$ and

$$\frac{\tau_k^2}{\beta_{k+1}^2} = \frac{\tau_k^4}{\beta_{k+1}^2}\frac{1}{\tau_k^2} \leq \left(\frac{1}{\beta_1(k+1)}\right)^2(k+1)^2 = \frac{1}{\beta_1^2},$$

using these estimates into the resulting bounds, we obtain (21). $\qquad\square$

Note that if we choose $\dot{y} := \mathbf{0}^m$ and $b_{\mathcal{Y}}(y, \dot{y}) = \frac{1}{2}\|y - \dot{y}\|_{\mathcal{Y}}^2$, then the bounds (21) can be further simplified as

$$(22) \quad \begin{cases} |f(\bar{x}^k) - f^\star| \leq \frac{1}{k}\left(\frac{L_A}{2\beta_1}\|\bar{x}^0 - x^\star\|_{\mathcal{X}}^2 + 3\beta_1\|y^\star\|_{\mathcal{Y}}^2 + \frac{\sqrt{L_A}}{\beta_1}\|\bar{x}^0 - x^\star\|_{\mathcal{X}}\|y^\star\|_{\mathcal{Y}}\right), \\ \|A\bar{x}^k - c\|_{\mathcal{Y},*} \leq \frac{\beta_1}{k+1}\left(2\|y^\star\|_{\mathcal{Y}} + \frac{\sqrt{L_A}}{\beta_1}\|\bar{x}^0 - x^\star\|_{\mathcal{X}}\right). \end{cases}$$

Clearly, the choice of β_1 in Theorem 3 trades off between $\|\bar{x}^0 - x^\star\|_{\mathcal{X}}^2$ and $\|y^\star - \dot{y}\|_{\mathcal{Y}}^2$ on the primal objective residual $f(\bar{x}^k) - f^\star$ and on the feasibility gap $\|A\bar{x}^k - c\|_{\mathcal{Y},*}$.

THEOREM 4. *Suppose that g is Lipschitz continuous and $\dot{y} = \bar{y}^c$ ensures $D_{\mathcal{Y}} = \sup_{y \in \text{dom}(g^*)} b_{\mathcal{Y}}(y, \dot{y}) < +\infty$. Let $\beta_1 > 0$ and $b_{\mathcal{Y}}$ be chosen such that $L_{b_{\mathcal{Y}}} = 1$. Let $\{\bar{x}^k\}$ be the primal sequence generated by Algorithm 1. Then the primal objective residual of (1) satisfies*

$$(23) \qquad P(\bar{x}^k) - P^\star \leq \frac{\bar{L}_A}{2\beta_1 k}\|\bar{x}^0 - x^\star\|_{\mathcal{X}}^2 + \frac{2\beta_1}{k+1}D_{\mathcal{Y}} \ \ \text{for all } k \geq 1.$$

Proof. If g is Lipschitz continous, then, for all $x \in \mathcal{X}$, $\partial g(Ax) \neq \emptyset$ and $\text{dom}(g^*)$ is bounded. For $y \in \text{dom}(g^*)$, we have $b_{\mathcal{Y}}(y, \dot{y}) \leq D_{\mathcal{Y}} < +\infty$. Let $y^*(x) \in \partial g(Ax)$. Then we can show that

$$g(Ax) = \langle Ax, y^*(x) \rangle - g^*(y^*(x)) \leq \max_{y \in \mathcal{Y}} \{\langle Ax, y \rangle - g^*(y) - \beta b_{\mathcal{Y}}(y, \dot{y})\} + \beta b_{\mathcal{Y}}(y^*(x), \dot{y})$$

$$\leq g_\beta(Ax) + \beta b_{\mathcal{Y}}(y^*(x), \dot{y}) \leq g_\beta(Ax) + \beta D_{\mathcal{Y}}.$$

Therefore, the bound (23) follows directly from (20) and this inequality. $\square$

Remark. If, in addition to g^*, f also has a bounded domain, we recover the assumptions of [56]. By choosing $\bar{x}^c := \bar{x}^0$ and $\beta_1 := \sqrt{\bar{L}_A \frac{D_{\mathcal{X}}}{2D_{\mathcal{Y}}}}$, we get a convergence bound as $P(\bar{x}^k) - P^\star \leq \frac{2\sqrt{2}\sqrt{\bar{L}_A D_{\mathcal{X}} D_{\mathcal{Y}}}}{k}$. The worst-case convergence rate bound (23) is the same as the one in [56] (up to a small constant factor). However, β_1 does not depend on the tolerance ε as in [56].

Remark. When Algorithm 1 is applied to solve the constrained convex problem (3) using $p_{\mathcal{Y}}(\cdot) := \frac{1}{2}\|\cdot\|_2^2$, we can simplify the update rule for τ_k at Step 3 and β_k at Step 6 as follows:

$$(24) \qquad \beta_{k+1} := (1 - \tau_k)\beta_k \quad \text{and} \quad \tau_{k+1} := \frac{\tau_k}{\tau_k + 1} = \frac{1}{k+2}.$$

This update rule does not improve the worst-case convergence guarantee in Theorem 4, but it is simple. The detail analysis can be found in Appendix A.4.

4. The accelerated dual smoothed gap reduction method. Algorithm 1 can be viewed as an accelerated proximal scheme applying to minimize the function $P_\gamma(\cdot; \dot{y})$ defined in (14). Now we exploit the smoothed gap function $G_{\gamma\beta}$ defined by (15) to develop a novel primal-dual method for solving (1) and (2). Our goal is to design a new scheme to compute a primal-dual sequence $\{\bar{w}^k\}$ and a parameter sequence $\{(\gamma_k, \beta_k)\}$ such that $\max\{0, G_{\gamma_k\beta_k}(\bar{w}^k; \dot{w})\}$ converges to zero.

4.1. The method. Given $\bar{w}^k := (\bar{x}^k, \bar{y}^k) \in \mathcal{W}$, we derive a scheme to compute a new point $\bar{w}^{k+1} := (\bar{x}^{k+1}, \bar{y}^{k+1})$ as follows:

$$\text{(ADSGARD)} \qquad \begin{cases} \hat{y}^k & := (1 - \tau_k)\bar{y}^k + \tau_k y^*_{\beta_k}(A\bar{x}^k; \dot{y}), \\ \bar{y}^{k+1} & := \text{prox}_{\gamma_{k+1}\bar{L}_A^{-1}g^*}\left(\hat{y}^k + \gamma_{k+1}\bar{L}_A^{-1}Ax^*_{\gamma_{k+1}}(\hat{y}^k; \dot{x})\right), \\ \bar{x}^{k+1} & := (1 - \tau_k)\bar{x}^k + \tau_k x^*_{\gamma_{k+1}}(\hat{y}^k; \dot{x}), \end{cases}$$

where $\tau_k \in (0, 1)$ and the parameters $\beta_k > 0$ and $\gamma_{k+1} > 0$ will be updated in what follows. The points $x^*_{\gamma_{k+1}}(\hat{y}^k; \dot{x})$ and $y^*_{\beta_k}(A\bar{x}^k; \dot{y})$ are computed by (11) and (13),

respectively. This scheme requires one primal step for $x^*_{\gamma_{k+1}}(\hat{y}^k; \dot{x})$, one dual step for $y^*_{\beta_k}(A\bar{x}^k; \dot{y})$, and one dual proximal-gradient step for $\bar{y}^{k+1}$. Since the accelerated step is applied to g_γ, we call this scheme the *Accelerated Dual Smoothed GAp ReDuction* (ADSGARD) scheme.

The following lemma, whose proof is in Appendix A.5, shows that $\bar{w}^{k+1}$ updated by ADSGARD decreases the smoothed gap $G_{\gamma_k \beta_k}(\bar{w}^k)$ with at least a factor of $(1-\tau_k)$.

LEMMA 5. *Let $\bar{w}^{k+1} := (\bar{x}^{k+1}, \bar{y}^{k+1})$ be updated by the* (ADSGARD) *scheme. Then if $\tau_k \in (0, 1]$, β_k, and γ_k are chosen such that $\beta_1 \gamma_1 \geq \bar{L}_A$ and*

$$(25) \quad (1 + \tau_k/L_{b_\mathcal{X}})\gamma_{k+1} \geq \gamma_k, \qquad \beta_{k+1} \geq (1-\tau_k)\beta_k, \quad and \quad \frac{\bar{L}_A}{\gamma_{k+1}} \leq \frac{(1-\tau_k)\beta_k}{\tau_k^2},$$

then $\bar{w}^{k+1} \in \mathcal{W}$ and satisfies $G_{\gamma_{k+1}\beta_{k+1}}(\bar{w}^{k+1}; \dot{w}) \leq (1-\tau_k)G_{\gamma_k\beta_k}(\bar{w}^k; \dot{w}) \leq 0$.

Let $\tau_0 := 1$. Then for all $k \geq 1$, if we choose $\tau_k \in (0, 1)$ to be the unique positive solution of the cubic equation $p_3(\tau) := \tau^3/L_{b_\mathcal{X}} + \tau^2 + \tau_{k-1}^2\tau - \tau_{k-1}^2 = 0$, then $\frac{1}{k+1} \leq \tau_k \leq \frac{2}{k+2}$ for $k \geq 1$. The parameters β_k and γ_k computed by $\beta_1\gamma_1 = \bar{L}_A$ and

$$(26) \qquad \gamma_{k+1} := \frac{\gamma_k}{1 + \tau_k/L_{b_\mathcal{X}}} \quad and \quad \beta_{k+1} := (1 - \tau_k)\beta_k$$

satisfy the conditions in (25).

In addition, if $L_{b_\mathcal{X}} = 1$, then $\gamma_k \leq \frac{2\gamma_1}{k+1}$ and $\frac{\bar{L}_A}{2\gamma_1(k+1)} \leq \beta_{k+1} \leq \frac{\beta_1}{k+1}$ for $k \geq 1$.

4.2. The primal-dual algorithmic template. We combine all the ingredients presented in the previous subsection to obtain a primal-dual algorithmic template for solving either (1) or (3) as shown in Algorithm 2.

Algorithm 2. (*Accelerated Dual Smoothed GAp ReDuction* (ADSGARD)).

Initialization:
1: Choose $\gamma_1 > 0$ (e.g., $\gamma_1 := \sqrt{\bar{L}_A}$, where $\bar{L}_A$ is given by (19)). Set $\beta_1 := \frac{\bar{L}_A}{\gamma_1}$ and $\tau_0 := 1$.
2: Take an initial point $\bar{y}^*_0 := \dot{y} \in \mathcal{Y}$.
For $k = 0$ to $k_{\max}$, perform:
3: Update $\hat{y}^k := (1 - \tau_k)\bar{y}^k + \tau_k\bar{y}^*_k$.
4: Compute $\hat{x}^*_{k+1}$ *in parallel* with

$$\hat{x}^*_{k+1} := \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ f(x) + \langle A^\top \hat{y}^k, x \rangle + \gamma_{k+1}b_\mathcal{X}(x, \dot{x}) \right\}.$$

5: Update the dual vector

$$\bar{y}^{k+1} := \operatorname{prox}_{\gamma_{k+1}\bar{L}_A^{-1}g^*} \left(\hat{y}^k + \gamma_{k+1}\bar{L}_A^{-1}A\hat{x}^*_{k+1} \right).$$

6: Update the primal vector $\bar{x}^{k+1} := (1 - \tau_k)\bar{x}^k + \tau_k\hat{x}^*_{k+1}$.
7: Compute

$$\bar{y}^*_{k+1} := \arg\max_{y \in \mathcal{Y}} \left\{ \langle A\bar{x}^{k+1}, y \rangle - g^*(y) - \beta_{k+1}b_\mathcal{Y}(y, \dot{y}) \right\}.$$

8: Compute $\tau_{k+1} \in (0, 1)$ the unique positive root of $\tau^3/L_{b_\mathcal{Y}} + \tau^2 + \tau_k^2\tau - \tau_k^2 = 0$.
9: Update $\gamma_{k+2} := \frac{\gamma_{k+1}}{1 + L_{b_\mathcal{X}}^{-1}\tau_{k+1}}$ and $\beta_{k+2} := (1 - \tau_{k+1})\beta_{k+1}$.
End for

Since $\tau_0 = 1$, Step 3 shows that $\hat{y}^0 = \bar{y}_0^*$, while Step 6 leads to $\bar{x}^1 = \hat{x}_1^*$. The main steps of Algorithm 2 are Steps 4, 5, and 7, where we need to solve the subproblem (11) and to update two dual steps, respectively. The first dual step requires the proximal operator $\mathrm{prox}_{\rho g^*}$ of g^*, while the second one computes $\bar{y}_{k+1}^* = y_{\beta_{k+1}}^*(A\bar{x}^{k+1}; \dot{y})$.

When $g = \delta_{\{c\}}$, the indicator of $\{c\}$ in the constrained problem (3), we have

$$y_{\beta_k}^*(A\bar{x}^k; \dot{y}) = \nabla b_{\mathcal{Y}}^* \left(\beta_k^{-1}(A\bar{x}^k - c), \dot{y} \right) \quad \text{and} \quad \bar{y}^{k+1} := \hat{y}^k + \gamma_{k+1} \left(Ax_{\gamma_{k+1}}^*(\hat{y}^k; \dot{x}) - c \right).$$

The first dual step only requires one matrix-vector multiplication Ax. Clearly, by Step 6, it follows that $A\bar{x}^{k+1} - c = (1 - \tau_k)(A\bar{x}^k - c) + \tau_k(A\hat{x}_{k+1}^* - c)$, and by Step 7, we have $\bar{y}_k^* = y_{\beta_k}^*(A\bar{x}^k; \dot{y}) = \nabla b_{\mathcal{Y}}^* \left(\beta_k^{-1}(A\bar{x}^k - c), \dot{y} \right)$, which is equivalent to $A\bar{x}^k - c = \beta_k \nabla b_{\mathcal{Y}}(\bar{y}_k^*, \dot{y})$. Hence, $A\bar{x}^{k+1} - c = (1 - \tau_k)\beta_k \nabla b_{\mathcal{Y}}(\bar{y}_k^*, \dot{y}) + \frac{\tau_k}{\gamma_{k+1}}(\bar{y}^{k+1} - \hat{y}^k)$ due to Step 5. Finally, we can derive an update rule for $\bar{y}_{k+1}^*$ as

$$(27) \qquad \bar{y}_{k+1}^* := \nabla b_{\mathcal{Y}}^* \left(\beta_{k+1}^{-1} \left((1 - \tau_k)\beta_k \nabla b_{\mathcal{Y}}(\bar{y}_k^*, \dot{y}) + \frac{\tau_k}{\gamma_{k+1}}(\bar{y}^{k+1} - \hat{y}^k) \right), \dot{y} \right).$$

Per-iteration complexity of Algorithm 2. From the above analysis, we can conclude that the per-iteration complexity of Algorithm 1 consists of the following:

- One adjoint matrix-vector multiplication $A^\top y$ and one mirror step in f at Step 4 to compute $\hat{x}_{k+1}^*$. If f is decomposable, then Step 4 can be implemented *in parallel*.
- One matrix-vector multiplication Ax, one proximal step of g^* at Step 5 to compute $\bar{y}^{k+1}$, and one mirror step of g^* at Step 7 to compute $\bar{y}_{k+1}^*$. If $g = \delta_{\{c\}}$ and $p_{\mathcal{Y}}(\cdot) := (1/2)\| \cdot \|_2^2$, then computing $\bar{y}^{k+1}$ using (27) requires only one Ax.

4.3. Convergence analysis. The following theorem shows the convergence of Algorithm 2. For the constrained setting (3), we still have the lower bound on $f(\bar{x}^k) - f^\star$ as in Theorem 4; i.e., $-\|y^\star\|_{\mathcal{Y}}\|A\bar{x}^k - c\|_{\mathcal{Y},*} \leq f(\bar{x}^k) - f^\star$ for any $\bar{x}^k \in \mathcal{X}$ and $y^\star \in \mathcal{Y}^\star$.

THEOREM 6. *Suppose that $g = \delta_{\{c\}}$. Let $b_{\mathcal{X}}$ be chosen such that $L_{b_{\mathcal{X}}} = 1$ and $\{\bar{w}^k\}$ be the sequence generated by Algorithm 2 for solving (3), where $\gamma_1 > 0$ is given. Then the following bounds for (3) hold:*

$$(28) \quad \begin{cases} f(\bar{x}^k) - f^\star & \geq -\|y^\star\|_{\mathcal{Y}}\|A\bar{x}^k - c\|_{\mathcal{Y},*}, \\ f(\bar{x}^k) - f^\star & \leq \frac{2\gamma_1}{k+1} b_{\mathcal{X}}(x^\star, \dot{x}) + \frac{\bar{L}_A}{\gamma_1 k} b_{\mathcal{Y}}(y^\star, \dot{y}) + \|y^\star\|_{\mathcal{Y}}\|A\bar{x}^k - c\|_{\mathcal{Y},*}, \\ \|A\bar{x}^k - c\|_{\mathcal{Y},*} & \leq \frac{\bar{L}_A}{\gamma_1 k} L_{b_{\mathcal{Y}}} \left[\|y^\star - \dot{y}\|_{\mathcal{Y}} + \left(\|y^\star - \dot{y}\|_{\mathcal{Y}}^2 + \frac{8\gamma_1^2}{L_{b_{\mathcal{Y}}}} b_{\mathcal{X}}(x^\star, \dot{x}) \right)^{1/2} \right]. \end{cases}$$

Proof. This set of inequalities is a consequence of Lemmas 1 and 5 using $\beta_k \leq \frac{\beta_1}{k}$, $\gamma_k \leq \frac{2\gamma_1}{k+1}$, and $\frac{\gamma_k}{\beta_k} \leq \frac{4\gamma_1^2 k}{k+1} \leq 4\gamma_1^2$. $\qquad\square$

THEOREM 7. *Suppose that g is Lipschitz continuous as in Theorem 4. Let $b_{\mathcal{X}}$ be chosen such that $L_{b_{\mathcal{X}}} = 1$ and $\{\bar{w}^k\}$ be the sequence generated by Algorithm 2 for solving (1), where $\gamma_1 > 0$ is given. Then the following convergence bound holds:*

$$(29) \qquad P(\bar{x}^k) - P^\star \leq \frac{2\gamma_1}{k+1} b_{\mathcal{X}}(x^\star, \dot{x}) + \frac{2\bar{L}_A}{\gamma_1 k} D_{\mathcal{Y}}.$$

Proof. Since $S_\beta(x; \dot{y}) \leq G_{\gamma\beta}(w; \dot{w}) + \gamma b_{\mathcal{X}}(x^\star, \dot{x})$, using Lemma 5 we can show that $S_{\beta_k}(\bar{x}^k; \dot{y}) \leq G_{\gamma_k \beta_k}(\bar{w}^k; \dot{w}) + \gamma_k b_{\mathcal{X}}(x^\star, \dot{x}) \leq \gamma_k b_{\mathcal{X}}(x^\star, \dot{x})$. Similar to the proof of Theorem 4, we obtain the bound (29) for the objective residual of (1). $\qquad\square$

Similar to Theorem 4, we can simplify the bound (28) to obtain a simple bound as in (22), where we omit the details here. The choice of γ_1 and β_1 in Theorem 7 also trades off the primal objective residual and the primal feasibility gap.

4.4. The choice of smoothers. For this algorithm, one needs to choose a norm $\|\cdot\|_{\mathcal{X}} = \|\cdot\|_{\mathbf{S}}$ and a smoother $p_{\mathcal{X}}$ such that $p_{\mathcal{X}}$ is strongly convex with respect to the norm $\|\cdot\|_{\mathbf{S}}$. One possibility is to choose $\|\cdot\|_{\mathbf{S}}$ in order to have a simple formula for $\hat{x}^*_{k+1} = x^*_\gamma(\hat{y}^k; \dot{x})$. A classical choice is a diagonal $\mathbf{S}$, and $b_{\mathcal{X}}(\cdot, \dot{x}) = \frac{1}{2}\|\cdot - \dot{x}\|^2_{\mathbf{S}}$ is a quadratic function for a given $\dot{x} \in \mathcal{X}$.

If f is decomposable as $f(x) = \sum_{i=1}^N f_i(x_i)$ and we choose $b_{\mathcal{X}}(x, \dot{x}) := \sum_{i=1}^N b_{\mathcal{X}_i}(x_i, \dot{x}_i)$, then the computation of $\hat{x}^*_{k+1}$ at Step 4 of Algorithm 2 can be carried out in *parallel*.

Another possibility is to choose $\mathbf{S} = A$ and $p_{\mathcal{X}}(\cdot) = \frac{1}{2}\|\cdot\|^2_{\mathbf{S}}$. In that case, the computation of x^*_{k+1} may require an iterative subsolver, but we are allowed to take $\dot{x} = x^\star$. Indeed, as $Ax^\star = c$, we have that for all x, $b_{\mathcal{X}}(x, x^\star) = \frac{1}{2}\|x - x^\star\|^2_A = \frac{1}{2}(Ax - c)^\top(Ax - c)$. Hence, we can consider $x^\star$ as a center even though we do not know it. We shall develop the consequences of such a choice in the section 5.1.

5. Special instances of the primal-dual gap reduction framework. We specify our ADSGARD scheme to handle two special cases: augmented Lagrangian method and strongly convex objective. Then we provide an extension of our algorithms to a general cone constraint.

5.1. Accelerated smoothing augmented Lagrangian gap reduction method. The augmented Lagrangian method is a classical optimization technique and has widely been used in various applications due to its emergingly practical performance. In this section, we customize Algorithm 2 using ADSGARD to solve the constrained convex problem (3). The inexact variant of this algorithm can be found in our early technical report [68, section 5.3].

The augmented Lagrangian smoother. We choose here $p_{\mathcal{X}} = \|\cdot\|^2_{\mathcal{X}} = \|\cdot\|^2_A$, $p_{\mathcal{Y}}(\cdot) = \|\cdot\|^2_{\mathcal{Y}} = \|\cdot\|^2_{\mathbb{I}}$ and $\dot{x} = x^\star$ and $b_{\mathcal{X}}(x, \dot{x}) := (1/2)\|A(x - x^\star)\|^2_{\mathcal{Y},*} = (1/2)\|Ax - c\|^2_{\mathcal{Y},*}$. This is indeed the augmented term for the Lagrange function of (3). Note that even though $\dot{x}$ is unknown, $b_{\mathcal{X}}(x, \dot{x})$ can be computed easily using the equality $Ax^\star = c$.

We specify the primal-dual ADSGARD scheme with the augmented Lagrangian smoother for fixed $\gamma_{k+1} = \gamma_0 > 0$ as follows:

$$(\text{ASALGARD}) \quad \begin{cases} \hat{y}^k & := (1 - \tau_k)\bar{y}^k + \tau_k y^*_{\beta_k}(A\bar{x}^k; \dot{y}), \\ \hat{x}^*_{\gamma_0}(\hat{y}^k) & := \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left\{ f(x) + \langle \hat{y}^k, Ax - c \rangle + \dfrac{\gamma_0}{2}\|Ax - c\|^2_{\mathcal{Y},*} \right\}, \\ \bar{y}^{k+1} & := \hat{y}^k + \gamma_0(A\hat{x}^*_{\gamma_0}(\hat{y}^k) - c), \\ \bar{x}^{k+1} & := (1 - \tau_k)\bar{x}^k + \tau_k \hat{x}^*_{\gamma_0}(\hat{y}^k), \end{cases}$$

where $\tau_k \in (0, 1)$, $\gamma_0 > 0$ is the penalty (or the primal smoothness) parameter, and β_k is the dual smoothness parameter. As a result, this method is called the *Accelerated Smoothing Augmented Lagrangian GAp ReDuction* (ASALGARD) scheme.

This scheme consists of two dual steps at lines 1 and 3. However, we can combine these steps as in (27) so that it requires only one matrix-vector multiplication Ax. Consequently, the per-iteration complexity of ASALGARD remains essentially the same as the standard augmented Lagrangian method [9].

The update rule for parameters. In our augmented Lagrangian method, we only need to update τ_k and β_k such that $\beta_{k+1} \geq (1 - \tau_k)\beta_k$ and $\gamma_0\beta_k(1 - \tau_k) \geq \tau_k^2$. Using the equality in these conditions and defining $\tau_k := t_k^{-1}$, we can derive

Algorithm 3. (*Accelerated Smoothing Augmented Lagrangian GAp ReDuction* (ASALGARD)).

Initialization:

1: Choose an initial value $\gamma_0 > 0$ and $\beta_0 := 1$. Set $t_0 := 1$ and $\beta_1 := \gamma_0^{-1}$.

2: Choose an initial point $(\bar{x}^0, \bar{y}^0) \in \mathcal{W}$.

For $k = 0$ **to** $k_{\max}$, **perform:**

3: Update $y^*_{\beta_k}(\bar{x}^k; \dot{y}) := \nabla b^*_{\mathcal{Y}}\left(\beta_k^{-1}(A\bar{x}^k - c), \dot{y}\right)$.

4: Update

$$(31) \qquad \hat{x}^*_{\gamma_0}(\hat{y}^k) := \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ f(x) + \langle \hat{y}^k, Ax - c \rangle + \frac{\gamma_0}{2} \|Ax - c\|^2_{\mathcal{Y},*} \right\}.$$

5: Update $\bar{y}^{k+1} := \hat{y}^k + \gamma_0(A\hat{x}^*_{\gamma_0}(\hat{y}^k) - c)$ and $\bar{x}^{k+1} := (1 - t_k^{-1})\bar{x}^k + t_k^{-1}\hat{x}^*_{\gamma_0}(\hat{y}^k)$.

6: Update $t_{k+1} := 0.5\left(1 + \sqrt{1 + 4t_k^2}\right)$ and $\beta_{k+2} := (t_{k+1} - 1)t_{k+1}^{-1}\beta_{k+1}$.

End for

$$(30) \qquad t_{k+1} := \frac{1}{2}\left(1 + \sqrt{1 + 4t_k^2}\right) \quad \text{and} \quad \beta_{k+1} := \frac{(t_k - 1)}{t_k}\beta_k.$$

Here we fix $\beta_1 > 0$ and choose $t_0 := 1$.

The algorithm template. We modify Algorithm 2 to obtain the following augmented Lagrangian variant, Algorithm 3.

The main step of Algorithm 3 is the solution of the primal convex subproblem (31). In general, solving this subproblem remains challenging due to the nonseparability of the quadratic term $\|Ax - c\|^2_{\mathcal{Y},*}$. We can numerically solve it by using either alternating direction optimization methods or other first-order methods. The convergence analysis of inexact augmented Lagrangian methods can be found in [51].

Convergence guarantee. The following proposition shows the convergence of Algorithm 3, whose proof is moved to Appendix A.6.

PROPOSITION 8. *Let* $\{\bar{w}^k\}$ *be the sequence generated by Algorithm* 3. *Then we have*

$$(32) \qquad \begin{cases} -\frac{8L_{b_{\mathcal{Y}}}\|y^\star\|_{\mathcal{Y}}\|y^\star - \dot{y}\|_{\mathcal{Y}}}{\gamma_0(k+2)^2} \leq f(\bar{x}^k) - f^\star \leq \frac{8L_{b_{\mathcal{Y}}}\|y^\star\|_{\mathcal{Y}}\|y^\star - \dot{y}\|_{\mathcal{Y}} + 4b_{\mathcal{Y}}(y^\star, \dot{y})}{\gamma_0(k+2)^2}, \\ \|A\bar{x}^k - c\|_{\mathcal{Y},*} \leq \frac{8L_{b_{\mathcal{Y}}}\|y^\star - \dot{y}\|_{\mathcal{Y}}}{\gamma_0(k+2)^2}. \end{cases}$$

As a consequence, the worst-case iteration-complexity of Algorithm 3 *to achieve an* ε*-primal solution* $\bar{x}^k$ *for* (3) *is* $\mathcal{O}(\sqrt{\frac{b_{\mathcal{Y}}(y^\star, \dot{y})}{\gamma_0\varepsilon}})$.

The estimate (32) guides us to choose a large value for γ_0 such that we obtain better convergence bounds. However, if γ_0 is too large, then the complexity of solving the subproblem (31) increases commensurately. In practice, γ_0 is often updated using a heuristic strategy [9, 11]. In general settings, since the solution $\hat{x}^*_{k+1}$ computed by (31) requires solving a generic convex problem, it no longer has a closed form expression.

5.2. The strongly convex objective case. If the objective function f of (1) is strongly convex with the convexity parameter $\mu_f > 0$, then it is well known [56] that its conjugate f^* is smooth and its gradient $\nabla f^*(\cdot) := x^*(\cdot)$ is Lipschitz continuous with the Lipschitz constant $L_{f^*} := \mu_f^{-1}$, where $x^*(\cdot)$ is given by

$$(33) \qquad x^*(u) := \arg\max_{x \in \mathcal{X}} \left\{ \langle u, x \rangle - f(x) \right\}.$$

In addition, if $f_A^*(\cdot) := f^*(-A^\top(\cdot))$, then ∇f_A is Lipschitz continuous with $L_{f_A^*} := \frac{\bar{L}_A}{\mu_f}$.

The primal-dual update scheme. In this subsection, we only illustrate the modification of ADSGARD to solve the strongly convex primal problem (1) as

$$(\text{ADSGARD}_\mu) \qquad \begin{cases} \hat{y}^k & := (1-\tau_k)\bar{y}^k + \tau_k y^*_{\beta_k}(A\bar{x}^k; \dot{y}) \\ \bar{x}^{k+1} & := (1-\tau_k)\bar{x}^k + \tau_k x^*(-A^\top \hat{y}^k) \\ \bar{y}^{k+1} & := \text{prox}_{L_{f_A^*}^{-1} g^*} \left(\hat{y}^k + L_{f_A^*}^{-1} A x^*(-A^\top \hat{y}^k) \right). \end{cases}$$

We note that we no longer have the dual smoothness parameter γ_k, which is fixed to $\mu_f > 0$. Hence, the conditions (25) of Lemma 5 reduce to $\beta_{k+1} \geq (1-\tau_k)\beta_k$ and $(1-\tau_k)\beta_k \geq L_{f_A^*}\tau_k^2$. From these conditions we can derive the update rule for τ_k and β_k as in Algorithm 3, which is

$$(34) \qquad t_{k+1} := \frac{1}{2}\left(1 + \sqrt{1+4t_k^2}\right), \qquad \beta_{k+1} := \frac{(t_k - 1)}{t_k}\beta_k, \quad \text{and} \quad \tau_k := t_k^{-1}.$$

Here we fix $\beta_1 := L_{f_A^*} = \frac{\|A\|^2}{\mu_f}$ and choose $t_0 := 1$.

Convergence guarantee. The following proposition shows the convergence of ADSGARD_μ, whose proof is in Appendix A.7.

PROPOSITION 9. *Suppose that the objective f of the constrained convex problem* (3) *is strongly convex with the convexity parameter $\mu_f > 0$. Let $\{\bar{w}^k\}$ be generated by* (ADSGARD_μ) *using the update rule* (34). *Then the following guarantees hold:*

$$(35) \qquad \begin{cases} -\frac{8L_{b_\mathcal{Y}}\bar{L}_A\|y^\star\|_\mathcal{Y}\|y^\star - \dot{y}\|_\mathcal{Y}}{\mu_f(k+2)^2} \leq f(\bar{x}^k) - f^\star \leq \frac{8L_{b_\mathcal{Y}}\bar{L}_A\|y^\star\|_\mathcal{Y}\|y^\star - \dot{y}\|_\mathcal{Y} + 4b_\mathcal{Y}(y^\star; \dot{y})}{\mu_f(k+2)^2}, \\ \|A\bar{x}^k - c\|_{\mathcal{Y},*} \leq \frac{8L_{b_\mathcal{Y}}\bar{L}_A\|y^\star - \dot{y}\|_\mathcal{Y}}{\mu_f(k+2)^2}. \end{cases}$$

This result shows that ADSGARD_μ has an $\mathcal{O}(1/k^2)$ convergence rate with respect to the objective residual and the feasibility gap. We note that in both Propositions 8 and 9, the bounds only depend on the quantities in the dual space and $\bar{L}_A$.

5.3. Extension to general cone constraints. The theory presented in the previous sections can be extended to solve the following general constrained convex optimization problem:

$$(36) \qquad f^\star := \min_{x \in \mathcal{X}} \left\{ f(x) \mid Ax - c \in \mathcal{K} \right\},$$

where f, A, and c are defined as in (3) and $\mathcal{K}$ is a nonempty, closed and convex set in $\mathbb{R}^m$.

If $\mathcal{K}$ is a nonempty, closed and convex set, then a simple way to process (36) is using a slack variable $r \in \mathcal{K}$ such that $r := Ax - c$ and $z := (x, r)$ as a new variable. Then we can transform (36) into (3) with respect to the new variable z. The primal subproblem corresponding to r is defined as $\min \left\{ \langle -y, r \rangle \mid r \in \mathcal{K} \right\}$, which is equivalent to the support function $s_\mathcal{K}(y) := \sup \left\{ \langle y, r \rangle \mid r \in \mathcal{K} \right\}$ of $\mathcal{K}$. Consequently, the dual function becomes $\tilde{g}(y) := g(y) - s_\mathcal{K}(y)$, where $g(y) := \min \left\{ f(x) + \langle Ax - c, y \rangle \mid x \in \mathcal{X} \right\}$. Now we can apply the algorithms presented in the previous sections to obtain an approximate solution $\bar{z}^k := (\bar{x}^k, \bar{r}^k)$ with a convergence guarantee on $f(\bar{x}^k) - f^\star$, $\|A\bar{x}^k - \bar{r}^k - c\|_{\mathcal{Y},*}$, $\bar{x}^k \in \mathcal{X}$ and $\bar{r}^k \in \mathcal{K}$, as in Theorem 3 or Theorem 6.

If $\mathcal{K}$ is a cone (e.g., $\mathcal{K} := \mathbb{R}_+^m$, $\mathcal{K} := \mathcal{L}_+^m$ is a second-order cone or $\mathcal{K} := \mathcal{S}_+^m$ is a semidefinite cone), then with the choice $p_{\mathcal{Y}}(\cdot) := (1/2)\|\cdot\|_{\mathcal{Y}}^2$, we can substitute the smoothed function g_β in (12) to obtain the following one:

$$(37) \qquad \hat{g}_\beta(Ax, \dot{y}) := \max\left\{\langle Ax - c, y\rangle - (\beta/2)\|y - \dot{y}\|_{\mathcal{Y}}^2 \mid y \in -\mathcal{K}^*\right\},$$

where $\mathcal{K}^*$ is the dual cone of $\mathcal{K}$, which is defined as $\mathcal{K}^* := \{z \mid \langle z, x\rangle \geq 0,\ x \in \mathcal{K}\}$. With this definition, we use the smoothed gap function $\hat{G}_{\gamma\beta}$ as $\hat{G}_{\gamma\beta}(w; \dot{w}) := \hat{P}_\beta(x; \dot{y}) - D_\gamma(y; \dot{x})$, where $D_\gamma(y; \dot{x}) := \min\{f(x) + \langle Ax - c, y\rangle + \gamma b_{\mathcal{X}}(x, \dot{x}) \mid x \in \mathcal{X}\}$ is the smoothed dual function defined as before and $\hat{P}_\beta(x; \dot{y}) := f(x) + \hat{g}_\beta(Ax, \dot{y})$.

In principle, we can apply one of the two previous schemes to solve (36). Let us demonstrate the ADSGARD for this case. Since $\mathcal{K}$ is a cone, we remain using the original scheme (ADSGARD) with the following changes:

$$\begin{cases} y_{\beta_k}^*(A\bar{x}^k; \dot{y}) & := \mathrm{proj}_{-\mathcal{K}^*}\left(\dot{y} + \beta_k^{-1}(A\bar{x}^k - c)\right), \\ \bar{y}^{k+1} & := \mathrm{proj}_{-\mathcal{K}^*}\left(\hat{y}^k + \frac{\gamma_{k+1}}{L_A}\left(Ax_{\gamma_{k+1}}^*(\hat{y}^k) - c\right)\right), \end{cases}$$

where $\mathrm{proj}_{-\mathcal{K}^*}$ is the projection onto the anti-cone $-\mathcal{K}^*$. In this case, we still have the convergence guarantee as in Theorem 7 for the objective residual $f(\bar{x}^k) - f^\star$ and the primal feasibility gap $\mathrm{dist}\left(A\bar{x}^k - c, \mathcal{K}\right)$, the Euclidean distance from $A\bar{x}^k - c$ to $\mathcal{K}$. We note that if $\mathcal{K}$ is a self-dual conic cone, then $\mathcal{K}^* = \mathcal{K}$. Hence, $y_{\beta_k}^*(A\bar{x}^k; \dot{y})$ and $\bar{y}^{k+1}$ can be either efficiently computed or a closed form.

5.4. Restarting techniques. Similar to other accelerated gradient algorithms in [31, 59, 66], restarting ASGARD and ADSGARD may lead to a better performance in practice. We discuss in this subsection how to restart these two algorithms using a fixed iteration restarting strategy [59].

If we consider ASGARD, then, when a restart takes place, we perform the following steps:

$$(38) \qquad \begin{cases} \tilde{x}^{k+1} & \leftarrow & \bar{x}^{k+1}, \\ \dot{y} & \leftarrow & y_{\beta_{k+1}}^*(A\bar{x}^{k+1}; \dot{y}), \\ \beta_{k+1} & \leftarrow & \beta_1, \\ \tau_{k+1} & \leftarrow & 1. \end{cases}$$

Restarting the primal variable at $\bar{x}^{k+1}$ is classical; see, e.g., [59]. For the dual center point $\dot{y}$, we suggest restarting it at the last dual variable computed. Indeed, by (16), we know that the distance between $y_{\beta_{k+1}}^*(A\bar{x}^{k+1}; \dot{y})$ and the optimal solution $y^\star$ will remain bounded. Hence, in the favorable cases, we will benefit from a smaller distance between the new center point and $y^\star$, while in the unfavorable cases, restarting should not affect too much the convergence. In practice, however, we observe that $y_{\beta_{k+1}}^*(A\bar{x}^{k+1}; \dot{y})$ converges to the dual solution $y^\star$. We note that the restarting strategy (38) does not increase the per-iteration complexity of the algorithm.

For ADSGARD, we suggest restarting it using the following steps:

$$(39) \qquad \begin{cases} \hat{y}^{k+1} & \leftarrow & \bar{y}^{k+1}, \\ \dot{y} & \leftarrow & \bar{y}^{k+1}, \\ \dot{x} & \leftarrow & x_{\gamma_{k+1}}^*(\hat{y}^k; \dot{x}), \\ \beta_{k+1} & \leftarrow & \beta_1, \\ \gamma_{k+1} & \leftarrow & \gamma_1, \\ \tau_{k+1} & \leftarrow & 1. \end{cases}$$

Understanding the actual consequences of the restart procedure as well as designing other conditions for restarting are still open questions, even for the unconstrained case. Yet we observe that it often significantly improves the convergence speed in practice.

6. Numerical experiments. In this section, we provide some key examples to illustrate the advantages of our new algorithms compared to existing states of the art. While other numerical experiments can be found in our technical reports [68], we instead focus on some extreme cases where existing methods may encounter an arbitrarily slow convergence rate due to lack of theory, while our methods exhibits an $\mathcal{O}(1/k)$ rate as predicted by the theory. We then compare our methods with [56] and provide one application to illustrate the advantages of the proposed algorithms.

6.1. A degenerate linear program. We aim at comparing different algorithms to solve the following simple linear program:

$$(40) \quad \begin{cases} \min_{x \in \mathbb{R}^n} & 2x_n \\ \text{s.t.} & \sum_{k=1}^{n-1} x_k = 1, \\ & x_n - \sum_{k=1}^{n-1} x_k = 0 \quad (2 \le j \le d), \\ & x_n \ge 0. \end{cases}$$

The second inequality is repeated $d-1$ times, which makes the problem degenerate. Yet qualification conditions hold since this is a feasible and bounded linear program. This fits into our framework with $f(x) := 2x_n + \delta_{\{x_n \ge 0\}}(x_n)$, $Ax := [\sum_{k=1}^{n-1} x_k; x_n - \sum_{k=1}^{n-1} x_k; \cdots; x_n - \sum_{k=1}^{n-1} x_k]$, $c := (1, 0, \ldots, 0)^\top \in \mathbb{R}^d$, and $g(\cdot) := \delta_{\{c\}}(\cdot)$. A primal and dual solution can be found explicitly, and by playing with the sizes n and d of the problem, one can control the degree of degeneracy.

In this test, we choose $n = 10$ and $d = 200$. We implement both ASGARD and ADSGARD and their restart variants. In Figure 1, we compare our methods against the Chambolle–Pock method [15]. We can see that the Chambolle–Pock method struggles with the degeneracy, while ASGARD still exhibits an $\mathcal{O}(1/k)$ sublinear convergence rate as predicted by our theory.
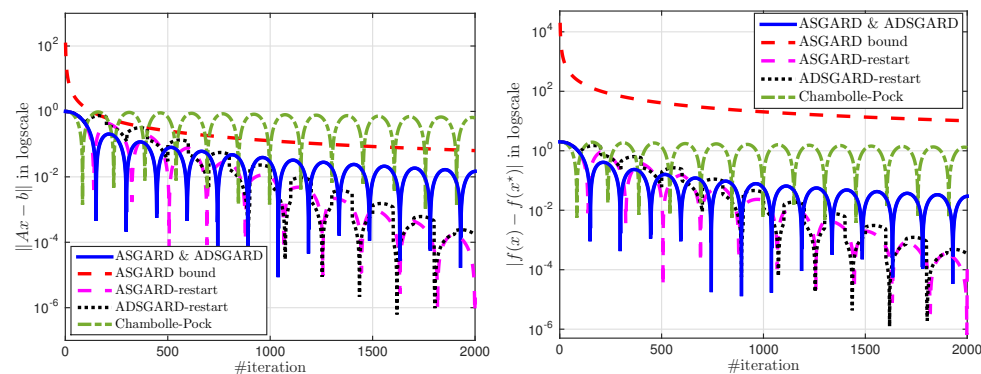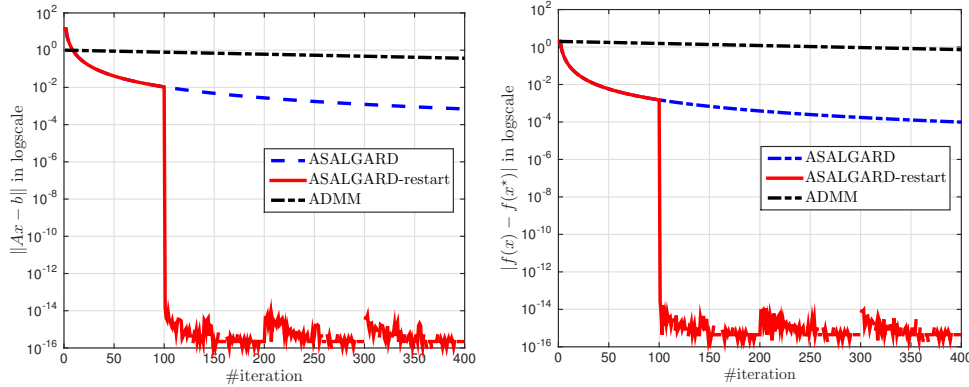

FIG. 1. *Comparison of the absolute feasibility violation (left) and the absolute objective residual (right) for* (ASGARD) *(solid blue line),* (ASGARD) *with a restart every* 100 *iterations using* (38) *(dashed pink line),* (ADSGARD) *with a restart every* 100 *iterations using* (39) *(black dotted line), and Chambolle–Pock (green dashed-dotted line). The dashed red line is the theoretical bound of* (ASGARD) *(Theorem 4).* (ADSGARD) *leads to similar results as* (ASGARD) *on this linear program* (40): *the difference is not perceptible on the figure.*

FIG. 2. *Comparison of the absolute feasibility violation* (*left*) *and the absolute objective residual* (*right*) *for* (ASALGARD) (*solid blue line*), (ASALGARD) *with a restart every* 100 *iterations* (*dashed red line*), *and ADMM* (*black dotted line*).

In Figure 2, we compare methods requiring the resolution of a nontrivial optimization subproblem at each iteration. In this case, the inversion of a rank-deficient linear system, we thus compare ASALGARD with and without restart against ADMM [11]. For ADMM, we selected the step-size parameter by sweeping from small values to large values and choosing the one that gives us the fastest performance. Again, our algorithm resists the degeneracy, and restarting strategies improve the performance, while ADMM has a very slow convergence rate.

6.2. Generalized convex feasibility problem. Given N nonempty, closed and convex sets $\mathcal{X}_i \subseteq \mathbb{R}^n$ for $i = 1, \ldots, N$, we consider the following optimization problem:

$$
(41) \qquad \min_{x:=(x_1^\top, \ldots, x_N^\top)^\top \in \mathbb{R}^{Nn}} \left\{ f(x) := \sum_{i=1}^N s_{\mathcal{X}_i}(x_i) \mid \sum_{i=1}^N A_i^\top x_i = \mathbf{0}^m \right\},
$$

where $s_{\mathcal{X}_i}$ is the support function of $\mathcal{X}_i$ and $A_i \in \mathbb{R}^{n \times m}$ is given for $i = 1, \ldots, N$.

It is trivial to show that the dual problem of (41) is the following generalization of a convex feasibility problem:

$$
(42) \qquad \text{Find } y^\star \in \mathbb{R}^m \text{ such that } A_i y^\star \in \mathcal{X}_i \ (i = 1, \ldots, N).
$$

Clearly, when $A_i = \mathbb{I}$ the identity matrix, (42) becomes the classical convex feasibility problem. When $A_i = \mathbb{I}$ for some $i \in \{1, \ldots, N\}$ and $A_i = A$ otherwise, (42) becomes a multiple-set split feasibility problem considered in the literature. Assume that (42) has a solution and $N \geq 2$. Hence, (41) and (42) satisfy Assumption A.1.

Our aim is to apply Algorithm 1 and Algorithm 2 to solve the primal problem (41) and compare them with the most state-of-the-art ADMM algorithm with multiple blocks [26]. Clearly, with nonorthogonal A_i, the primal subproblem of computing x_i in the parallel-ADMM scheme [26] does not have a closed form solution, and we need to solve it iteratively up to a given accuracy. In addition, by a change of variable, we can rescale the iterates such that ADMM does not depend on the penalty parameter when solving (41). With the use of Euclidean distance for our smoother, Algorithm 1 and Algorithm 2 can solve the primal subproblem (11) in x_i with a closed form solution, which only requires one projection onto $\mathcal{X}_i$.
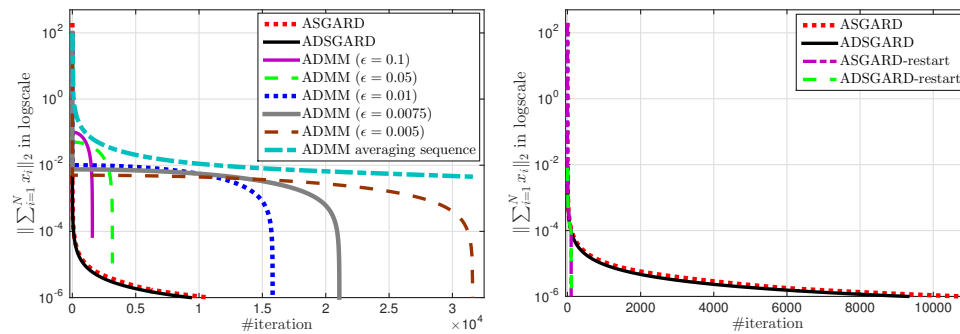
FIG. 3. *Comparison of Algorithm* 1, *Algorithm* 2, *and ADMM with different values of* ϵ (*left*). *Comparison of Algorithm* 1, *Algorithm* 2, *and their restart variant* (*restarting after every* 100 *iterations*) (*right*). *The number of variables is* 20,000.

The first experiment is for $N = 2$. We choose $\mathcal{X}_1 := \{y \in \mathbb{R}^n \mid \epsilon y_1 - \sum_{j=2}^{n} y_j \leq 1\}$ and $\mathcal{X}_2 := \{y \in \mathbb{R}^n \mid \sum_{i=2}^{n} y_j \leq -1\}$ to be two half-planes, where $\epsilon > 0$ is fixed. The constant ϵ represents the angle between these half-planes. It is well-known [69] that the ADMM algorithm can be written equivalently to an alternating projection method on the dual space. The convergence of this algorithm strongly depends on the angle between these sets. By varying ϵ, we observe the convergence speed of ADMM is also varying, while our algorithms seem not to depend on ϵ. Figure 3 shows the convergence rate on the absolute feasibility gap $\|\sum_{i=1}^{N} x_i\|_2$ of three algorithms for $n = 10,000$. Since the objective value is always zero, we omit its plot here.

The theoretical version of Algorithm 1 and Algorithm 2 exhibits a convergence rate slightly better than $\mathcal{O}(1/k)$ and is independent of ϵ, while ADMM can be arbitrarily slow as ϵ decreases. ADMM very soon drops to a certain accuracy and then is saturated at that level before it converges. Algorithm 1 and Algorithm 2 also quickly converge to the 10^{-5} accuracy level and then make a slower progress to achieve the 10^{-6} accuracy but still obey our theoretical guarantee. We notice that the averaging sequence of ADMM converges at the $\mathcal{O}(1/k)$ rate, but it remains far away from our theoretical rate in Algorithm 1 and Algorithm 2 due to a big constant factor. If we combine these two algorithms with our restart strategy, both algorithms need 102 iterations to reach the desired accuracy. We can see that Algorithm 1 performs very similar to Algorithm 2. We can also observe that the performance of our algorithms depends on $\bar{L}_A$ and initial points, but it is relatively independent of the geometric structure of problems as opposed to the ADMM for solving the generalized convex feasibility problem (41).

Now we extend to the cases of $N = 3$ and $N = 4$, where we add two more sets $\mathcal{X}_3$ and $\mathcal{X}_4$. We choose $\mathcal{X}_3 := \{y \in \mathbb{R}^n \mid 0.5\epsilon y_1 - \sum_{j=2}^{n} y_j = 1\}$ to be a hyperplane in $\mathbb{R}^n$ and $\mathcal{X}_4 := \{y \in \mathbb{R}^n \mid -y_1 + \sum_{j=3}^{n} y_j \leq 1\}$ to be a half-plane in $\mathbb{R}^n$. We test our algorithms and the multiblock-ADMM method in [26] again.

The results are plotted in Figure 4 for the case $n = 10,000$. In both cases, the ADMM algorithm still makes slow progress as ϵ is decreasing and N is increasing. Algorithm 1 and Algorithm 2 seem to scale slightly to N, the number of blocks. We note that since $A_i = \mathbb{I}$ for $i = 1, \ldots, N$, the per-iteration complexity of three algorithms in our experiment is essentially the same.
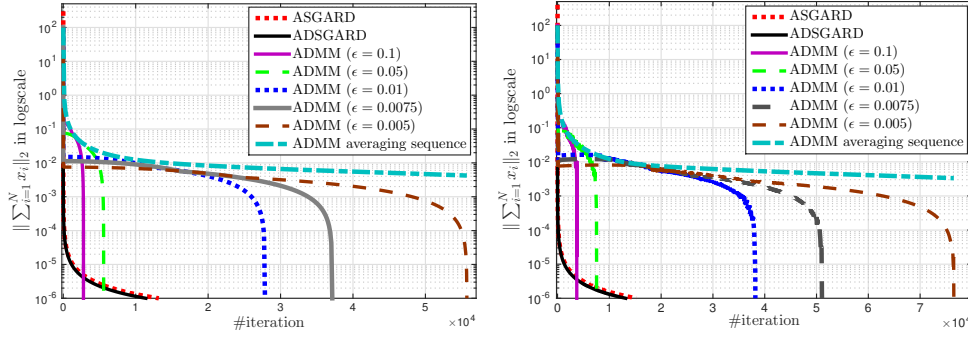
FIG. 4. *Comparison of Algorithm* 1*, Algorithm* 2*, and ADMM with different values of ϵ. The left plot is for $N = 3$, and the right one is for $N = 4$.*

6.3. A comparison with [56]. To see the advantages of our homotopy strategy, we consider the following square-root LASSO problem considered in the literature, e.g., [6]:

$$(43) \qquad P^\star = \min_{x \in \mathbb{R}^n} \left\{ P(x) := \tfrac{1}{\sqrt{m}} \|Ax - b\|_2 + \lambda \|x\|_1 \right\},$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $\lambda > 0$ is a given regularization parameter. As suggested in [6], we can choose $\lambda := \frac{1.1}{\sqrt{m}} \Phi^{-1}(1 - 0.5\alpha/n)$, where $\alpha = 0.05$ and Φ is the standard normal distribution. By letting $f(x) := \lambda \|x\|_1$ and $g(u) := \|u - b\|_2 = \max \left\{ \langle u, y \rangle - \langle b, y \rangle \mid \|y\|_2 \leq 1 \right\}$, we can easily check that f and g satisfy Assumption 1. Clearly, the conjugate function $g^*(y) := \langle b, y \rangle + \delta_{\mathcal{B}_2(0;1)}(y)$, where $\delta_{\mathcal{B}_2(0;1)}$ is the indicator function of the ℓ_2-norm ball $\mathcal{B}_2(0;r) := \{y \in \mathbb{R}^m \mid \|y\|_2 \leq r\}$.

Since the solutions of (43) are sparse, we apply Algorithm 1 to solve this problem and compare it with Nesterov's method in [56]. Let us choose $b_{\mathcal{X}}(x, \dot{x}) := \frac{1}{2} \|x - \dot{x}\|_2^2$ and $b_{\mathcal{Y}}(y, \dot{y}) := \frac{1}{2} \|y - \dot{y}\|_2^2$ with $\dot{x} = 0$ and $\dot{y} = 0$. In this case, $y_\beta^*(x; \dot{y})$ can be computed as the projection on the unit ℓ_2-ball, while $x_\gamma^*(y; \dot{x})$ is computed from the proximal operator of the ℓ_1-norm (a soft-thresholding operator). We initialize the algorithms at $x^0 = 0$.

Let us try to tune the smoothness parameter in both algorithms. Nesterov suggested to tune it as follows. Given an iteration budget K and an a priori bound on the distance to the solution, choose the smoothness parameter that minimizes the known theoretical bound. In Figure 5, this corresponds to "β with guarantees." For this square-root LASSO problem, we take $K := 10^5$. Theoretically, we can show that $\|x^\star - x^0\|_2 \leq \|x^\star\|_1 \leq \frac{\|b\|_2}{\lambda \sqrt{m}}$. We can also compute $D_{\mathcal{Y}} := \frac{1}{2}$, where $\mathcal{Y} := \{y \in \mathbb{R}^m \mid \|y\|_2 \leq 1\}$ is an ℓ_2-unit ball. The theoretical bound derived from [56] becomes $P(x^K) - P^\star \leq \frac{4\|A\|\|b\|_2 \sqrt{D_{\mathcal{Y}}}}{\lambda \sqrt{m}(K+1)}$. Similarly, in our algorithms, we set β_1 and γ_1 as suggested by (23) and (29), respectively, and we obtain a slightly better theoretical bound.

In practice, the a priori bound on the distance $\|x^\star - x^0\|_2$ may be very conservative. Hence, we will also use the quantity $\frac{\|b\|_2}{\lambda \sqrt{mn}}$ as an estimate of $R_0 := \|x^\star - x^0\|_2$. In Figure 5, this corresponds to "β a priori tuned."

We generate matrix A using standard Gaussian distribution $\mathcal{N}(0, 1)$ with 25% correlated columns. The true parameter $x^\natural$ is a given s-sparse vector. We generate the observation b as $b := Ax^\natural + \mathcal{N}(0, 0.005)$, where the last term represents Gaussian noise.
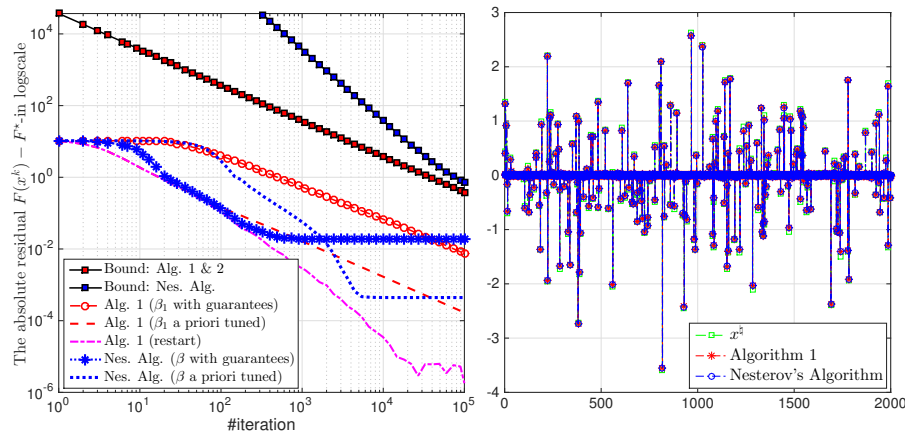
FIG. 5. *Comparison of Algorithm* 1 (Alg. 1) *and Nesterov's smoothing algorithm in* [56] (Nes. Alg.). *Left: convergence of the algorithms and their theoretical bounds. Right: recovered solutions and* $x^\natural$. (ASGARD) (*red color plots*) *features a steady* $\mathcal{O}(1/k)$ *decrease. Nesterov's smoothing algorithm* (*blue color plots*) *has a slower start than* (ASGARD)*, then enters a quick decrease phase, and finally stagnates when it has reached the minimum of the smoothed problem. At the end of the iteration budget,* (ASGARD) *returns a more accurate solution. This behavior can be seen for both parameter tuning strategies and is consistent with the theoretical bound.*

Figure 5 illustrates the performance of the two algorithms for solving (43), where $m = 700$, $n = 2000$, and $s = 200$. The left-plot shows the convergence behavior of both algorithms and their theoretical bounds. We clearly see that for each tuning strategy, Algorithm 1 reaches a smaller final objective value than the one in [56]. Moreover, Nesterov's method has the disadvantage of stagnating after a given moment, while Algorithm 1 makes steady progress. Restarting the method every 25 iterations gives improvement again in the performance. We can finally observe that both algorithms have better performance than their theoretical worst-case bounds. Figure 5 (right) shows the solutions of both algorithms and compares them with the true parameter $x^\natural$. We see that the obtained solutions fit well $x^\natural$, and they are both sparse solutions.

6.4. Application to image reconstruction. In this example, we propose to use the following total variation norm optimization formulation to reconstruct images from compressive measurements b obtained via a linear operator $\mathcal{L}$:

$$(44) \qquad \min_{Z \in \mathbb{R}^{p_1 \times p_2}} \left\{ f(Z) := \|D(Z)\|_1 \mid \mathcal{L}(Z) = b \right\},$$

where D is a two-dimensional discrete gradient operator, $\mathcal{L} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$ is a linear transformation obtained from a subsampling-FFT transformation, and $b \in \mathbb{R}^n$ is a compressive measurement vector. We first reformulate problem (44) into the form (3) using a splitting trick as follows:

$$(45) \qquad f^\star := \begin{cases} \min_{x := [u^\top, \mathrm{vec}(Z)^\top]^\top} & \left\{ f(x) := \|u\|_1 \right\} \\ \text{s.t.} & \mathcal{L}(Z) = b, \quad D(Z) - u = 0. \end{cases}$$

We now apply Algorithm 1 and Algorithm 2 to solve this problem and compare them with the Chambolle–Pock method in [15, 72]. We also compare our methods with a line-search variant of the Chambolle–Pock method recently proposed in [43]. We note that our algorithms and the standard Chambolle–Pock method have the

TABLE 1
Performance and results of the five algorithms on two MRI images.

Algorithms	MRI-knee (650 × 650)					MRI-brain-tumor (630 × 611)				
	$f(Z^k)$	$\frac{\|\mathcal{L}(Z^k)-b\|}{\|b\|}$	Error	PSNR	Time[s]	$f(Z^k)$	$\frac{\|\mathcal{L}(Z^k)-b\|}{\|b\|}$	Error	PSNR	Time[s]
ASGARD	39.927	2.426e-03	2.305e-02	90.07	172.85	54.186	2.206e-03	3.179e-02	85.81	101.94
ASGARD-restart	40.126	6.443e-04	2.290e-02	90.12	173.41	54.655	7.122e-04	3.109e-02	86.00	103.07
ADSGARD	39.372	3.580e-03	2.336e-02	89.95	210.36	53.727	3.353e-03	3.238e-02	85.65	115.36
Chambolle–Pock	39.931	3.710e-02	6.689e-02	80.82	160.12	54.408	5.748e-02	1.232e-01	74.04	107.71
Linesearch CP	41.291	4.514e-03	2.563e-02	89.15	469.35	54.720	4.005e-03	3.520e-02	84.92	317.09

same per-iteration complexity. We first test all the algorithms on two MRI images:
`MRI-brain-tumor` and `MRI-of-knee`.[1] We follow the procedure in [37] to generate the
samples using a sample rate of 20%. Then the vector of measurements b is computed
from $b := \mathcal{L}(Z^\natural)$, where $Z^\natural$ is the original image. Our experiment is implemented
in Matlab 2014b running on a MacBook Pro (Retina, 2.7 GHz Intel Core i5, 16 GB
1867 MHz).

In the Chambollle–Pock method, we use the parameters as suggested in [15] with
$\tau = \sigma = \|A\|^{-1}$ (see [43]). For the line-search variant of the of the Chambollle–Pock
method in [43] (denoted by `Linesearch CP`), we tune the parameters to obtain the
best performance on a set of sample images. These parameters are set to $\beta = 10^5$,
$\mu = 0.7$, and $\delta = 0.99$. Since we aim at reducing the feasibility gap, as guided by our
theoretical results above, we use $\beta_1 = 10^{-3}\|A\|$ in our algorithms. The performance
and results of these algorithms are summarized in Table 1, where $\text{Error} := \frac{\|Z^k - Z^\natural\|_F}{\|Z^\natural\|_F}$
presents the error between the original image $Z^\natural$ to the reconstruction Z^k after $k = 500$
iterations.

As we can see, both Algorithm 1 and Algorithm 2 have comparable performance
with the line-search variant of the Chambolle–Pock method in terms of accuracy while
outperforming the standard variant. In fact, our standard ASGARD method is still
slightly better than this line-search version. The ASGARD with restart gives the
best performance in terms of accuracy as well as peak signal-to-noise ratio (PSNR).
The `Linesearch CP` is more than three times slower than the other methods in this
experiment.

The reconstructed images are revealed in Figure 6. As we can see from this plot,
the quality of recovery image is very close to the original image for the sampling rate
of 20%. Our algorithms give slightly higher PSNR for both images. The Chambolle–
Pock algorithm has the worst performance compared to the others. However, the
performance of this algorithm can slightly be changed and depends on the tuning
strategy of the step-size parameters [15], which is often very hard to tune a priori in
practice without using a heuristic strategy.

7. A comparison between our results and existing methods. We have
presented a new primal-dual framework and two main algorithms (one with a pri-
mal flavor and one with a dual flavor) together with two special cases. Now let us
summarize the main differences between our approach and existing methods in the
literature.

The composite convex problem (1) can be written as a convex-concave saddle
point problem:

$$(46) \qquad \min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} \left\{ \Phi(x, y) := f(x) + \langle Ax, y \rangle - g^*(y) \right\}.$$

---

[1]These images are from https://radiopaedia.org/cases/4090/studies/6567 and https://www.
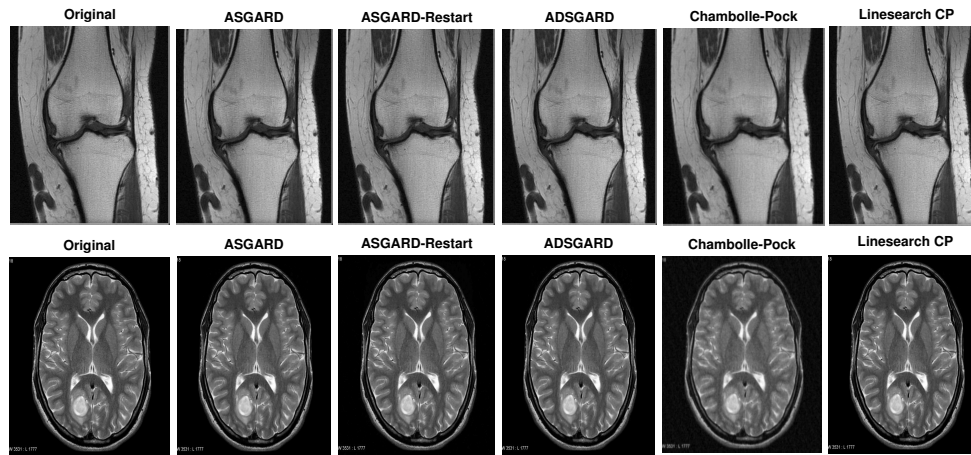nibib.nih.gov.

Fig. 6. *The original image and the reconstructed images of the five algorithms.*

The optimality condition of this problem is a maximal monotone inclusion and can be reformulated as a variational inequality (VIP) [2, 29, 62]. While (46) is classical, it has broad applications in image processing, machine learning, and game theory, among many others [2, 20, 29, 52]. Recent developments in solution methods for solving (46) has attracted great attention. Let us briefly survey some notable works which we find most related to ours.

Nemirovskii in [52] proposed an averaging scheme to solve (46) based on its VIP formulation. His algorithm requires a proximal step at each iteration, which is usually not easy to compute in applications. He proved an $\mathcal{O}(1/k)$ convergence rate in an ergodic sense for the primal-dual gap function under the boundedness of both the primal and the dual domains. Nesterov in [57] proposed a similar method to solve VIP that covers (46) as a special case. By using smoothing techniques, he could prove an $\mathcal{O}(1/k)$ convergence rate for the gap function in an ergodic sense as in [52]. While this method has a simple subproblem, it requires the underlying operator to be Lipschitz continuous, and both primal and dual domains are bounded.

One of the most celebrated works for solving nonsmooth convex problem (1) is due to Nesterov in [56]. By combining both the smoothing technique and his accelerated gradient-type method, he proposed an algorithm to solve (1) just using proximal operators of f and g^*. The method achieves an $\mathcal{O}(1/\varepsilon)$ complexity to obtain an ε-solution. However, this algorithm has two disadvantages. First, it requires the boundedness of both the primal and dual domains. Second, the smoothness parameter depends on both the accuracy ε and the diameter of the primal and dual domains. An improvement was proposed in [55] to remove the second disadvantage. But this algorithm requires a symmetric update which leads to a different per-iteration complexity than [56].

Another remarkable work was proposed by Chambolle and Pock in [15]. This algorithm solves (46) just using the proximal operator of f and g^*. Similar to [52], they also proved an $\mathcal{O}(1/k)$ convergence rate on the gap function in an ergodic sense requiring the boundedness of both the primal and dual domains. An improvement on the parameter range was proposed in [35]. However, the convergence guarantee remains preserved under the same assumptions. Extensions of [15] can be found in several papers, including [21, 22, 43].

Shefi and Teboulle provided a comprehensive study on the convergence rate of proximal-type methods for solving (1) in [64], which extended the work of [18]. They discussed different variants of the primal-dual proximal-type methods including the Chambole–Pock scheme, alternating minimization algorithms (AMA), and alternating direction methods of multipliers (ADMM). With a proper choice of metric, they showed an $\mathcal{O}(1/k)$-convergence guarantee on the primal-dual gap function in an ergodic sense. Their convergence guarantee indeed unifies several schemes but is different from our results in this paper. First, they used different metric for proximal terms depending on A. This makes the subproblem much harder to solve. Second, they provided a guarantee for convergence rate on the gap function, which is not obvious to transform it into a separated guarantee for the objective and constraints in constrained convex optimization settings such as (3). Moreover, the rate on the feasibility in the constrained setting reduces to $\mathcal{O}(1/\sqrt{k})$ (see [64, Theorem 5]). Finally, the gap function is defined on a given domain, and it is not clear how to choose the radius of this domain.

Other methods for solving (46) can be found in the literature including [15, 19, 23, 24, 25, 36, 60, 61]. Each method requires different structure assumptions and achieves different guarantees mostly in an ergodic or averaging sense. For instance, in [19], the authors required f to have a Lipschitz gradient which is much more limited than our assumptions. The authors in [36] specified a so-called hybrid proximal extragradient framework proposed in [65] to solve (46). While this method achieves an $\mathcal{O}(1/\varepsilon)$ complexity without any boundedness assumption on the domains, it is rather complicated due to a double loop of inner and outer iterations.

Regarding the constrained setting (3), primal-dual first-order methods for directly solving large-scaled settings of this problems are also well developed. Let us briefly discuss some of these methods here. A natural approach is due to dual gradient-type methods. Such methods often use directly the subgradients of the dual function or smooth the dual function using proximity terms [50, 49]. While the former gives a slow convergence rate, which is $\mathcal{O}(1/\sqrt{k})$, the latter uses Nesterov's smoothing technique in [56] and therefore faces the same drawbacks. In addition, for the setting (3), the dual domain is unbounded, which leads to a difficulty to estimate the worst-case complexity bounds. Another approach is using penalty or augmented Lagrangian as considered in [38, 39, 51], which often leads a two-loop algorithm and is much more complicated to control the specified parameters and accuracies in practice. Alternating direction methods are perhaps the most common use for solving (3); see [11, 15, 23, 24, 25, 60, 61]. This method often requires an additional structure assumption, such as f being the sum of two separable convex functions. Without this structure, auxiliary variables need to be introduced, which is again equivalent to the two-block case; see, e.g., [11, 74, 75]. Two common methods in this direction are AMA and ADMM. While AMA can be viewed as a forward-backward splitting scheme for the dual formulation (2) and requires strict assumptions to guarantee convergence (e.g., one objective component is strongly convex), ADMM is equivalent to the Douglas–Rachford splitting method applying to the dual and has a convergence guarantee under mild assumptions. Recently, the convergence rate of ADMM has been attracted great attention. Many papers have studied ADMM and its variants, including [11, 24, 34, 60, 61]. We would also like to mention that during the revision process of this paper, variants of ALM and ADMM with similar convergence rates as ours have been proposed in [76].

In summary, this paper has tried to overcome several issues we have mentioned above in recent primal-dual methods. Let us highlight the following characterizations.

Problem structure assumptions. Our approach requires the convexity and the existence of primal and dual solutions of (1). In the unconstrained setting (1), we require g to be Lipschitz continuous, which is often the case in practice. We argue that such assumptions are mild for (1) and (2) and can be verified a priori. We emphasize that existing primal-dual methods in [15, 23, 24, 25, 60, 61, 64] require other structure assumptions on either f, such as Lipschitz gradient, strong convexity, error bound conditions, or the boundedness of both the primal and dual feasible sets, which may not be satisfied for (1) and especially for (3) [15, 23, 24, 25, 60, 61].

Convergence characterization. We characterize an $\mathcal{O}(1/k)$-convergence rate for both the objective residual $f(x^k) - f^\star$ and the feasibility violation $\|Ax^k - c\|_{\mathcal{Y},*}$ for the constrained convex problem (3). Our finding is the best-known result under very mild assumptions and low per-iteration complexity. In the composite form (1), we also achieve the same $\mathcal{O}(1/k)$-convergence rate as in the seminal work [56]. Tables 2 and 3 compare our theoretical convergence rate results with the most recent selected algorithms in the literature for solving (1) and (3), respectively.

Decomposition methods. Our algorithms naturally support decomposable structures in f without either reformulating the problem as in ADMM or requiring additional assumptions as in parallel and multiblock ADMM [42]. Both algorithms simply require only one proximal operator of f and g^*, one matrix vector multiplication, and one adjoint per iteration.

Smoothing and smoothness parameter updates. In contrast to proximal-type approaches in [45, 46, 64, 65], where the subproblem is often more complicated to solve, we instead exploit Nesterov's smoothing technique [56], which allows us to use proximal operators of f and g^*. However, we use differentiable smoothing functions as compared to Nesterov's smoothing technique in [56]. We propose explicit rules to update the smoothness parameters simultaneously at each iteration. We emphasize that this is one of the key contributions of this paper. To the best of our knowledge, this is the first adaptive primal-dual algorithms for smoothness parameters without sacrificing the $\mathcal{O}(1/k)$ rate and requiring additional assumptions.

Averaging vs. nonaveraging. Most existing methods employ either nonweighted averaging [15, 34, 35, 64] or weighted averaging schemes [24, 60, 61] to guarantee the $\mathcal{O}(1/k)$ rate on the primal sequence. While we also provide a weighted averaging scheme (Algorithm 2), we alternatively derive a method (Algorithm 1) without any averaging in the primal for solving (1). The nonaveraging schemes are important since taking averages may destroy key structures, such as the sparsity or low-rankness in sparse or low-rank optimization. Our weighted averaging scheme has increasing weight at the later iterates compared to nonweighted averaging schemes [15, 34, 35, 64]. As indicated in [23, 25], weighted averaging schemes have a better performance guarantee than nonweighted ones.

We have attempted to review various primal-dual methods which are most related to our work. It is still worth mentioning other primal-dual methods that are based on augmented Lagrangian methods, such as alternating direction methods (e.g., AMA, ADMM, and their variants) [11, 38, 39, 71], Bregman and other splitting methods [2, 21, 28, 32, 48, 46, 47], and using variational inequality frameworks [15, 35, 33]. While most of these works have not considered the global convergence rate of the proposed algorithms, a few of them characterized the convergence rate in unweighted averaging schemes or used a more general variational inequality/monotone inclusion framework to study (1)–(3). Hence, the results achieved in these papers are distinct from our findings.

TABLE 2

A comparison of convergence rates between our algorithms and selected existing methods for solving (1) and (3). Here all algorithms do not involve any large matrix inversion or "complex" convex subproblem, and $w_k := \frac{1}{k}\sum_{l=1}^{k} w^l$, K is the iteration budget (see subsection 6.3), and σ is the step-size in [15].

Paper	dom (f) bounded and g Lipschitz	g Lipschitz	$g = \delta_{\{c\}}$ (optimality and feasibility)
Nesterov [56]	$P(x^k) - P^\star \leq \frac{2\sqrt{L_A D_\mathcal{X} D_\mathcal{Y}}}{K}\left(1 + \frac{K^2}{k^2}\right)$	$P(x^k) - P^\star \leq \frac{\epsilon}{2} + \frac{8\bar{L}_A\|x^0 - x^\star\|^2 D_\mathcal{Y}}{\epsilon(k+1)^2}$	not applicable
Chambolle and Pock [15]	$G(w_k) \leq \frac{\sigma\bar{L}_A D_\mathcal{X} + \sigma^{-1}D_\mathcal{Y}}{k}$	convergence	convergence
ASGARD (section 3)	$P(x^k) - P^\star \leq \frac{2\sqrt{2}\sqrt{L_A D_\mathcal{Y}}D_\mathcal{X}}{k}$	$P(x^k) - P^\star \leq \frac{\bar{L}_A}{2\beta_1 k}\|\bar{x}^0 - x^\star\|^2 + \frac{2\beta_1}{k}D_\mathcal{Y}$	$\|f(x^k) - f^\star\| \leq \frac{\bar{L}_A}{\beta_1 k}\|\bar{x}^0 - x^\star\|^2 + \frac{3\beta_1}{k}\|\bar{y} - y^\star\|^2 + \frac{\beta_1}{k}\|y^\star\|^2$ $\|Ax^k - c\|_{\mathcal{Y},*} \leq \frac{\beta_1}{k+1}\left(2\|\bar{y} - y^\star\| + \frac{\sqrt{L_A}}{\beta_1}\|\bar{x}^0 - x^\star\|\right)$
ADSGARD (section 4)	$G(w^k) \leq \frac{2\sqrt{L_A D_\mathcal{Y}}D_\mathcal{X}}{k}$	$G(w^k) \leq \frac{\gamma_1}{k+1}\|\hat{x} - x^\star\|^2 + \frac{2\bar{L}_A}{\gamma_1 k}D_\mathcal{Y}$	$\|f(x^k) - f^\star\| \leq \frac{3\gamma_1}{k}\|\hat{x} - x^\star\|^2 + \frac{2\bar{L}_A}{\gamma_1 k}\|\hat{y} - y^\star\|^2 + \frac{\bar{L}_A}{\gamma_1 k}\|y^\star\|^2$ $\|Ax^k - c\|_{\mathcal{Y},*} \leq \frac{\bar{L}_A}{\gamma_1 k}\left(2\|\hat{y} - y^\star\| + 2\gamma_1\|\hat{x} - x^\star\|\right)$

TABLE 3

A comparison of convergence rates between our algorithms and selected existing methods for solving (3). *Here all algorithms may involve "complex" convex subproblems or matrix inversions, and ρ is the penalty parameter in* [39] *and* [47].

Paper	$g = \delta_{\{c\}}$
ALM [39]	$\|f(x^k) - f^\star\| \leq \frac{6}{\rho\sqrt{k}}\|y^\star\|\|y^0 - y^\star\|$
	$\|Ax^k - c\|_{\nu,*} \leq \frac{3}{\rho\sqrt{k}}\|y^0 - y^\star\|$
ADMM[1] [47][2]	$\|f_1((x_1)_k) + f_2((x_2)_k) - f_1(x_1^\star) - f_2(x_2^\star)\| \leq \frac{6+4\sqrt{2}}{k}\left(\frac{1}{\rho}\|y^0 - y^\star\|^2 + \rho\|x_1^0 - x_1^\star\|^2_{A_1^*A_1}\right)$
	$\|A_1(x_1)_k + A_2(x_2)_k - c\| \leq \frac{2}{k}\sqrt{\frac{1}{\rho^2}\|y^0 - y^\star\|^2 + \|x_1^0 - x_1^\star\|^2_{A_1^*A_1}}$
ASALGARD (section 5.1)	$\|f(x^k) - f^\star\| \leq \frac{10\|y^\star\|_\nu\|\hat{y}-y^\star\|_\nu}{\gamma_0(k+1)^2}$
	$\|Ax^k - c\|_{\nu,*} \leq \frac{8\|\hat{y}-y^\star\|_\nu}{\gamma_0(k+2)^2}$

[1] *Note that ADMM splits the objective in two parts, which may make it very efficient for some problems.* [2] *The original result is stated by means of ϵ-subdifferentials.*

Appendix A. The proof of theoretical results.   This section provides the full proof of lemmas and theorems in the main text.

A.1. Technical results. We first prove the following basic lemma, which will be used to analyze the convergence of our algorithms in the main text.

LEMMA 10. *Let h be a proper, closed and convex function defined on $\mathcal{Z}$ and h^* its Fenchel conjugate. Let $b_{\mathcal{Z}}$ be a Bregman distance as defined in (7) with a weighted norm. We define a smoothed approximation of h as*

$$(47) \qquad h_\beta(z; \dot{z}) := \max_{\hat{z} \in \mathcal{Z}} \left\{ \langle z, \hat{z} \rangle - h^*(\hat{z}) - \beta b_{\mathcal{Z}}(\hat{z}, \dot{z}) \right\},$$

where $\dot{z} \in \mathcal{Z}$ is fixed and $\beta > 0$ is a smoothness parameter. We also denote by $z_\beta^*(z; \dot{z})$ the solution of the maximization problem in (47). Then the following facts hold:
(a) *We have a relation between the partial derivatives of $(z, \beta) \mapsto h_\beta(z; \dot{z})$ as*

$$\frac{\partial h_\beta(z; \dot{z})}{\partial \beta}(\beta) = -b_{\mathcal{Z}}(z_\beta^*(z; \dot{z}), \dot{z}) = -b_{\mathcal{Z}}(\nabla h_\beta(z; \dot{z}), \dot{z}).$$

(b) *For all $z \in \mathcal{Z}$, $\beta \mapsto h_\beta(z; \dot{z})$ is convex, and for $\beta_{k+1}, \beta_k > 0$ and $\bar{z} \in \mathcal{Z}$, we have*

$$(48) \qquad h_{\beta_{k+1}}(\bar{z}; \dot{z}) \leq h_{\beta_k}(\bar{z}; \dot{z}) - (\beta_k - \beta_{k+1}) \frac{\partial h_\beta(\bar{z}; \dot{z})}{\partial \beta}(\beta_{k+1})$$
$$= h_{\beta_k}(\bar{z}; \dot{z}) + (\beta_k - \beta_{k+1}) b_{\mathcal{Z}}(\nabla h_{\beta_{k+1}}(\bar{z}; \dot{z}), \dot{z}).$$

(c) *$h_\beta(\cdot; \dot{z})$ has a $1/\beta$-Lipschitz gradient in $\| \cdot \|_{\mathcal{Z}, *}$. Hence, for all $\bar{z}, \hat{z} \in \mathcal{Z}$, we have*

$$(49) \qquad h_\beta(\bar{z}; \dot{z}) \leq h_\beta(\hat{z}; \dot{z}) + \langle \nabla h_\beta(\hat{z}; \dot{z}), \bar{z} - \hat{z} \rangle + \frac{1}{2\beta} \|\bar{z} - \hat{z}\|_{\mathcal{Z}, *}^2$$

$$(50) \quad h_\beta(\hat{z}; \dot{z}) + \langle \nabla h_\beta(\hat{z}; \dot{z}), \bar{z} - \hat{z} \rangle \leq h_\beta(\bar{z}; \dot{z}) - \frac{\beta}{2} \|\nabla h_\beta(\hat{z}; \dot{z}) - \nabla h_\beta(\bar{z}; \dot{z})\|_{\mathcal{Z}}^2.$$

(d) *Both functions h and h_β evaluated at different points $z, \hat{z} \in \mathcal{Z}$ satisfy*

$$(51) \qquad h_\beta(\hat{z}; \dot{z}) + \langle \nabla h_\beta(\hat{z}; \dot{z}), z - \hat{z} \rangle \leq h(z) - \beta b_{\mathcal{Z}}(\nabla h_\beta(\hat{z}; \dot{z}), \dot{z}).$$

(e) *If $\| \cdot \|_{\mathcal{Z}}$ is derived from a scalar product, then for all $\tau > 0$, $\bar{z}, \hat{z} \in \mathcal{Z}$, we have*

$$(52) \qquad (1 - \tau) \|\nabla h_\beta(\hat{z}; \dot{z}) - \nabla h_\beta(\bar{z}; \dot{z})\|_{\mathcal{Z}}^2 + \tau \|\nabla h_\beta(\hat{z}; \dot{z}) - \dot{z}\|_{\mathcal{Z}}^2$$
$$\geq \tau(1 - \tau) \|\nabla h_\beta(\bar{z}; \dot{z}) - \dot{z}\|_{\mathcal{Z}}^2.$$

(f) *We can control the influence of a change in the center points from $\dot{z}_1$ to $\dot{z}_2$ using the following estimate:*

$$(53) \qquad h_\beta(z; \dot{z}_2) \leq h_\beta(z; \dot{z}_1) - \frac{\beta}{2} \|z_\beta^*(z; \dot{z}_1) - z_\beta^*(y; \dot{z}_2)\|_{\mathcal{Z}}^2$$
$$+ \beta \left[b_{\mathcal{Z}}(z_\beta^*(z; \dot{z}_2), \dot{z}_1) - b_{\mathcal{Z}}(z_\beta^*(z; \dot{z}_2), \dot{z}_2) \right].$$

Proof. We prove from item (a) to item (f) as follows.
(a) Since $h_\beta(y)$ is defined by the maximization of a strongly convex program in (47), where the function in the max operator is linear in β and convex in z, the

minimizer $z^*_\beta(z; \dot{z})$ is unique. By the classical marginal derivative theorem [63], the function is differentiable with respect to β and z. In addition, $\nabla_z h_\beta(z; \dot{z}) = z^*_\beta(z; \dot{z})$.

(b) The function $\beta \mapsto h_\beta(z; \dot{z})$ is the maximization of a linear function in β indexed by y and $\dot{y}$. Hence, it is convex. The remaining statement follows by the convexity of h_β with respect to β and item (a).

(c) Since $\beta b_\mathcal{Z}(\cdot, \dot{z})$ is β-strongly convex in the weighted norm $\|\cdot\|_\mathcal{Z}$, $h_\beta(\cdot; \dot{z})$ is $\frac{1}{\beta}$-Lipschitz [56] in the corresponding dual norm. The inequalities (49) and (50) are classical for convex functions with Lipschitz gradient [54].

(d) Let us denote here $\hat{z}^*_\beta := z^*_\beta(\hat{z}; \dot{z})$. Then we can derive

$$h_\beta(\hat{z}; \dot{z}) + \langle \nabla h_\beta(\hat{z}; \dot{z}), z - \hat{z} \rangle = \left(\langle \hat{z}, \hat{z}^*_\beta \rangle - h^*(\hat{z}^*_\beta) - \beta b_\mathcal{Z}(\hat{z}^*_\beta, \dot{z}) \right) + \langle \hat{z}^*_\beta, z - \hat{z} \rangle$$
$$= \langle z, \hat{z}^*_\beta \rangle - h^*(\hat{z}^*_\beta) - \beta b_\mathcal{Z}(\hat{z}^*_\beta, \dot{z})$$
$$\leq \max_{u \in \mathcal{Z}} \{ \langle z, u \rangle - h^*(u) \} - \beta b_\mathcal{Z}(z^*_\beta, \dot{z})$$
$$= h(z) - \beta b_\mathcal{Z}(\nabla h_\beta(\hat{z}; \dot{z}), \dot{z}).$$

(e) The elementary equality $\|(1-\tau)a + \tau c\|^2 = (1-\tau)\|a\|^2 + \tau\|c\|^2 - \tau(1-\tau)\|a-c\|^2$ directly implies the result for any norm $\|\cdot\|$ deriving from a scalar product.

(f) Let us denote by $z^*_{\beta,1} = z^*_\beta(z; \dot{z}_1)$ and $z^*_{\beta,2} := z^*_\beta(z; \dot{z}_2)$. Using the definition of h_β in (47) and its optimality condition, we can derive

$$h_\beta(z; \dot{z}_2) = \max_{\hat{z} \in \mathcal{Z}} \{ \langle z, \hat{z} \rangle - h^*(\hat{z}) - \beta b_\mathcal{Z}(\hat{z}, \dot{z}_2) \} = \langle z, z^*_{\beta,2} \rangle - h^*(z^*_{\beta,2}) - \beta b_\mathcal{Z}(z^*_{\beta,2}, \dot{z}_2)$$
$$= \left(\langle z, z^*_{\beta,2} \rangle - h^*(z^*_{\beta,2}) - \beta b_\mathcal{Z}(z^*_{\beta,2}, \dot{z}_1) \right) + \beta b_\mathcal{Z}(z^*_{\beta,2}, \dot{z}_1) - \beta b_\mathcal{Z}(z^*_{\beta,2}, \dot{z}_2)$$
$$\leq \langle z, z^*_{\beta,1} \rangle - h^*(z^*_{\beta,1}) - \beta b_\mathcal{Z}(z^*_{\beta,1}, \dot{z}_1)$$
$$\quad - \frac{\beta}{2} \|z^*_{\beta,1} - z^*_{\beta,2}\|^2_\mathcal{Y} + \beta b_\mathcal{Z}(z^*_{\beta,2}, \dot{z}_1) - \beta b_\mathcal{Z}(z^*_{\beta,2}, \dot{z}_2)$$
$$= h_\beta(z; \dot{z}_1) - \frac{\beta}{2} \|z^*_{\beta,1} - z^*_{\beta,2}\|^2_\mathcal{Y} + \beta \left(b_\mathcal{Z}(z^*_{\beta,2}, \dot{z}_1) - b_\mathcal{Z}(z^*_{\beta,2}, \dot{z}_2) \right),$$

which proves (53). $\square$

A.2. The proof of Lemma 1: Key bounds for approximate solutions. We consider the smooth objective residual $S_\beta(x; \dot{y}) := \left(f(x) + g_\beta(Ax; \dot{y}) \right) - \left(f(x^\star) + g(Ax^\star) \right)$. By using the definition of g_β, we can derive that

$$g_\beta(Ax; \dot{y}) = \max_{\hat{y} \in \mathcal{Y}} \{ \langle Ax, \hat{y} \rangle - g^*(\hat{y}) - \beta b_\mathcal{Y}(\hat{y}, \dot{y}) \}$$
$$\geq \langle Ax, y^\star \rangle - g^*(y^\star) - \beta b_\mathcal{Y}(y^\star, \dot{y})$$
$$= \langle Ax - Ax^\star, y^\star \rangle + \langle Ax^\star, y^\star \rangle - g^*(y^\star) - \beta b_\mathcal{Y}(y^\star, \dot{y})$$
$$\text{(54)} \qquad = \langle A(x - x^\star), y^\star \rangle + g(Ax^\star) - \beta b_\mathcal{Y}(y^\star, \dot{y}),$$

where the last line is the equality case in the Fenchel–Young inequality using the fact that $Ax^\star \in \partial g^*(y^\star)$. Similarly, we have

$$f^*_\gamma(-A^\top y; \dot{x}) = \max_{\hat{x} \in \mathcal{X}} \{ \langle -A^\top y, \hat{x} \rangle - f(\hat{x}) - \gamma b_\mathcal{X}(\hat{x}, \dot{x}) \}$$
$$\text{(55)} \qquad \geq \langle A^\top(y^\star - y), x^\star \rangle + f^*(-A^\top y^\star) - \gamma b_\mathcal{X}(x^\star, \dot{x}).$$

Combining (55), the definition (15) of $G_{\gamma\beta}(\cdot; \dot{w})$, and the strong duality condition (8), we can show that

$$
\begin{aligned}
G_{\gamma\beta}(w; \dot{w}) &:= P_\beta(x; \dot{y}) - D_\gamma(y; \dot{x}) \\
&= f(x) + g_\beta(Ax; \dot{y}) + f_\gamma^*(-A^\top y; \dot{x}) + g^*(y) \\
&\overset{(8)}{=} S_\beta(x; \dot{y}) + f_\gamma^*(-A^\top y; \dot{x}) + g^*(y) - f^*(-A^\top y^\star) - g^*(y^\star) \\
&\overset{(55)}{\geq} S_\beta(x; \dot{y}) + \langle A^\top(y^\star - y), x^\star \rangle + g^*(y) - g^*(y^\star) - \gamma b_{\mathcal{X}}(x^\star, \dot{x}) \\
&\geq S_\beta(x; \dot{y}) - \gamma b_{\mathcal{X}}(x^\star, \dot{x}),
\end{aligned}
$$
(56)

where the last inequality holds because g^* is convex and $Ax^\star \in \partial g^*(y^\star)$ due to (9). This proves the first inequality of (16).

Since $b_{\mathcal{Y}}(\cdot, \dot{y})$ is 1-strongly convex with respect to the weighted norm, using the optimality condition of the maximization problem in (12) at $y := y^\star$ and $u := Ax$, we obtain

$$
g_\beta(Ax; \dot{y}) \geq \langle Ax, y^\star \rangle - g^*(y^\star) - \beta b_{\mathcal{Y}}(y^\star, \dot{y}) + \frac{\beta}{2} \|y_\beta^*(Ax; \dot{y}) - y^\star\|_{\mathcal{Y}}^2.
$$
(57)

By (9), we have $-A^\top y^\star \in \partial f(x^\star)$. Using this and the convexity of f, we have $f(x) \geq f(x^\star) - \langle A(x - x^\star), y^\star \rangle$. Summing up the last inequality and (57), then using the definition of $S_\beta(x; \dot{y})$, we obtain

$$
\begin{aligned}
\frac{\beta}{2} \|y_\beta^*(Ax; \dot{y}) - y^\star\|_{\mathcal{Y}}^2 &\leq \beta b_{\mathcal{Y}}(y^\star, \dot{y}) + S_\beta(x; \dot{y}) + g(Ax^\star) + g^*(y^\star) - \langle Ax^\star, y^\star \rangle \\
&\leq \beta b_{\mathcal{Y}}(y^\star, \dot{y}) + S_\beta(x; \dot{y}),
\end{aligned}
$$

which implies the second estimate in (16), where the last inequality is due to the Fenchel–Young equality $g(Ax^\star) + g^*(y^\star) = \langle Ax^\star, y^\star \rangle$ and $Ax^\star \in \partial g^*(y^\star)$.

Now we consider the choice $g(\cdot) := \delta_{\{c\}}(\cdot)$ in the constrained setting (3). Under Assumption A.1, any $w^\star := (x^\star, y^\star) \in \mathcal{W}^\star$ is a saddle point of the Lagrange function $\mathcal{L}(x, y) := f(x) + \langle Ax - c, y \rangle$, i.e., $\mathcal{L}(x^\star, y) \leq \mathcal{L}(x^\star, y^\star) \leq \mathcal{L}(x, y^\star)$ for all $x \in \mathcal{X}$ and $y \in \mathbb{R}^m$. The dual function D in (2) becomes $D(y) := -f^*(-A^\top y) - c^\top y = \min_x \{f(x) + \langle Ax - c, y \rangle\}$. It leads to $D(y) \leq D(y^\star) = f(x^\star) \leq f(x) + \langle y^\star, Ax - c \rangle$, and hence

$$
f(x) - D(y) \geq f(x) - f(x^\star) \geq \langle c - Ax, y^\star \rangle \geq -\|y^\star\|_{\mathcal{Y}} \|Ax - c\|_{\mathcal{Y},*}
$$
(58)

for all $(x, y) \in \mathcal{W}$, which proves (17).

Finally, we prove (18). Indeed, using the definition of g and g_β and $Ax^\star = c$, we can write

$$
\begin{aligned}
f(x) - f(x^\star) &= f(x) + g_\beta(Ax; \dot{y}) - f(x^\star) - g(Ax^\star) - g_\beta(Ax; \dot{y}) + g(Ax^\star) \\
&= S_\beta(x; \dot{y}) - g_\beta(Ax; \dot{y}) + g(Ax^\star) \\
&\overset{(54)}{\leq} S_\beta(x; \dot{y}) - \langle A(x - x^\star), y^\star \rangle + \beta b_{\mathcal{Y}}(y^\star, \dot{y}) \\
&\overset{(56)}{\leq} G_{\gamma\beta}(w; \dot{w}) + \langle c - Ax, y^\star \rangle + \beta b_{\mathcal{Y}}(y^\star, \dot{y}) + \gamma b_{\mathcal{X}}(x^\star, \dot{x}).
\end{aligned}
$$

We then use the second inequality of (58) to get

$$
\begin{aligned}
\langle y^\star, c - Ax \rangle &\leq f(x) - f(x^\star) = S_\beta(x; \dot{y}) - g_\beta(Ax; \dot{y}) + g(Ax^\star) \\
&= S_\beta(x; \dot{y}) - g_\beta(Ax; \dot{y}),
\end{aligned}
$$
(59)

where $g(Ax^\star) = 0$ due to the feasibility of $x^\star$, i.e., $Ax^\star = c$. Now it is obvious that

$$g_\beta(Ax; \dot{y}) := \sup_{\hat{y} \in \mathcal{Y}} \{\langle Ax - c, \hat{y} \rangle - \beta b_{\mathcal{Y}}(\hat{y}, \dot{y})\} \geq \langle Ax - c, y^\star \rangle - \beta b_{\mathcal{Y}}(y^\star, \dot{y}).$$

Hence, combining this estimate and (59), we obtain the first inequality in (18).

As $\nabla b_{\mathcal{Y}}(\cdot, \dot{y})$ is $L_{b_{\mathcal{Y}}}$-Lipschitz continuous, $b_{\mathcal{Y}}(\dot{y}, \dot{y}) = 0$ and $\nabla b_{\mathcal{Y}}(\dot{y}, \dot{y}) = 0$, we have

$$g_\beta(Ax; \dot{y}) = \sup_{\hat{y} \in \mathcal{Y}} \{\langle Ax - c, \hat{y} \rangle - \beta b_{\mathcal{Y}}(\hat{y}, \dot{y})\} \geq \sup_{\hat{y} \in \mathcal{Y}} \left\{ \langle Ax - c, \hat{y} \rangle - \frac{\beta L_{b_{\mathcal{Y}}}}{2} \|\hat{y} - \dot{y}\|_{\mathcal{Y}}^2 \right\}$$
$$= \frac{1}{2\beta L_{b_{\mathcal{Y}}}} \|Ax - c\|_{\mathcal{Y},*}^2 + \langle \dot{y}, Ax - c \rangle.$$

The last equality comes from the formula of the Fenchel conjugate of the squared norm. Combining this inequality and (59), we obtain

$$\langle y^\star, c - Ax \rangle \leq S_\beta(x; \dot{y}) - \frac{1}{2L_{b_y}\beta} \|Ax - c\|_{\mathcal{Y},*}^2 - \langle \dot{y}, Ax - c \rangle.$$

Rearranging this expression and using the Cauchy–Schwarz inequality, we obtain $-\|y^\star - \dot{y}\|_{\mathcal{Y}} \|Ax - c\|_{\mathcal{Y},*} \leq S_\beta(x; \dot{y}) - (2L_{b_{\mathcal{Y}}}\beta)^{-1} \|Ax - c\|_{\mathcal{Y},*}^2$, which leads to

$$\|Ax - c\|_{\mathcal{Y},*}^2 - 2\beta L_{b_{\mathcal{Y}}} \|y^\star - \dot{y}\|_{\mathcal{Y}} \|Ax - c\|_{\mathcal{Y},*} - 2L_{b_{\mathcal{Y}}}\beta S_\beta(x; \dot{y}) \leq 0.$$

Let $t := \|Ax - c\|_{\mathcal{Y},*}$. The last inequality becomes $t^2 - 2\beta L_{b_{\mathcal{Y}}} \|y^\star - \dot{y}\|_{\mathcal{Y}} t - 2L_{b_{\mathcal{Y}}}\beta S_\beta(x; \dot{y}) \leq 0$. This inequation in t has solution. Hence, $\|y^\star - \dot{y}\|_{\mathcal{Y}}^2 + 2L_{b_{\mathcal{Y}}}^{-1}\beta^{-1} S_\beta(x; \dot{y}) \geq 0$ and

$$t := \|Ax - c\|_{\mathcal{Y},*} \leq \beta L_{b_{\mathcal{Y}}} \left[\|y^\star - \dot{y}\|_{\mathcal{Y}} + \left(\|y^\star - \dot{y}\|_*^2 + 2L_{b_{\mathcal{Y}}}^{-1}\beta^{-1} S_\beta(x; \dot{y})\right)^{1/2} \right],$$

which is the second estimate of (18). $\qquad \square$

A.3. The convergence analysis of the ASGARD method. In this appendix, we provide the full convergence analysis of ASGARD. First, we prove a key inequality to guarantee the optimality gap reduction condition.

LEMMA 11. *Let us define* $S_{\beta_k}(\bar{x}^k; \dot{y}) := P_{\beta_k}(\bar{x}^k; \dot{y}) - P^* = f(\bar{x}^k) + g_{\beta_k}(A\bar{x}^k; \dot{y}) - f(x^\star) - g(Ax^\star)$. *If* $\tau_k \in (0, 1]$, *then*

$$S_{\beta_{k+1}}(\bar{x}^{k+1}; \dot{y}) + \frac{\bar{L}_A \tau_k^2}{\beta_{k+1}} \|\tilde{x}^{k+1} - x^\star\|_{\mathcal{X}}^2 \leq (1 - \tau_k) S_{\beta_k}(\bar{x}^k; \dot{y}) + \frac{\bar{L}_A \tau_k^2}{\beta_{k+1}} \|\tilde{x}^k - x^\star\|_{\mathcal{X}}^2$$

$$(60) \qquad\qquad + (1 - \tau_k) \left[(\beta_k - \beta_{k+1}) L_{b_{\mathcal{Y}}} - \beta_{k+1}\tau_k\right] \left\|\nabla g_{\beta_{k+1}}(A\bar{x}^k; \dot{y}) - \dot{y}\right\|_{\mathcal{Y}}^2.$$

Proof. Using Lemma 10 with $h := g$, $h_\beta := g_\beta$, $\mathcal{Z} := \mathcal{Y}$, and $z := Ax$, we can proceed as

$$f(\bar{x}^{k+1}) + g_{\beta_{k+1}}(A\bar{x}^{k+1}; \dot{y}) \overset{(49)}{\leq} f(\bar{x}^{k+1}) + g_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) + \langle \nabla g_{\beta_{k+1}}(A\hat{x}^k; \dot{y}), A\bar{x}^{k+1} - A\hat{x}^k \rangle$$

$$+ \frac{1}{2\beta_{k+1}} \|A\hat{x}^k - A\bar{x}^{k+1}\|_{\mathcal{Y},*}^2$$

$$(61) \qquad \overset{\nabla g_\beta = y_\beta^*}{\leq} f(\bar{x}^{k+1}) + g_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) + \langle A^\top y_{\beta_{k+1}}^*(A\hat{x}^k; \dot{y}), \bar{x}^{k+1} - \hat{x}^k \rangle + \frac{\bar{L}_A}{2\beta_{k+1}} \|\hat{x}^k - \bar{x}^{k+1}\|_{\mathcal{X}}^2$$

$$\overset{\text{def. of } \bar{x}_{k+1}}{\leq} f(x) + g_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) + \langle A^\top y_{\beta_{k+1}}^*(A\hat{x}^k; \dot{y}), x - \hat{x}^k \rangle + \frac{\bar{L}_A}{2\beta_{k+1}} \left[\|\hat{x}^k - x\|_{\mathcal{X}}^2 - \|\bar{x}^{k+1} - x\|_{\mathcal{X}}^2\right],$$

where the last inequality comes from the definition of $\bar{x}^{k+1}$ by using its optimality condition and the functions value at $x \in \mathcal{X}$.

Our next step is to choose $x := (1 - \tau_k)\bar{x}^k + \tau_k x^\star$. In this case, we have

$$x - \hat{x}^k = (1 - \tau_k)\bar{x}^k + \tau_k x^\star - (1 - \tau_k)\bar{x}^k - \tau_k \tilde{x}^k \;\; = \tau_k(x^\star - \tilde{x}^k),$$

$$x - \hat{x}^k = (1 - \tau_k)\bar{x}^k + \tau_k x^\star - (1 - \tau_k)\hat{x}^k - \tau_k \hat{x}^k \;\; = (1 - \tau_k)(\bar{x}^k - \hat{x}^k) + \tau_k(x^\star - \hat{x}^k),$$

$$x - \bar{x}^{k+1} = (1 - \tau_k)\bar{x}^k + \tau_k x^\star - \hat{x}^k - \tau_k(\tilde{x}_{k+1} - \tilde{x}_k) = \tau_k(x^\star - \tilde{x}^{k+1}).$$

Now we plug these expressions into (61), and using the convexity of f, we can derive

$$
\begin{aligned}
f(\bar{x}^{k+1}) + g_{\beta_{k+1}}(A\bar{x}^{k+1}; \dot{y}) &\leq (1 - \tau_k)f(\bar{x}^k) + \tau_k f(x^\star) + g_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) \\
&\quad + \tau_k \langle A^\top y^*_{\beta_{k+1}}(A\hat{x}^k; \dot{y}), x^\star - \hat{x}^k \rangle + (1 - \tau_k)\langle A^\top y^*_{\beta_{k+1}}(A\hat{x}^k; \dot{y}), \bar{x}^k - \hat{x}^k \rangle \\
&\quad + \frac{\bar{L}_A \tau_k^2}{2\beta_{k+1}}\|\tilde{x}^k - x^\star\|^2_{\mathcal{X}} - \frac{\bar{L}_A \tau_k^2}{2\beta_{k+1}}\|\tilde{x}^{k+1} - x^\star\|^2_{\mathcal{X}} \\
&\overset{(50)+(51)}{\leq} (1 - \tau_k)f(\bar{x}^k) + \tau_k f(x^\star) + \tau_k g(Ax^\star) - \tau_k \beta_{k+1} b_{\mathcal{Y}}(\nabla g_{\beta_{k+1}}(A\hat{x}^k; \dot{y}), \dot{y}) \\
&\quad + (1 - \tau_k)g_{\beta_{k+1}}(A\bar{x}^k; \dot{y}) - (1 - \tau_k)\frac{\beta_{k+1}}{2}\|\nabla g_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) - \nabla g_{\beta_{k+1}}(A\bar{x}^k; \dot{y})\|^2_{\mathcal{Y}} \\
&\quad + \frac{\bar{L}_A \tau_k^2}{2\beta_{k+1}}\|\tilde{x}^k - x^\star\|^2_{\mathcal{X}} - \frac{\bar{L}_A \tau_k^2}{2\beta_{k+1}}\|\tilde{x}^{k+1} - x^\star\|^2_{\mathcal{X}} \\
&\overset{(48)}{\leq} (1 - \tau_k)f(\bar{x}^k) + \tau_k f(x^\star) + \tau_k g(Ax^\star) + (1 - \tau_k)g_{\beta_k}(A\bar{x}^k; \dot{y}) \\
&\quad - \frac{\tau_k \beta_{k+1}}{2}\|\nabla g_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) - \dot{y}\|^2_{\mathcal{Y}} + (1 - \tau_k)(\beta_k - \beta_{k+1}) \\
&\quad \quad b_{\mathcal{Y}}(\nabla g_{\beta_{k+1}}(A\bar{x}^k; \dot{y}), \dot{y}) \\
&\quad - (1 - \tau_k)\frac{\beta_{k+1}}{2}\|\nabla g_{\beta_{k+1}}(A\hat{x}^k; \dot{y}) - \nabla g_{\beta_{k+1}}(A\bar{x}^k; \dot{y})\|^2_{\mathcal{Y}} \\
&\quad + \frac{\bar{L}_A \tau_k^2}{2\beta_{k+1}}\|\tilde{x}^k - x^\star\|^2_{\mathcal{X}} - \frac{\bar{L}_A \tau_k^2}{2\beta_{k+1}}\|\tilde{x}^{k+1} - x^\star\|^2_{\mathcal{X}}.
\end{aligned}
$$

By using (52) from Lemma 10, we can further estimate this inequality as

$$
\begin{aligned}
f(\bar{x}^{k+1}) + g_{\beta_{k+1}}(A\bar{x}^{k+1}; \dot{y}) &\overset{(52)}{\leq} (1 - \tau_k)f(\bar{x}^k) + (1 - \tau_k)g_{\beta_k}(A\bar{x}^k; \dot{y}) + \tau_k f(x^\star) + \tau_k g(Ax^\star) \\
&\quad + (\beta_k - \beta_{k+1})(1 - \tau_k)b_{\mathcal{Y}}(\nabla g_{\beta_{k+1}}(A\bar{x}^k; \dot{y}), \dot{y}) \\
&\quad - \frac{\beta_{k+1}}{2}\tau_k(1 - \tau_k)\|\nabla g_{\beta_{k+1}}(A\bar{x}^k; \dot{y}) - \dot{y}\|^2_{\mathcal{Y}} \\
&\quad + \frac{\bar{L}_A \tau_k^2}{2\beta_{k+1}}\|\tilde{x}^k - x^\star\|^2_{\mathcal{X}} - \frac{\bar{L}_A \tau_k^2}{2\beta_{k+1}}\|\tilde{x}^{k+1} - x^\star\|^2_{\mathcal{X}}.
\end{aligned}
$$

Finally, using the $L_{b_{\mathcal{Y}}}$-Lipschitz continuity of $\nabla b_{\mathcal{Y}}$ in the weighted norm $\|\cdot\|_{\mathcal{Y}}$ and the fact that $\nabla b_{\mathcal{Y}}(\dot{y}, \dot{y}) = 0$, we obtain (60) from the last derivation. □

A.3.1. The proof of Lemma 2: Small smoothed primal optimality gap. Let us denote $S_{\beta_k}(\bar{x}^k; \dot{y}) := P_{\beta_{k+1}}(\bar{x}^k; \dot{y}) - P^\star = f(\bar{x}^k) + g_{\beta_k}(A\bar{x}^k; \dot{y}) - f(x^\star) - g(Ax^\star)$. Using (60) from Lemma 11, we have

$$
S_{\beta_{k+1}}(\bar{x}^{k+1}; \dot{y}) + \frac{\bar{L}_A \tau_k^2}{\beta_{k+1}}\|\tilde{x}^{k+1} - x^\star\|^2_{\mathcal{X}} \leq (1 - \tau_k)S_{\beta_k}(\bar{x}^k; \dot{y}) + \frac{\bar{L}_A \tau_k^2}{\beta_{k+1}}\|\tilde{x}^k - x^\star\|^2_{\mathcal{X}}
$$

$$
(62) \qquad + \frac{(1 - \tau_k)}{2}\big[(\beta_k - \beta_{k+1})L_{b_{\mathcal{Y}}} - \beta_{k+1}\tau_k\big]\|\nabla g_{\beta_{k+1}}(A\bar{x}^k; \dot{y}) - \dot{y}\|^2_{\mathcal{Y}}.
$$

In order to remove the last term in this estimate and to get a telescoping sum, we can impose the following conditions:

$$(63) \qquad (\beta_k - \beta_{k+1})L_{b_y} = \beta_{k+1}\tau_k \quad \text{and} \quad (1-\tau_k)\frac{\beta_{k+1}}{\tau_k^2} = \frac{\beta_k}{\tau_{k-1}^2}.$$

By eliminating β_k and β_{k+1} from these equalities, we obtain $\tau_k^2(1+\tau_k/L_{b_y}) = \tau_{k-1}^2(1-\tau_k)$. Hence, we can compute τ_k by solving the cubic equation

$$(64) \qquad p_3(\tau) := \tau^3/L_{b_y} + \tau^2 + \tau_{k-1}^2\tau - \tau_{k-1}^2 = 0.$$

At the same time, we also obtain from (63) an update rule $\beta_{k+1} := \frac{\beta_k}{1+\frac{\tau_k}{L_{b_y}}} < \beta_k$.

Now we show that (64) has a unique positive solution $\tau_k \in (0,1)$ for any $L_{b_y} \geq 1$ and $\tau_{k-1} \in (0,1]$. We consider the cubic polynomial $p_3(\tau)$ defined by the left-hand side of (64). Clearly, for any $\tau > 0$, we have $p_3'(\tau) = 3\tau^2/L_{b_y} + 2\tau + \tau_{k-1}^2 > 0$. Hence, $p_3(\cdot)$ is monotonically increasing on $(0, +\infty)$. In addition, since $p_3(0) = -\tau_{k-1}^2 < 0$ and $p_3(1) = 1/L_{b_y} + 1 > 0$, the equation (64) has only one positive solution $\tau_k \in (0,1)$.

Next we show that $\tau_k \leq \frac{2}{k+2}$. Indeed, by (64), we have $p_3(\tau) \geq \tau^2 + \tau_{k-1}^2\tau - \tau_{k-1}^2 := p_2(\tau)$. Since the unique positive root of $p_2(\tau) = 0$ is $\tilde{\tau}_k := \frac{\tau_{k-1}}{2}(\sqrt{\tau_{k-1}^2+4} - \tau_{k-1})$, we have $p_3(\tau) \geq p_2(\tilde{\tau}_k) = 0$ for $\tau \geq \tilde{\tau}_k$. As $p_3(\tau)$ is monotonically increasing on $\mathbb{R}_+$, its positive solution τ_k must be in $(0, \tilde{\tau}_k]$. Hence, we have $\tau_k \leq \frac{\tau_{k-1}}{2}(\sqrt{\tau_{k-1}^2+4} - \tau_{k-1})$. By induction, we can easily show that $\tau_k \leq \frac{2}{k+2}$.

We show by induction that $\tau_k \geq \frac{1}{k+1}$. First of all, by the choice of τ_0, we have $\tau_0 = 1 \geq \frac{1}{0+1}$. Suppose that $\tau_{k-1} \geq \frac{1}{k}$; we show that $\tau_k \geq \frac{1}{k+1}$. Assume by contradiction that $\tau_k < \frac{1}{k+1}$. Then using (63), we have

$$\frac{1}{k^2} \leq \tau_{k-1}^2 = \tau_k^2 \frac{1+\tau_k/L_{b_y}}{1-\tau_k} < \frac{1}{(k+1)^2}\frac{1+\frac{L_{b_y}^{-1}}{k+1}}{1-\frac{1}{k+1}} = \frac{1}{(k+1)^2}\frac{k+1+L_{b_y}}{k}.$$

This is equivalent to $(k+1)^2 < k(k+1+L_{b_y})$, which contradicts the assumption that $L_{b_y} = 1$ in Lemma 2. Hence, if $\tau_{k-1} \geq \frac{1}{k}$, then we have $\tau_k \geq \frac{1}{k+1}$. We have $\frac{1}{k+1} \leq \tau_k \leq \frac{2}{k+2}$ for $k \geq 0$.

By the update rule $\beta_{k+1} := \frac{\beta_k}{1+\frac{\tau_k}{L_{b_y}}}$ of β_k, we can show that

$$\beta_{k+1} = \frac{\beta_k}{1+\tau_k/L_{b_y}} \leq \beta_k \frac{k+1}{k+1+L_{b_y}^{-1}} \leq \beta_1 \prod_{l=1}^{k} \frac{l+1}{l+1+L_{b_y}^{-1}} = \mathcal{O}\Big(\frac{1}{k^{1/L_{b_y}}}\Big) \underset{k\to\infty}{\longrightarrow} 0.$$

Clearly, if $L_{b_y} = 1$, then $\beta_{k+1} = \frac{\beta_k}{1+\tau_k} \leq \frac{k+1}{k+2}\beta_k \leq \frac{2\beta_1}{k+2}$ by induction.

Finally, we upper bound the ratio τ_k^2/β_{k+1} by using the second equality in (63) as

$$\frac{\tau_k^2}{\beta_{k+1}} = \frac{\tau_{k-1}^2}{\beta_k}(1-\tau_k) = \frac{\tau_0^2}{\beta_1}\prod_{l=1}^{k}(1-\tau_l) \leq \frac{\tau_0^2}{\beta_1}\prod_{l=1}^{k}\Big(1-\frac{1}{l+1}\Big) = \frac{\tau_0^2}{\beta_1}\prod_{l=1}^{k}\frac{l}{l+1} = \frac{\tau_0^2}{\beta_1(k+1)}.$$

Using these relations in (62) and letting $S_k := S_{\beta_k}(\bar{x}^k; \dot{y})$, we obtain

$$\frac{\beta_{k+1}}{\tau_k^2}S_{k+1} + \frac{\bar{L}_A}{2}\|\tilde{x}^{k+1}-x^\star\|_{\mathcal{X}}^2 \leq \frac{\beta_k}{\tau_{k-1}^2}S_k + \frac{\bar{L}_A}{2}\|\tilde{x}^k-x^\star\|_{\mathcal{X}}^2 \leq \frac{\beta_0(1-\tau_0)}{\tau_0^2}S_0 + \frac{\bar{L}_A}{2}\|\tilde{x}^0-x^\star\|_{\mathcal{X}}^2$$

and get (20), noting that $\tau_0 = 1$, the bound on $\frac{\tau_k^2}{\beta_{k+1}}$, and $S_k := P_{\beta_k}(\bar{x}^k; \dot{y}) - P^\star$. $\square$

A.4. The analysis of the update rule (24). If we choose $p_{\mathcal{Y}}(y) := \frac{1}{2}\|y\|_2^2$, then $b_{\mathcal{Y}}(y, \dot{y}) = \frac{1}{2}\|y - \dot{y}\|_2^2$. We can compute $y_\beta^*(u; \dot{y})$ from (13) explicitly as $y_\beta^*(u; \dot{y}) = \dot{y} + \frac{1}{\beta}(u - c)$ and $g_\beta(u; \dot{y})$ from (12) as $g_\beta(u; \dot{y}) = \frac{1}{2\beta}\|u - c\|_2^2 + \langle \dot{y}, u - c \rangle$. Hence, $g_{\beta_{k+1}}(u; \dot{y}) = g_{\beta_k}(u; \dot{y}) + \frac{(\beta_k - \beta_{k+1})}{\beta_{k+1}\beta_k}\|u - c\|_2^2$. Using this relation in the proof of Lemma 11 instead of (48), we obtain

$$S_{\beta_{k+1}}(\bar{x}^{k+1}; \dot{y}) + \frac{\bar{L}_A \tau_k^2}{\beta_{k+1}}\|\tilde{x}^{k+1} - x^\star\|_{\mathcal{X}}^2 \leq (1 - \tau_k)S_{\beta_k}(\bar{x}^k; \dot{y}) + \frac{\bar{L}_A \tau_k^2}{\beta_{k+1}}\|\tilde{x}^k - x^\star\|_{\mathcal{X}}^2$$
$$+ \frac{(1 - \tau_k)}{2\beta_k\beta_{k+1}}[\beta_{k+1} - (1 - \tau_k)\beta_k]\|\nabla g_{\beta_{k+1}}(A\bar{x}^k; \dot{y}) - \dot{y}\|_{\mathcal{Y}}^2.$$

Hence, if we choose $\beta_{k+1} = (1 - \tau_k)\beta_k$, then we can remove the last term in the above estimate. Combining this rule and the second condition of (63), we obtain the update rule (24). □

A.5. The proof of Lemma 5: Gap reduction in ADSGARD. For simplicity of notation, we denote by $f_k^*(y) := f_{\gamma_{k+1}}^*(-A^\top y; \dot{x})$ using (10), $\bar{y}_k^* := y_{\beta_k}^*(A\bar{x}^k; \dot{y})$, $\hat{x}_{k+1}^* := x_{\gamma_{k+1}}^*(\hat{y}^k; \dot{x})$, and $\bar{x}_{k+1}^* := x_{\gamma_{k+1}}^*(\bar{y}^k; \dot{x})$. By (49), ∇f_γ^* is Lipschitz continuous with the Lipschitz constant $L_{f_\gamma^*} := \gamma^{-1}$, and thus ∇f_k^* is Lipschitz continuous with the Lipschitz constant $\gamma_{k+1}^{-1}\bar{L}_A$.

First, using the optimality condition for problem (14), we obtain

$$(65) \qquad f(\bar{x}^k) + \langle A\bar{x}^k, y \rangle - \beta_k b_{\mathcal{Y}}(y, \dot{y}) - g^*(y) \leq P_{\beta_k}(\bar{x}^k; \dot{y}) - (\beta_k/2)\|y - \bar{y}_k^*\|_{\mathcal{Y}}^2.$$

Second, using the definition of $f_\gamma^*(\cdot; \dot{x})$ in (10), we can show that

$$\langle A\hat{x}_{k+1}^*, y \rangle + f(\hat{x}_{k+1}^*) = -\gamma_{k+1}b_{\mathcal{X}}(\hat{x}_{k+1}^*, \dot{x}) - f_k^*(\hat{y}^k) + \langle A\hat{x}_{k+1}^*, y - \hat{y}^k \rangle$$
$$(66) \qquad\qquad = -\gamma_{k+1}b_{\mathcal{X}}(\hat{x}_{k+1}^*, \dot{x}) - f_k^*(\hat{y}^k) - \langle \nabla f_k^*(\hat{y}^k), y - \hat{y}^k \rangle.$$

Third, using (48) for f_γ^* and the inequality (50) of $f_{\gamma_{k+1}}^*(\cdot; \dot{x})$, we can derive

$$-D_{\gamma_k}(\bar{y}^k; \dot{x}) = f_{\gamma_k}^*(-A^\top \bar{y}^k; \dot{x}) + g^*(\bar{y}^k)$$
$$\overset{(48)}{\geq} f_{\gamma_{k+1}}^*(-A^\top \bar{y}^k; \dot{x}) + g^*(\bar{y}^k) - (\gamma_k - \gamma_{k+1})b_{\mathcal{X}}(\bar{x}_{k+1}^*; \dot{x})$$
$$\overset{(50)}{\geq} f_{\gamma_{k+1}}^*(-A^\top \hat{y}^k; \dot{x}) + \langle \nabla f_{\gamma_{k+1}}^*(-A^\top \hat{y}^k; \dot{x}), A^\top(\hat{y}^k - \bar{y}^k) \rangle$$
$$+ \frac{\gamma_{k+1}}{2}\|\nabla f_{\gamma_{k+1}}^*(-A^\top \bar{y}^k; \dot{x}) - \nabla f_{\gamma_{k+1}}^*(-A^\top \hat{y}^k; \dot{x})\|_{\mathcal{X}}^2$$
$$+ g^*(\bar{y}^k) - (\gamma_k - \gamma_{k+1})b_{\mathcal{X}}(\bar{x}_{k+1}^*, \dot{x})$$
$$= f_k^*(\hat{y}^k) + \langle \nabla f_k^*(\hat{y}^k), \bar{y}^k - \hat{y}^k \rangle + \frac{\gamma_{k+1}}{2}\|\bar{x}_{k+1}^* - \hat{x}_{k+1}^*\|_{\mathcal{X}}^2$$
$$(67) \qquad\qquad + g^*(\bar{y}^k) - (\gamma_k - \gamma_{k+1})b_{\mathcal{X}}(\bar{x}_{k+1}^*, \dot{x}).$$

Then by the definition of $\bar{y}^{k+1}$, we can write

$$D_{\gamma_{k+1}}(\bar{y}^{k+1}; \dot{x}) = -g^*(\bar{y}^{k+1}) - f_{\gamma_{k+1}}^*(-A^\top \bar{y}^{k+1}, \dot{x})$$
$$\geq -g^*(\bar{y}^{k+1}) - f_k^*(\hat{y}^k) - \langle \nabla f_k^*(\hat{y}^k), \bar{y}^{k+1} - \hat{y}^k \rangle - \frac{\bar{L}_A}{2\gamma_{k+1}}\|\bar{y}^{k+1} - \hat{y}^k\|_{\mathcal{Y}}^2$$
$$(68) \qquad = -\min_{u \in \mathcal{Y}}\left\{ g^*(u) + f_k^*(\hat{y}^k) + \langle \nabla f_k^*(\hat{y}^k), u - \hat{y}^k \rangle + \frac{\bar{L}_A}{2\gamma_{k+1}}\|u - \hat{y}^k\|_{\mathcal{Y}}^2 \right\}.$$

Using these relations, the definition of $\bar{x}^{k+1}$, and the convexity of f, we have

$$P_{\beta_{k+1}}(\bar{x}^{k+1}; \dot{y}) = f(\bar{x}^{k+1}) + \max_{y \in \mathcal{Y}} \left\{ \langle A\bar{x}^{k+1}, y \rangle - g^*(y) - \beta_{k+1}b_{\mathcal{Y}}(y, \dot{y}) \right\}$$

$$\overset{(25)}{\leq} \max_{y \in \mathcal{Y}} \left\{ (1-\tau_k) \left[f(\bar{x}^k) + \langle A\bar{x}^k, y \rangle - \beta_k b_{\mathcal{Y}}(y, \dot{y}) - g^*(y) \right] \right.$$
$$\left. + \tau_k \left[\langle A\hat{x}_{k+1}^*, y \rangle + f(\hat{x}_{k+1}^*) - g^*(y) \right] \right\}$$

$$\overset{(65)+(66)}{\leq} (1-\tau_k)P_{\beta_k}(\bar{x}^k; \dot{y}) - \tau_k\gamma_{k+1}b_{\mathcal{X}}(\hat{x}_{k+1}^*, \dot{x})$$
$$- \min_{y \in \mathcal{Y}} \left\{ \tau_k f_k^*(\hat{y}^k) + \tau_k\langle \nabla f_k^*(\hat{y}^k), y - \hat{y}^k \rangle + \frac{(1-\tau_k)\beta_k}{2}\|y - \bar{y}_k^*\|_{\mathcal{Y}}^2 + \tau_k g^*(y) \right\}$$

$$\overset{(67)}{\leq} (1-\tau_k) \left[P_{\beta_k}(\bar{x}^k; \dot{y}) - D_{\gamma_k}(\bar{y}^k; \dot{x}) \right] - \tau_k\gamma_{k+1}b_{\mathcal{X}}(\hat{x}_{k+1}^*, \dot{x})$$
$$- \frac{(1-\tau_k)\gamma_{k+1}}{2}\|\bar{x}_{k+1}^* - \hat{x}_{k+1}^*\|_{\mathcal{X}}^2 + (1-\tau_k)(\gamma_k - \gamma_{k+1})b_{\mathcal{X}}(\bar{x}_{k+1}^*, \dot{x})$$
$$- \min_{y \in \mathcal{Y}} \left\{ f_k^*(\hat{y}^k) + \langle \nabla f_k^*(\hat{y}^k), (1-\tau_k)\bar{y}^k + \tau_k y - \hat{y}^k \rangle \right.$$
$$\left. + \frac{(1-\tau_k)\beta_k}{2}\|y - \bar{y}_k^*\|_{\mathcal{Y}}^2 + g^*((1-\tau_k)\bar{y}^k + \tau_k y) \right\}.$$

Let us define the auxiliary term $\mathcal{T}_k$ as

$$(69) \qquad \begin{aligned} \mathcal{T}_k &:= (1-\tau_k)\frac{\gamma_{k+1}}{2}\|\bar{x}_{k+1}^* - \hat{x}_{k+1}^*\|_{\mathcal{X}}^2 - (1-\tau_k)(\gamma_k - \gamma_{k+1})b_{\mathcal{X}}(\bar{x}_{k+1}^*, \dot{x}) \\ &\quad + \tau_k\gamma_{k+1}b_{\mathcal{X}}(\hat{x}_{k+1}^*, \dot{x}). \end{aligned}$$

Now we consider the change of variable $u := (1-\tau_k)\bar{y}^k + \tau_k y$ for $y \in \mathcal{Y}$. Then $u \in \mathcal{Y}$ and $u - \hat{y}^k = \tau_k(y - \bar{y}_k^*)$. We have

$$P_{\beta_{k+1}}(\bar{x}^{k+1}; \dot{y}) \leq (1-\tau_k)G_{\gamma_k\beta_k}(\bar{w}^k; \dot{w}) - \mathcal{T}_k$$
$$- \min_{u \in \mathcal{Y}} \left\{ f_k^*(\hat{y}^k) + \langle \nabla f_k^*(\hat{y}^k), u - \hat{y}^k \rangle + \frac{(1-\tau_k)\beta_k}{2\tau_k^2}\|u - \hat{y}^k\|_{\mathcal{Y}}^2 + g^*(u) \right\}$$

$$(70) \qquad \overset{(68)+(25)}{\leq} (1-\tau_k)G_{\gamma_k\beta_k}(\bar{w}^k; \dot{w}) + D_{\gamma_{k+1}}(\bar{y}^{k+1}; \dot{x}) - \mathcal{T}_k.$$

Finally, we estimate $\mathcal{T}_k$ in (69) using the strong convexity of $b_{\mathcal{X}}(\cdot, \dot{x})$ as follows:

$$2\mathcal{T}_k \geq (1-\tau_k)\gamma_{k+1}\|\bar{x}_{k+1}^* - \hat{x}_{k+1}^*\|_{\mathcal{X}}^2 + \tau_k\gamma_{k+1}\|\hat{x}_{k+1}^* - \dot{x}\|_{\mathcal{X}}^2$$
$$- (1-\tau_k)(\gamma_k - \gamma_{k+1})L_{b_{\mathcal{X}}}\|\bar{x}_{k+1}^* - \dot{x}\|_{\mathcal{X}}^2$$
$$\overset{(52)}{\geq} (1-\tau_k)\left[\tau_k\gamma_{k+1} - (\gamma_k - \gamma_{k+1})L_{b_{\mathcal{X}}}\right]\|\bar{x}_{k+1}^* - \dot{x}\|_{\mathcal{X}}^2$$

$$(71) \qquad \overset{(25)}{\geq} 0.$$

Substituting (71) into (70), we get $G_{\gamma_{k+1}\beta_{k+1}}(\bar{w}^{k+1}; \dot{w}) \leq (1-\tau_k)G_{\gamma_k\beta_k}(\bar{w}^k; \dot{w})$.

Note that this is valid for all $k \geq 1$. Using similar ideas together with the relations $\bar{x}^1 = \hat{x}_1^*$ and $\hat{y}^0 = \bar{y}_0^*$, we also get

$$G_{\gamma_1,\beta_1}(\bar{w}^1; \dot{w}) \leq -\gamma_1 b_{\mathcal{X}}(\bar{x}^1, \dot{x}) + \frac{\bar{L}_A}{2\gamma_1}\|\bar{y}_1^* - \bar{y}_0^*\|_{\mathcal{Y}}^2 - \beta_1 b_{\mathcal{Y}}(\bar{y}_1^*, \dot{y}).$$

As $\beta_1\gamma_1 \geq \bar{L}_A$ and $\bar{y}_0^* := \dot{y}$, we obtain $G_{\gamma_1\beta_1}(\bar{w}^1; \dot{w}) \leq 0$.

Next we set the equality in three conditions of (25) to get $\gamma_{k+1} = \gamma_k(1 + \tau_k/L_{b_\mathcal{X}})^{-1}$, $\beta_{k+1} = (1-\tau_k)\beta_k$, and $(1-\tau_k)\beta_k\gamma_{k+1} = \tau_k^2\bar{L}_A$. In particular, $\gamma_{k+1}\beta_{k+1} = \tau_k^2\bar{L}_A$, and thus $\gamma_1\beta_1 = \bar{L}_A$. By eliminating γ_k and β_k, we obtain $\tau_k^3/L_{b_\mathcal{X}} + \tau_k^2 + \tau_{k-1}^2\tau_k - \tau_{k-1}^2 = 0$. Hence, similar to the proof of Lemma 2, we can show that $\tau_k \in (0,1)$ is the unique positive solution of the cubic equation $p_3(\tau) := \tau^3/L_{b_\mathcal{X}} + \tau^2 + \tau_{k-1}^2\tau - \tau_{k-1}^2 = 0$. In addition, $\frac{1}{k+1} \le \tau_k \le \frac{2}{k+2}$ for $k \ge 1$ and $\tau_0 = 1$. If $L_{b_\mathcal{X}} = 1$, then $\gamma_{k+1} = \frac{\gamma_k}{1+\tau_k} \le \frac{\gamma_k(k+1)}{k+2} \le \frac{2\gamma_1}{k+2}$. Similarly, $\beta_{k+1} = (1-\tau_k)\beta_k \le \frac{k}{k+1}\beta_k \le \frac{\beta_1}{k+1}$. Finally, we note that $\beta_{k+1} = \frac{\tau_k^2\bar{L}_A}{\gamma_{k+1}} \ge \frac{\bar{L}_A}{(k+1)^2}\frac{k+2}{2\gamma_1} \ge \frac{\bar{L}_A}{2\gamma_1(k+1)}$. □

A.6. The proof of Proposition 8: The accelerated augmented Lagrangian method. First of all, with the choice of norm associated with the Lagrangian smoother, we have

$$\bar{L}_A := \|A\|^2 = \max_{x\in\mathbb{R}^n}\left\{\frac{\|Ax\|_{\mathcal{Y},*}^2}{\|x\|_{\mathcal{X}}^2}\right\} = \max_{x\in\mathbb{R}^n}\left\{\frac{\|Ax\|_{\mathcal{Y},*}^2}{\|Ax\|_{\mathcal{Y},*}^2}\right\} = 1.$$

Next note that the conclusions of Lemma 5 are valid for any seminorm. In particular, if we choose $\beta_1\gamma_0 \ge \bar{L}_A = 1$,

$$\gamma_{k+1} = \gamma_0 \ge \frac{\gamma_0}{1+\tau_k/L_{b_\mathcal{X}}}, \quad \beta_{k+1} = (1-\tau_k)\beta_k, \quad \text{and} \quad \frac{\bar{L}_A}{\gamma_0} = \frac{(1-\tau_k)\beta_k}{\tau_k},$$

then $G_{\gamma_0,\beta_{k+1}}(\bar{w}^{k+1};\dot{w}) \le (1-\tau_k)G_{\gamma_0,\beta_k}(\bar{w}^k;\dot{w}) \le 0$.

Eliminating β_{k+1} and β_k in these equalities, we get $\frac{\tau_{k+1}^2}{1-\tau_{k+1}} = \tau_k^2$. One can easily check by induction that $\beta_k = \beta_1\prod_{l=1}^k(1-\tau_l) = \beta_1\frac{\tau_k^2}{\tau_0^2} = \frac{\tau_k^2}{\gamma_0}$ and $\tau_k \le \frac{2}{k+2}$. We then conclude, using Lemma 1 and the fact that $b_\mathcal{X}(x^\star,\dot{x}) = 0$, that

$$S_{\beta_k}(\bar{x}^k;\dot{y}) \le G_{\gamma_0\beta_k}(\bar{w}^k;\dot{w}) \le 0,$$
$$\|A\bar{x}^k - c\|_{\mathcal{Y},*} \le \beta_k L_{b_\mathcal{Y}}\left[\|y^\star - \dot{y}\|_\mathcal{Y} + \left(\|y^\star - \dot{y}\|_\mathcal{Y}^2 + 2L_{b_\mathcal{Y}}^{-1}\beta_k^{-1}S_{\beta_k}(\bar{x}^k;\dot{y})\right)^{1/2}\right]$$
$$\le \frac{8L_{b_\mathcal{Y}}\|y^\star - \dot{y}\|_\mathcal{Y}}{\gamma_0(k+2)^2},$$

and

$$f(\bar{x}^k) - f^\star \le S_{\beta_k}(\bar{x}^k;\dot{y}) - \langle y^\star, A\bar{x}^k - c\rangle + \beta_k b_\mathcal{Y}(y^\star,\dot{y})$$
$$\le \|y^\star\|_\mathcal{Y}\|A\bar{x}^k - c\|_{\mathcal{Y},*} + \beta_k b_\mathcal{Y}(y^\star,\dot{y}) \le \frac{8L_{b_\mathcal{Y}}\|y^\star\|_\mathcal{Y}\|y^\star - \dot{y}\|_\mathcal{Y} + 4b_\mathcal{Y}(y^\star,\dot{y})}{\gamma_0(k+2)^2},$$
$$f(\bar{x}^k) - f^\star \ge -\|y^\star\|_\mathcal{Y}\|Ax - c\|_{\mathcal{Y},*} \ge -\frac{8L_{b_\mathcal{Y}}\|y^\star\|_\mathcal{Y}\|y^\star - \dot{y}\|_\mathcal{Y}}{\gamma_0(k+2)^2}.$$

The proposition is proved. □

A.7. The proof of Proposition 9: The strongly convex objective case. The proof follows the same arguments as the proof of Lemma 5. We only need to replace the Lipschitz continuity coefficient $\frac{\bar{L}_A}{\gamma_{k+1}}$ by $L_{f_A^*} = \frac{\bar{L}_A}{\mu_f}$ in (68) and replace all other occurrences of γ_{k+1} by zero. Under a choice of parameters satisfying (34), we obtain the gap reduction condition

$$G_{0,\beta_{k+1}}(\bar{w}^{k+1};\dot{w}) \le (1-\tau_k)G_{0,\beta_k}(\bar{w}^k;\dot{w}) \le 0$$

as in Lemma 5. We can also check by induction that $\beta_k \le \frac{4}{(k+2)^2}\frac{\bar{L}_A}{\mu_f}$. Hence, we obtain the conclusion of Proposition 9 by using Lemma 1. □

Acknowledgments. We are thankful Alacaoglu Ahmet, Baran Gozcü, and Alp Yurtsever for their help on the last numerical example and Van Quang Nguyen for his careful proofreading.

REFERENCES

[1] A. AUSLENDER, *Optimisation: Méthodes Numériques*, Masson, Paris, 1976.

[2] H. BAUSCHKE AND P. COMBETTES, *Convex Analysis and Monotone Operators Theory in Hilbert Spaces*, Springer-Verlag, Berlin, 2011.

[3] A. BECK AND M. TEBOULLE, *A fast iterative shrinkage-thresholding agorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[4] A. BECK AND M. TEBOULLE, *Smoothing and first order methods: A unified framework*, SIAM J. Optim., 22 (2012), pp. 557–580.

[5] A. BECK AND M. TEBOULLE, *A fast dual proximal gradient algorithm for convex minimization and applications*, Oper. Res. Lett., 42 (2014), pp. 1–6.

[6] A. BELLONI, V. CHERNOZHUKOV, AND L. WANG, *Square-root LASSO: Pivotal recovery of sparse signals via conic programming*, Biometrika, 94 (2011), pp. 791–806.

[7] A. BEN-TAL AND A. NEMIROVSKI, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS/SIAM Series on Optimization 3, SIAM, Philadelphia, PA, 2001.

[8] D. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and Distributed Computation: Numerical Methods*, Prentice Hall, Englewood Chiffs, NJ, 1989.

[9] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Athena Scientific, Nashua, NH, 1996.

[10] R. I. BOŢ AND C. HENDRICH, *A variable smoothing algorithm for solving convex optimization problems*, TOP, 23 (2012), pp. 124–150.

[11] S. BOYD, N. PARIKH, E. CHU, B. PELEATO, AND J. ECKSTEIN, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), pp. 1–122.

[12] S. BOYD AND L. VANDENBERGHE, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[13] X. CAI, D. HAN, AND X. YUAN, *On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function*, Comput. Optim. Appl., (2016), pp. 1–35.

[14] V. CEVHER, S. BECKER, AND M. SCHMIDT, *Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics*, IEEE Signal Process. Mag., 31 (2014), pp. 32–43.

[15] A. CHAMBOLLE AND T. POCK, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145.

[16] V. CHANDRASEKARANM, B. RECHT, P. A. PARRILO, AND A. S. WILLSKY, *The convex geometry of linear inverse problems*, Found. Comput. Math., 12 (2012), pp. 805–849.

[17] C. CHEN, B. HE, Y. YE, AND X. YUAN, *The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent*, Math. Program., 155 (2016), pp. 57–79.

[18] G. CHEN AND M. TEBOULLE, *A proximal-based decomposition method for convex minimization problems*, Math. Program., 64 (1994), pp. 81–101.

[19] Y. CHEN, G. LAN, AND Y. OUYANG, *Optimal primal-dual methods for a class of saddle-point problems*, SIAM J. Optim., 24 (2014), pp. 1779–1814.

[20] P. COMBETTES AND P. J.-C., *Signal recovery by proximal forward-backward splitting*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer-Verlag, Berlin, 2011, pp. 185–212.

[21] P. L. COMBETTES AND J.-C. PESQUET, *Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators*, Set-Valued Var. Anal., 20 (2012), pp. 307–330.

[22] L. CONDAT, *A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms*, J. Optim. Theory Appl., 158 (2013), pp. 460–479.

[23] D. DAVIS, *Convergence rate analysis of primal-dual splitting schemes*, SIAM J. Optim., 25 (2015), pp. 1912–1943.

[24] D. DAVIS, *Convergence rate analysis of the forward-Douglas-Rachford splitting scheme*, SIAM J. Optim., 25 (2015), pp. 1760–1786.

[25] D. DAVIS AND W. YIN, *Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions*, Math. Oper. Res., (2014).

[26] W. Deng, M.-J. Lai, Z. Peng, and W. Yin, *Parallel multi-block ADMM with o(1/k) convergence*, J. Scientific Computing, https://doi.org/10.1007/s10915-s016-s0318-s2, 2016.

[27] W. Deng and W. Yin, *On the global and linear convergence of the generalized alternating direction method of multipliers*, J. Sci. Comput., 66 (2016), pp. 889–916.

[28] J. E. Esser, *Primal-Dual Algorithm for Convex Models and Applications to Image Restoration, Registration and Nonlocal Inpainting*, Ph.D. thesis, University of California, Los Angeles, Los Angeles, CA, 2010.

[29] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, vols. 1–2, Springer-Verlag, Berlin, 2003.

[30] O. Fercoq and Z. Qu, *Restarting Accelerated Gradient Methods with a Rough Strong Convexity Estimate*, preprint, arXiv:1609.07358, 2016.

[31] P. Giselsson and S. Boyd, *Monotonicity and restart in fast gradient methods*, in IEEE Conference on Decision and Control, CDC, Los Angeles, CA, December 2014.

[32] T. Goldstein, E. Esser, and R. Baraniuk, *Adaptive Primal-Dual Hybrid Gradient Methods for Saddle Point Problems*, Technical report, 2013, pp. 1–26, arXiv:1305.0546v1.pdf.

[33] B. He and X. Yuan, *Convergence analysis of primal-dual algorithms for saddle-point problem: From contraction perspective*, SIAM J. Imaging Sci., 5 (2012), pp. 119–149.

[34] B. He and X. Yuan, *On non-ergodic convergence rate of Douglas–Rachford alternating direction method of multipliers*, Numer. Math., 130 (2012), pp. 567–577.

[35] B. He and X. Yuan, *On the O(1/n) convergence rate of the Douglas-Rachford alternating direction method*, SIAM J. Numer. Anal., 50 (2012), pp. 700–709.

[36] Y. He and R.-D. Monteiro, *An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems*, SIAM J. Optim., 26 (2016), pp. 29–56.

[37] F. Knoll, C. Clason, C. Diwoky, and R. Stollberger, *Adapted random sampling patterns for accelerated MRI*, Magnetic Resonance Mat. Phys. Biol. Med., 24 (2011), pp. 43–50.

[38] G. Lan and R. Monteiro, *Iteration complexity of first-order penalty methods for convex programming*, Math. Program., 138 (2013), pp. 115–139.

[39] G. Lan and R. Monteiro, *Iteration-complexity of first-order augmented Lagrangian methods for convex programming*, Math. Program., 155 (2016), pp. 511–547.

[40] T. Lin, S. Ma, and S. Zhang, *Iteration complexity analysis of multi-block ADMM for a family of convex minimization without strong convexity*, J. Sci. Comput., (2015), pp. 1–30.

[41] T. Lin, S. Ma, and S. Zhang, *On the global linear convergence of the admm with multiblock variables*, SIAM J. Optim., 25 (2015), pp. 1478–1497.

[42] T. Lin, S. Ma, and S.-Z. Zhang, *On the sublinear convergence rate of multi-block ADMM*, J. Oper. Res. Soc. China, 3 (2015), pp. 251–274.

[43] Y. Malitsky and T. Pock, *A First-Order Primal-Dual Algorithm with Linesearch*, preprint, arXiv:1608.08883, 2016.

[44] M. B. McCoy, V. Cevher, Q. Tran-Dinh, A. Asaei, and L. Baldassarre, *Convexity in source separation: Models, geometry, and algorithms*, IEEE Signal Process. Mag., 31 (2014), pp. 87–95.

[45] R. Monteiro and B. Svaiter, *On the complexity of the hybrid proximal extragradient method for the interates and the ergodic mean*, SIAM J. Optim., 20 (2010), pp. 2755–2787.

[46] R. Monteiro and B. Svaiter, *Complexity of variants of Tseng's modified F-B splitting and Korpelevicha's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems*, SIAM J. Optim., 21 (2011), pp. 1688–1720.

[47] R. Monteiro and B. Svaiter, *Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers*, SIAM J. Optim., 23 (2013), pp. 475–507.

[48] R. Monteiro and B. Svaiter, *Iteration-complexity of block-decomposition algorithms and the alternating minimization augmented Lagrangian method*, SIAM J. Optim., 23 (2013), pp. 475–507.

[49] I. Necoara and A. Patrascu, *Iteration complexity analysis of dual first order methods for convex programming*, J. Optim. Theory Appl., arXiv:1409.1462, 2014.

[50] I. Necoara and J. Suykens, *Applications of a smoothing technique to decomposition in convex optimization*, IEEE Trans. Automat. Control, 53 (2008), pp. 2674–2679.

[51] V. Nedelcu, 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.

[52] A. Nemirovskii, *Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM J. Optim., 15 (2004), pp. 229–251.

[53] Y. Nesterov, *A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$*, Dokl. Akad. Nank, 269 (1983), pp. 543–547.

[54] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Applied Optimization 87, Kluwer Academic Publishers, Dordrecht, 2004.

[55] Y. Nesterov, *Excessive gap technique in nonsmooth convex minimization*, SIAM J. Optim., 16 (2005), pp. 235–249.

[56] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.

[57] Y. Nesterov, *Dual extrapolation and its applications to solving variational inequalities and related problems*, Math. Program., 109 (2007), pp. 319–344.

[58] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed., Springer Series in Operations Research and Financial Engineering, Springer, Berlin, 2006.

[59] B. O'Donoghue and E. Candes, *Adaptive restart for accelerated gradient schemes*, Found. Comput. Math., 15 (2015), pp. 715–732.

[60] H. Ouyang, N. He, L. Q. Tran, and A. Gray, *Stochastic alternating direction method of multipliers*, JMLR Workshop Conf. Proc., 28 (2013), pp. 80–88.

[61] Y. Ouyang, Y. Chen, G. Lan, and E. J. Pasiliao, *An accelerated linearized alternating direction method of multiplier*, SIAM J. Imaging Sci., 8 (2015), pp. 644–681.

[62] R. Rockafellar and R. Wets, *Variational Analysis*, Vol. 317, Springer, Berlin, 2004.

[63] R. T. Rockafellar, *Convex Analysis*, Princeton Mathematics Series 28, Princeton University Press, Princeton, NJ, 1970.

[64] 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.

[65] M. Solodov and B. Svaiter, *A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator*, Set-Valued Var. Anal., 7 (1999), pp. 323–345.

[66] W. Su, S. Boyd, and E. Candes, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, in Advances in Neural Information Processing Systems (NIPS), 2014, pp. 2510–2518.

[67] Q. Tran-Dinh and V. Cevher, *Constrained convex minimization via model-based excessive gap*, in Proceeding of the Neural Information Processing Systems (NIPS), Vol. 27, Montreal, Canada, December 2014, pp. 721–729.

[68] Q. Tran-Dinh and V. Cevher, *A Primal-Dual Algorithmic Framework for Constrained Convex Minimization*, Technical report, LIONS, 2014, pp. 1–54.

[69] Q. Tran-Dinh and V. Cevher, *Smooth Alternating Direction Methods for Nonsmooth Constrained Convex Optimization*, Technical report, LIONS, EPFL, 2015, arXiv:1507.03734.

[70] Q. Tran-Dinh, A. Kyrillidis, and V. Cevher, *Composite self-concordant minimization*, J. Mach. Learn. Res., 15 (2015), pp. 374–416.

[71] P. Tseng, *Applications of splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM J. Control Optim., 29 (1991), pp. 119–138.

[72] B. C. Vu, *A variable metric extension of the forward–backward–forward algorithm for monotone operators*, Numer. Funct. Anal. Optim., 34 (2013), pp. 1050–1065.

[73] M. J. Wainwright, *Structured regularizers for high-dimensional problems: Statistical and computational issues*, Annu. Rev. Stat. Appl., 1 (2014), pp. 233–253.

[74] X. Wang, M. Hong, S. Ma, and Z.-Q. Luo, *Solving multiple-block separable convex minimization problems using two-block alternating direction method of multipliers*, preprint, arXiv:1308.5294, 2013.

[75] E. Wei and A. Ozdaglar, *On the $\mathcal{O}(1/k)$-convergence of asynchronous distributed alternating direction method of multipliers*, in Global Conference on Signal and Information Processing (GlobalSIP), IEEE, 2013, pp. 551–554.

[76] Y. Xu, *Accelerated First-Order Primal-Dual Proximal Methods for Linearly Constrained Composite Convex Programming*, preprint, arXiv:1606.09155, 2016.