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

QUOC TRAN-DINH∗, OLIVIER FERCOQ†, AND VOLKAN CEVHER‡

Abstract. We propose a new first-order primal-dual optimization framework for a convex optimization template with broad applications. Our optimization algorithms feature optimal convergence guarantees under a variety of common structure assumptions on the problem template. 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 non-smooth 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 numerical evidence with two examples and illustrate that the new methods can outperform the state-of-the-art, including Chambolle-Pock, and the alternating direction method-of-multipliers algorithms.

Keywords: 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

1. Introduction. We propose 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:

\[ P^* := \min_{x \in \mathcal{X}} \left\{ P(x) := f(x) + g(Ax) \right\}, \]  

(1.1)

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; \( \mathcal{X} = \text{dom}(P) \) is the domain of \( P \), 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 (1.1) as a nonsmooth composite minimization problem.

Associated with the primal problem (1.1), we define the following dual formulation:

\[ D^* := \max_{y \in \mathcal{Y}} \left\{ D(y) := -f^*(-A^\top y) - g^*(y) \right\}, \]

(1.2)

where \( f^* (\cdot) \) and \( g^* \) are the Fenchel conjugate of \( f \) and \( g \), respectively, and \( \mathcal{Y} = \text{dom}(D) \) is the domain of \( D \). Clearly, (1.2) has exactly the same form as (1.1) in the dual space.

The templates (1.1)-(1.2) provide a unified formulation for a broad set of applications in various disciplines, see, e.g., [6, 10, 12, 14, 36, 45, 57]. While problem (1.1) is presented in the unconstrained form, it automatically covers constrained settings by means of indicator functions. For example, (1.1) covers the following prototypical optimization template via \( g(z) := \delta_{\{c\}}(z) \) (i.e., the indicator function of the convex set \( \{c\} \)):

\[ f^* := \min_{x \in \mathcal{X}} \left\{ f(x) + \delta_{\{c\}}(Ax) \right\} \equiv \min_{x \in \mathcal{X}} \left\{ f(x) : Ax = c \right\}, \]

(1.3)

where \( f \) is a proper, closed and convex function as in (1.1). Note that (1.3) is sufficiently general to cover standard convex optimization subclasses, such as conic programming, monotropic programming, and geometric programming, as specific instances [5, 17, 9].

Among classical convex optimization methods, the primal-dual approach is perhaps one of the best candidates to solve the primal-dual pair (1.1)-(1.2). Theory and methods along this

*Department of Statistics and Operations Research, University of North Carolina at Chapel Hill (UNC), USA. Email: quoctd@email.unc.edu.
†LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France. E-mail: olivier.fercoq@telecom-paristech.fr.
‡Laboratory for Information and Inference Systems (LIONS), École Polytechnique Fédérale de Lausanne (EPFL), CH1015 - Lausanne, Switzerland. E-mail: volkan.cevher@epfl.ch.
approach have been developed for several decades and have led to various algorithms, see, e.g., [1, 9, 13, 15, 16, 17, 18, 19, 20, 21, 23, 25, 28, 29, 31, 32, 35, 37, 39, 43, 47, 48, 50, 56, 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 nonsmooth composite minimization that have the best convergence rate guarantees?" To our knowledge, this question have never been addressed fully unattended first-order primal-dual methods for nonsmooth composite minimization that have the best convergence rate guarantees?" To our knowledge, this question have never been addressed fully unpredictably computational costs. A vast list of references can be found, e.g., in [13, 50].

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 references can be found, e.g., in [13, 50].

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 convergence rate guarantees?" To our knowledge, this question have never been addressed fully in a unified fashion in this generality. It is obvious that our theory presented in this paper is still applicable to the smooth cases of \( f \) without requiring neither Lipschitz gradient nor strongly convex-type assumption. Such a model covers several important applications such as graphical learning models and Poisson imaging reconstruction [55].

1.1. Our approach. Associated with the primal problem (1.1) and the dual one (1.2), we define

\[
G(w) := P(x) - D(y),
\]  

(1.4)

as a primal-dual gap function, where \( w \) is the primal-dual variable. The gap function \( G \) in (1.4) is convex in terms of the concatenated primal-dual variable \( w := (x, y) \). Under strong duality, we have \( G(w^*) = 0 \) if and only if \( w^* = (x^*, y^*) \) is a primal-dual solution of (1.1) and (1.2).

The gap function (1.4) is widely used in convex optimization and variational inequalities, see, e.g., [24]. 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 [13, 29, 48].

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 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 \( g \) [3, 44]. In the sequel, we denote the smoothed gap function \( G_{\gamma\beta}(w) := P_\beta(x) - D_\gamma(y) \) to \( G(w) \), where \( P_\beta \) is a smoothed approximation to \( P \) depending on the smoothness parameter \( \beta > 0 \), and \( D_\gamma \) is a smoothed approximation to \( D \) depending on the smoothness parameter \( \gamma > 0 \). By smoothed approximation, we mean the same max-form approximation as [44]. However, it is still unclear how to properly choose these smoothness parameters.

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

\[
G_{\gamma k + 1, \beta k + 1}(\bar{w}^{k+1}) \leq (1 - \tau_k)G_{\gamma k, \beta k}(\bar{w}^{k}) + \psi_k, \tag{1.5}
\]

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 [43] is a special case of the gap reduction model (1.5) when \( \psi_k \leq 0 \).

Homotopy: We will design algorithms to maintain (1.5) while simultaneously updating \( \beta_k \), \( \gamma_k \) and \( \tau_k \) to zero to achieve the optimal convergence rate based on the assumptions on the
problem template. This strategy will also allow our theoretical guarantees not to depend on the diameter of the feasible sets. A similar technique is also proposed in [43], but only for symmetric primal-dual methods. It is also used in conjunction with Nesterov’s smoothing technique in [8] for unconstrained problem but had only the $O(\ln(k)/k)$ 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 the primal and dual problems, which may not applicable to (1.1) or (1.3) due to the unboundedness of the dual feasible domain.

1.2. Our contributions. To this end, the main contributions of this paper can be summarized as follows, which consists of three parts:

(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 optimal first-order primal-dual methods that can operate in a black-box fashion (in the sense of [44]). Our analysis technique unifies several classical concepts in convex optimization, from Auslander’s gap function and Nesterov’s smoothing technique 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 (1.3), 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 (1.1) and (1.3). The first algorithm requires to perform only one primal step and one dual step without using 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 the $O(1/k)$ convergence rate on the objective residual $P(\bar{x}^k) - P^*$ of (1.1) for both algorithms, which is the best known in the literature for the fully nonsmooth setting case. For the constrained case (1.3), we also prove the convergence of both algorithms in terms of the primal objective residual and the feasibility violation, both achieve the $O(1/k)$ convergence rate, and are independent of the prox-diameters compared to existing smoothing techniques [3, 43, 44].

(c) (Special cases) We illustrate that the new techniques enable us to exploit additional structures, including the augmented Lagrangian smoothing, 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 in [4, 13, 22, 28, 29, 48, 50] thanks to the separation of the convergence rates for primal optimality and the feasibility. We believe this is important since the separate constraint feasibility guarantee can act as a consensus rate in distributed optimization. Second, our assumptions cover a much broader class of problems: we can trade-off primal optimality and constraint feasibility without any heuristic strategy, and our convergence rates seem to be the best known rate for the class of fully nonsmooth problems (1.1) so far. Third, our augmented Lagrangian algorithm generates simultaneously both the primal-dual sequence compared to existing augmented Lagrangian algorithms, while it maintains its $O\left(\frac{1}{k^2}\right)$-optimal 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 constraint components, such as strong convexity. Fifth, this work significantly expands on our earlier conference work [52] not only with new methods but also by demonstrating the impact of warm-start and restart. Finally, our follow up work [54] also demonstrates how our analysis framework and uncertainty principles extend to cover alternating direction optimization methods.

1.3. Paper organization. In Section 2 we propose a smoothing technique with proximity functions for (1.1) 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 \((1.1) - (1.2)\) and its convergence theory. Section 4 provides the second primal-dual algorithmic framework using averaged sequences for solving \((1.1) - (1.2)\) and its convergence theory. A comparison between our approach and existing methods is given in Section 7. For clarity of exposition, technical proofs of the results in the main text are moved to the appendix.

2. Smoothed gap function and optimality characterization. We propose to smooth the primal-dual gap function defined by \((1.4)\) by proximity functions. Then, we provide a key lemma to characterize the optimality condition for \((1.1)\) and \((1.2)\).

2.1. Basic notation. We use \(\|x\|_2\) for the Euclidean norm. Given a matrix \(S\), we define a semi-norm of \(x\) as \(\|x\|_S := \sqrt{\langle Sx, Sx \rangle}\). When \(S\) is the identity matrix \(I\), we recover the standard Euclidean norm. When \(S^{-1}\) is positive definite, the semi-norm becomes the weighted-norm. Its dual norm exists and is defined by \(\|u\|_{S^*} = \max \{\langle u, v \rangle : \|v\|_S = 1\}\). When \(S^{-1}\) is not positive definite, we still consider the quantity \(\|u\|_{S^*} = \max \{\langle u, v \rangle : \|v\|_S = 1\}\), although \(\|u\|_{S^*}\) is finite if only if \(u \in \text{Ran}(S^{-1})\).

We also use \(\|\cdot \|_X\) (resp. \(\|\cdot \|_Y\)) and \(\|\cdot \|_{X^*}\) (resp. \(\|\cdot \|_{Y^*}\)) for the (semi) norm and the corresponding dual norm in the primal space \(X\) (resp. the dual space \(Y\)). Given a proper, closed, and convex function \(f\), we use \(\text{dom}(f)\) and \(\partial f(x)\) to denote its domain and its subdifferential at \(x\). If \(f\) is differentiable, then we use \(\nabla f(x)\) for its gradient at \(x\). For a given set \(C\), \(\delta_C(x) := 0\) if \(x \in C\) and \(\delta_C(x) := +\infty\), otherwise, denotes the indicator function of \(C\).

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

2.2. Smoothed proximity functions and Bregman distance. We use the following two mathematical tools in the sequel.

2.2.1. Proximity functions. Given a nonempty, closed and convex set \(Z\), a continuous, and \(\mu_p\)-strongly convex function \(p\) w.r.t. the semi-norm \(\|\cdot\|_Z\) is called a proximity function or prox-function of \(Z\) if \(Z \subseteq \text{dom}(p)\). We also denote

\[
\bar{z}^c := \arg\min \{p(z) : z \in \text{dom}(p)\} \quad \text{and} \quad D_Z := \sup \{p(z) : z \in Z\},
\]

as the prox-center of \(p\) and the prox-diameter of \(Z\), respectively. Without loss of generality, we can assume that \(\mu_p = 1\) and \(p(\bar{z}^c) = 0\). Otherwise, we can shift and rescale \(p\). Moreover, \(D_Z \geq 0\), and it is finite if \(Z\) 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 this class of prox-functions by \(S_{L,1}^{1,1}\). For example, \(p_Z(z) := (1/2)\|z\|_Z^2\) is a simple prox-function in \(\mathbb{R}^n\), i.e.,

\[
\frac{1}{2}\|z\|_Z^2 \in S_{L,1}^{1,1}.
\]

2.2.2. Bregman distance. Instead of smoothing the primal-dual problems \((1.1) - (1.2)\) by smooth proximity functions, we use a Bregman distance defined via \(p_Z\) as

\[
b_Z(z, \hat{z}) := p_Z(z) - p_Z(\hat{z}) - \langle \nabla p_Z(\hat{z}), z - \hat{z} \rangle, \quad \forall z, \hat{z} \in \mathbb{R}^n,
\]

where \(p_Z \in S_{L,1}^{1,1}\). Clearly, if we fix \(\hat{z} = \bar{z}^c\) at the center point of \(p_Z\), then \(b_Z(z, \bar{z}^c) = p_Z(z)\). In addition, \(\nabla_1 b_Z(z, z) = 0\) for all \(z \in Z\). We use in the sequel \(\nabla b_Z\) for \(\nabla_1 b_Z\).
2.3. Basic assumption. Our main assumption for problems (1.1)-(1.2) is to guarantee the strong duality, which essentially requires the following assumption (see, Proposition 15.22).

**Assumption A.1.** The solution set \( X^* \) of the primal problem (1.1) (or (1.3)) is nonempty. In addition, the following assumption holds for either (1.1) or (1.3):

(a) The condition \( 0 \in \text{ri} (\text{dom} (g) - A(\text{dom} (f))) \) for (1.1) holds.
(b) The Slater condition \( \text{ri} (\text{dom} (f)) \cap \{Ax = c\} \neq \emptyset \) for (1.3) holds.

Under Assumption A.1, the strong duality for (1.1)-(1.2) holds, see, e.g., [1]. The solution set \( \mathcal{Y}^* \) of the primal-dual (or the saddle point) set of (1.1)-(1.2) is nonempty, and

\[
P^* = f(x^*) + g(Ax^*) = D^* = -f^*(-A^T y^*) - g^*(y^*), \ \forall x^* \in X^*, \ \forall y^* \in \mathcal{Y}^*.  \tag{2.3}
\]

Let \( \mathcal{W}^* := X^* \times \mathcal{Y}^* \) be the primal-dual (or the saddle point) set of (1.1)-(1.2). Then, (2.3) is equivalent to \( f(x^*) + g(Ax^*) + f(-A^T y^*) + g^*(y^*) = 0 \) for all \((x^*, y^*) \in X^* \times \mathcal{Y}^*\). In addition, we can write the optimality condition of (1.1)-(1.2) as follows:

\[
-A^T y^* \in \partial f(x^*) \quad \text{and} \quad y^* \in \partial g(Ax^*).  \tag{2.4}
\]

Note that this condition can be written as \( 0 \in \partial f(x^*) + A^T \partial g(Ax^*) \) for the primal problem (1.1), and \( 0 \in \partial g^*(y^*) - ADf^*(-A^T y^*) \) for the dual problem (1.2).

2.4. Smoothed primal-dual gap function. The gap function \( G \) defined in (1.4) 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 \( b_X \) be a Bregman distance defined on \( X \), and \( \hat{x} \in \mathbb{R}^n \) be given, we consider an approximation to \( D(\cdot) \) as

\[
D_\gamma (y; \hat{x}) := - \min_{x \in \mathbb{R}^n} \{ f(x) + \langle y, Ax \rangle + \gamma b_X(x, \hat{x}) \} - g^*(y) \equiv -f^*_\gamma (-A^T y) - g^*(y),  \tag{2.5}
\]

where \( \gamma > 0 \) is a dual smoothness parameter. The minimization subproblem in (2.5) always admits a solution, which is denoted by

\[
x^*_\gamma (y; \hat{x}) \in \arg \min_{x \in \mathbb{R}^n} \{ f(x) + \langle y, Ax \rangle + \gamma b_X(x, \hat{x}) \}.  \tag{2.6}
\]

We emphasize that our algorithms presented in the next sections support parallel and distributed computation for the decomposable setting of (1.1) or (1.3), where \( f \) is decomposed into \( N \) terms as \( f(x) := \sum_{i=1}^N f_i(x_i) \) with the \( i \)-th 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 \( b_X(x, \hat{x}) := \sum_{i=1}^N b_{X_i}(x_i, \hat{x}_i) \) to approximate the dual function \( D \) defined in (1.2). 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 the sequel, see, Section 5.

2.4.2. The second smoothed approximation. Let \( b_Y \) be a Bregman distance defined on \( \mathcal{Y} \) the feasible set of the dual problem (1.2) and \( \hat{y} \in \mathbb{R}^m \). We consider an approximation to the objective \( g(\cdot) \) in (1.1) as

\[
g_\beta (u; \hat{y}) := \max_{y \in \mathbb{R}^m} \{ \langle u, y \rangle - g^*(y) - \beta b_Y(y, \hat{y}) \},  \tag{2.7}
\]

where \( \beta > 0 \) is a primal smoothness parameter. We also denote the solution of the maximization problem in (2.7) by \( y^*_\beta (u; \hat{y}) \), i.e.: \( y^*_\beta (u; \hat{y}) := \arg \max_{y \in \mathbb{R}^m} \{ \langle u, y \rangle - g^*(y) - \beta b_Y(y, \hat{y}) \}. \)

We consider an approximation to the primal objective \( P \) as

\[
P_\beta (x; \hat{y}) := f(x) + g_\beta (Ax; \hat{y}).  \tag{2.9}
\]

This function is the second smoothed approximation for the primal problem. We note that if \( g(\cdot) := \delta (\cdot) \) and \( b_Y(\cdot) := (1/2) \| \cdot \|_2^2 \), then \( y^*_\beta (u; \hat{y}) = \hat{y} + \beta^{-1} (u - c) \), which has a closed form.
2.5. Smoothed gap function and its properties. Given $D_\gamma$ and $P_\beta$ defined by (2.5) and (2.9), 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

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

where $\gamma$ and $\beta$ are two smoothness parameters.

The following lemma provides fundamental bounds of the objective residual $P(x) - P^*$ for the unconstrained form (1.1), and the objective residual $f(x) - f^*$ and the feasibility gap $\|Ax - b\|$ for the constrained form (1.3). For clarity of exposition, we move its proof to Appendix A.2.

Lemma 2.1. Let $G_{\gamma, \beta}$ be the smoothed gap function defined by (2.10) and $S_\beta(x) := P_\beta(x; \bar{y}) - P^* = f(x) + g_\beta(Ax; \bar{y}) - P^*$ be the smoothed objective residual. Then, we have

$$S_\beta(x) \leq G_{\gamma, \beta}(w, \bar{w}) + \gamma b_x(x^*, \bar{x}) \quad \text{and} \quad \frac{1}{\beta} \|y_\beta^*(Ax; \bar{y}) - y^*\|_Y^2 \leq b_Y(y^*; \bar{y}) + \frac{1}{\beta} S_\beta(x).$$

(2.11)

Suppose that $g(\cdot) := \delta_{\{\ell_1\}}(\cdot)$. Then, for any $y^* \in Y^*$ and $x \in X$, one has

$$- \|y^*\|_Y \|Ax - c\|_{Y^*} \leq f(x) - f^* \leq f(x) - g(y).$$

(2.12)

The following primal objective residual and feasibility gap estimates hold for (1.3):

$$\left\{ \begin{array}{l}
 f(x) - f^* \leq S_\beta(x) - (y^*, Ax - c) + \beta b_Y(y^*; \bar{y}), \\
 \|Ax - c\|_{Y^*} \leq \beta L b_Y \left[ \|y^* - \bar{y}\|_Y + \left( \|y^* - \bar{y}\|_Y^2 + 2L b_Y^{-1} S_\beta(x) \right)^{1/2} \right],
 \end{array} \right.$$

(2.13)

where the quantity under the square root is always nonnegative.

The estimates (2.12) and (2.13) 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 $\gamma_k$ and $\beta_k$. Hence, the convergence rate of the objective residual $f(\bar{x}^k) - f^*$ and feasibility gap $\|Ax^k - c\|$ depends on the rate of $\{\gamma_k, \beta_k\}$.

The second inequality in (2.11) shows that the distance between $y_\beta^*(Ax; \bar{y})$ and $y^*$ is controlled by quantities that will remain bounded. In practice, we observed that $y_\beta^*(Ax; \bar{y})$ seems to be converging to $y^*$. Hence, restarting the algorithm with $\bar{y}' = y_\beta^*(Ax; \bar{y})$, as we will do in the experiments will not hurt too much convergence, even in unfavorable cases.

3. The accelerated primal-dual gap reduction algorithm. Our new scheme builds upon Nesterov’s acceleration idea [41, 42]. 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 $\tau_k$ and $\beta_k$ simultaneously at each iteration with analytical updating formulas.

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

$$\left\{ \begin{array}{l}
 \bar{x}^k = (1 - \tau_k)\bar{x}^k + \tau_k \bar{x}^{k+1}, \\
 y_{\beta_k}^* (A\bar{x}^k; \bar{y}) := \arg \max \{(A\bar{x}^k, y) - g^*(y) - \beta_{k+1} b_Y(y; \bar{y})\}, \\
 \bar{x}^{k+1} := \text{prox}_{\beta_{k+1} L_A^{-1} f} \left( \bar{x}^k - \beta_{k+1} L_A^{-1} A^\top y_{\beta_k}^* (A\bar{x}^k; \bar{y}) \right), \\
 \bar{x}^{k+1} := \bar{x}^k - \tau_{k}^{-1}(\bar{x}^k - \bar{x}^{k+1}).
 \end{array} \right.$$  

(ASGARD)

where $\tau_k \in (0, 1]$ and $\beta_k > 0$ are parameters that will be defined in the sequel. The constant $L_A$ is defined as $L_A := \|A\|^2 = \max_{x \in X} \left\{ \frac{\|Ax\|_{Y^*}^2}{\|x\|_X^2} \right\}$. 

6
The ASGARD scheme requires a mirror step in $q^*$ to get $y^*_{βk+1}(·; ˆ{y})$ in (2.8) and a proximal step of $f$. Computing this proximal step can be implemented in parallel when the decomposition structure in (1.1) is available as discussed above.

The following lemma shows that $x^{k+1}$ updated by ASGARD decreases the smoothed objective residual $P_{βk+1}(x^{k+1}) − P^*$, whose proof can be found in Appendix A.3.1.

**Lemma 3.1.** Let us choose $τ_0 := 1$. If $τ_k ∈ (0, 1)$ is the unique positive root of the cubic polynomial equation $τ^3/L_{by} + τ^2 + τ^2_k τ − τ^2 = 0$ for $k ≥ 1$, and $β_k := \frac{β_k−1}{1 + τ_k / L_{by}}$, then $β_k → 0$ as $k → ∞$, and

$$P_{βk+1}(x^{k+1}) − P^* + \frac{τ^2}{β_{k+1}} \frac{L_A}{2} \|x^{k+1} − x^∗\|^2 ≤ \frac{τ^2}{β_{k+1}} \frac{L_A}{2} \|x^0 − x^∗\|^2. \tag{3.1}$$

Moreover, $\frac{1}{k+1} ≤ τ_k ≤ \frac{2}{k+2}$ and if $L_{by} = 1$, then $β_k ≤ \frac{2β_k−1}{k+1}$.

### 3.2. The primal-dual algorithmic template.

Similarly to the accelerated scheme [2, 41], we can eliminate $x^k$ in ASGARD by combining its first and last line to obtain

$$x^{k+1} = x^{k+1} + \frac{(1 − τ_k)τ_k+1}{τ_k} (x^{k+1} − x^k).$$

Now, we combine all the ingredients presented previously and this step to obtain a primal-dual algorithmic template for solving (1.1) as in Algorithm 1.

**Algorithm 1 (Accelerated Smoothed GAP Reduction (ASGARD) algorithm)**

**Initialization:**
1. Choose $β_1 > 0$ (e.g., $β_1 := 0.5√L_A$) and set $τ_0 := 1$.
2. Choose $x^0 ∈ X$ arbitrarily, and set $x^0 := x^0$.

**For** $k = 0$ **to** $k_{max}$, **perform:**
3. Compute $τ_{k+1} ∈ (0, 1)$ the unique positive root of $τ^3/L_{by} + τ^2 + τ^2_k τ − τ^2 = 0$.
4. Compute the dual step by solving

$$y^*_{βk+1}(Ax^k; ˆ{y}) := \arg \max_{y ∈ Y} \{ ⟨Ax^k, ˆ{y}⟩ − g^*(y) − β_{k+1} b_Y(ˆ{y}, ˆ{y}) \}.$$

5. Compute the primal step $x^{k+1}$ using the prox $f$ of $f$ as

$$x^{k+1} := \text{prox}_{β_{k+1}L_A−1} f \left( x^k − β_{k+1}L_A−1 A^T y^*_{βk+1}(Ax^k; ˆ{y}) \right).$$

6. Update $x^{k+1} = x^{k+1} + \frac{τ_{k+1}(1−τ_k)}{τ_k} (x^{k+1} − x^k)$ and $β_{k+2} := \frac{β_{k+1}}{1 + L_{by} τ_{k+1}}$.

**End for**

The computationally heavy steps of Algorithm 1 are given by Steps 3 and 5. At Step 4, $y^*_{βk+1}(Ax^k; ˆ{y})$ needs a matrix-vector multiplication $Ax$ and one prox $g^*$. If $g(·) := δ(c)(·)$ and $b_Y(·) := (1/2)∥ · ∥^2$, then we can compute $y^*_{βk+1}(Ax^k; ˆ{y}) = ˆ{y} + β_{k+1}−1(Ax^k − c)$. At Step 5, the algorithm requires one proximal step on $f$, which can be implemented in parallel when $f$ has a decomposable structure. For this step, we need one adjoint matrix-vector multiplication $A^T ˆ{y}$.

### 3.3. Convergence analysis.

Our first main result is the following theorem, which shows the $O(1/k)$ convergence rate of Algorithm 1 for both the unconstrained problem (1.1) and the constrained setting (1.3).

**Theorem 3.2.** Let $\{x^k\}$ be the primal sequence generated by Algorithm 1 for any $β_1 > 0$, and $by$ be chosen such that $L_{by} = 1$.
If \( \text{dom}(g) = \mathcal{Y} \), then the primal objective residual of (1.1) satisfies
\[
P(\bar{x}^k) - P^* \leq \frac{L_A}{2\beta_1 k} \|\bar{x}^0 - x^*\|_X^2 + \frac{2\beta_1 k}{k + 1} b_y(y^*(\bar{x}^k), \hat{y}) \quad \text{for all } k \geq 1.
\] (3.2)

where \( y^*(\bar{x}^k) \in \partial g(A\bar{x}^k) \).

If \( g(\cdot) := \delta_{\{c\}}(\cdot) \), then the following bounds hold for problem (1.3):
\[
\begin{align*}
  f(\bar{x}^k) - f^* & \geq -\|y^*\|_Y \|A\bar{x}^k - c\|_{\mathcal{Y}^*}, \\
  f(\bar{x}^k) - f^* & \leq \frac{L_A}{2\beta_1 k} \|\bar{x}^0 - x^*\|_X^2 + \|y^*\|_Y \|A\bar{x}^k - c\|_{\mathcal{Y}^*} + \frac{2\beta_1 k}{k + 1} b_y(y^*, \hat{y}), \\
  \|A\bar{x}^k - c\|_{\mathcal{Y}^*} & \leq \frac{\beta_1}{k + 1} \left( \|y^* - \hat{y}\|_Y + \left( \|y^* - \hat{y}\|_Y^2 + \frac{\beta_1}{\beta_2} \|\bar{x}^0 - x^*\|_X^2 \right)^{1/2} \right).
\end{align*}
\] (3.3)

Proof. If \( \text{dom}(g) = \mathcal{Y} \), then \( \partial g(Ax) \neq \emptyset \). Let \( y^*(x) \in \partial g(Ax) \). We can show that
\[
g(Ax) = \langle Ax, y^*(x) \rangle - g^*(y^*(x)) \leq \max_{y \in \mathcal{Y}} \{ \langle Ax, y \rangle - g^*(y) - \beta b_y(y, \hat{y}) \} + \beta b_y(y^*(x), \hat{y}) = g_{\beta}(Ax) + \beta b_y(y^*(x), \hat{y}).
\]
The bound (3.2) follows directly from (3.1) and this inequality.

If \( g(\cdot) = \delta_{\{c\}}(\cdot) \), then we apply Lemma 2.1 and use the bound on the smoothed optimality gap given in Lemma 3.1 with noting that \( \tilde{x}^0 = \bar{x}^0 = \bar{x}_0 \). Since \( \beta_k \leq \frac{2\beta_1}{k + 1} \) and
\[
\frac{\tau_k^2}{\beta_{k+1}^2} = \frac{\beta_k^2}{\beta_{k+1}^2} \frac{1}{\beta_k^2} \leq \left( \frac{1}{\beta_1(k + 1)} \right)^2 (k + 1)^2 = \frac{1}{\beta_1^2},
\]
we obtain the bounds as in (3.3). \( \square \)

From (3.2), if \( \text{dom}(\partial g) \) is bounded, then \( b_y(y^*(\bar{x}^k), \hat{y}) \leq \sup_{y \in \text{dom}(\partial g)} b_y(y, \hat{y}) < +\infty \). As a special case, if \( g \) is Lipschitz continuous, then \( \partial g \) is bounded \( \mathbb{1} \). For (3.3), note that if we choose \( \hat{y} := 0^m \), then the bounds (3.3) can be further simplified as
\[
\begin{align*}
  f(\bar{x}^k) - f^* & \leq \frac{1}{k} \left( \frac{L_A}{2\beta_1} \|\bar{x}^0 - x^*\|_X^2 + 3\beta_1 \|y^*\|_Y^2 + \sqrt{\frac{L_A}{\beta_1}} \|\bar{x}^0 - x^*\|_X \|y^*\|_Y \right), \\
  \|A\bar{x}^k - c\|_{\mathcal{Y}^*} & \leq \frac{\beta_1}{k + 1} \left( 2\|y^*\|_Y + \sqrt{\frac{L_A}{\beta_1}} \|\bar{x}^0 - x^*\|_X \right).
\end{align*}
\] (3.4)

Clearly, the choice of \( \beta_1 \) in Theorem 3.2 trades off between \( \|\bar{x}^0 - x^*\|_X^2 \) and \( \|y^* - \hat{y}\|_Y^2 \) on the primal objective residual \( f(\bar{x}^k) - f^* \) and on the feasibility gap \( \|A\bar{x}^k - c\|_{\mathcal{Y}^*} \).

4. The accelerated dual smoothed gap reduction method. Algorithm [1] can be viewed as an accelerated proximal scheme applying to minimize the function \( P_y(\cdot; \hat{y}) \) defined in (2.9).

Now, we exploit the smoothed gap function \( G_{\gamma, \beta} \) defined by (2.10) to develop a novel primal scheme for solving (1.1). Our goal is to design a new scheme to compute a primal-dual sequence \( \{\hat{x}^k, \hat{y}^k\} \) and a parameter sequence \( \{(\gamma_k, \beta_k)\} \) such that max \( \{0, G_{\gamma_k, \beta_k}(\hat{x}^k, \hat{y}^k)\} \) converges to zero.

4.1. The method. Given \( \hat{x}^k := (\bar{x}^k, \hat{y}^k) \in \mathcal{W} \), we derive a scheme to compute a new point \( \hat{x}^{k+1} := (\bar{x}^{k+1}, \hat{y}^{k+1}) \) as follows:
\[
\begin{align*}
  \hat{y}^k & := (1 - \tau_k)\hat{y}^k + \tau_k y_{\beta_k}(A\bar{x}^k; \hat{y}), \\
  \hat{y}^{k+1} & := \text{prox}_{\gamma_k L_A^y} g^* \left( \hat{g}^k + \gamma_k L_A^y A\bar{x}^k \right), \\
  \bar{x}^{k+1} & := (1 - \tau_k)\bar{x}^k + \tau_k x_{\gamma_k \beta_k}^* (\hat{y}^k; \hat{x}), \quad \text{(ADSG)}
\end{align*}
\]

where \( \tau_k \in (0, 1) \) and the parameters \( \beta_k > 0 \) and \( \gamma_k > 0 \) will be updated in the sequel. The points \( x_{\gamma_{k+1}}^*(\hat{y}^k; \hat{x}) \) and \( y_{\beta_k}(A\bar{x}^k; \hat{y}) \) are computed by (2.6) and (2.8), respectively. This scheme requires one primal step for \( x_{\gamma_{k+1}}^*(\hat{y}^k; \hat{x}) \), one dual step for \( y_{\beta_k}^*(A\bar{x}^k; \hat{y}) \), and one dual proximal-gradient step for \( \hat{x}^{k+1} \). Since the accelerated step is applied to \( g_{\gamma} \), we call this scheme the Accelerated Dual Smoothened Gap Reduction (ADSG) scheme.
The following lemma, whose proof is in Appendix A.4, shows that \( \bar{w}^{k+1} \) updated by ADSGARD decreases the smoothed gap \( G_{\tau_k b_k} (\bar{w}^k) \) with at least a factor of \((1 - \tau_k)\).

**Lemma 4.1.** Let \( \bar{w}^{k+1} := (\bar{x}^{k+1}, \bar{y}^{k+1}) \) be updated by the ADSGARD scheme. Then, if \( \tau_k \in (0, 1] \), \( \beta_k \) and \( \gamma_k \) are chosen such that \( \beta_k \gamma_k \geq \hat{L}_A \) and

\[
(1 + \tau_k/L_bx) \gamma_{k+1} \geq \gamma_k, \quad \beta_{k+1} \geq (1 - \tau_k) \beta_k, \quad \text{and} \quad \frac{\hat{L}_A}{\gamma_{k+1}} \leq \frac{(1 - \tau_k) \beta_k}{\tau_k^2},
\]

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

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 \( p_3(\tau) := \tau^3/L_{bx} + \beta^2 + \tau^2_{k-1} \tau - \tau_0^2 = 0 \), then \( \frac{1}{\tau_0} \leq \tau_k \leq \frac{2}{\tau_1} \) for \( k \geq 1 \). The parameters \( \beta_k \) and \( \gamma_k \) computed by \( \beta_1 \gamma_1 = \hat{L}_A \) and

\[
\gamma_{k+1} := \frac{\gamma_k}{1 + \tau_k/L_{bx}} \quad \text{and} \quad \beta_{k+1} := (1 - \tau_k) \beta_k,
\]

satisfy the conditions in (4.1).

In addition, if \( L_{bx} = 1 \), then \( \gamma_k \leq \frac{2 \gamma_1}{k+1} \) and \( \frac{\hat{L}_A}{\gamma_2 (k+1)} \leq \beta_{k+1} \leq \frac{\beta_1}{k+1} \) for \( k \geq 1 \).

### 4.2. The primal-dual algorithmic template.

We combine all the ingredients presented in the previous subsections to obtain a primal-dual algorithmic template for solving (1.1) as shown in Algorithm 2.

**Algorithm 2 (Accelerated Dual Smoothed GAP Reduction (ADSGARD))**

**Initialization:**
1. Choose \( \gamma_1 > 0 \) (e.g., \( \gamma_1 := \sqrt{\hat{L}_A} \)). Set \( \beta_1 := \hat{L}_A/\gamma_1 \) and \( \tau_0 := 1 \).
2. Take an initial point \( \bar{y}_0 := \bar{y} \in \mathcal{Y} \).

**For** \( k = 0 \) to \( k_{\text{max}} \), **perform:**
3. Update \( \bar{y}_k := (1 - \tau_k) \bar{y}_k + \tau_k \bar{y}_k^* \).
4. Compute \( \bar{x}_{k+1}^* \) in parallel with
   \[
   \bar{x}_{k+1}^* := \arg \min_{x \in \mathcal{X}} \{ f(x) + (A^\top \bar{y}_k^* x) + \gamma_{k+1} b_{AX}(x, \bar{x}) \}.
   \]
5. Update the dual vector
   \[
   \bar{y}^{k+1} := \text{prox}_{\gamma_{k+1} \hat{L}_A^{-1} g^*} \left( \bar{y}^k + \gamma_{k+1} \hat{L}_A^{-1} A \bar{x}_{k+1}^* \right).
   \]
6. Update the primal vector:
   \[
   \bar{x}^{k+1} := (1 - \tau_k) \bar{x}^k + \tau_k \bar{x}_{k+1}^*.
   \]
7. Compute
   \[
   \bar{y}_{k+1}^* := \arg \max_{y \in \mathcal{Y}} \{ (A \bar{x}^{k+1}) y - g^* (y) - \beta_{k+1} b_{by} (y, \bar{y}) \}.
   \]
8. Compute \( \gamma_{k+1} := (1 - \tau_k) \) the unique positive root of \( \frac{\tau^3}{L_{by}} + \beta^2 + \tau_{k-1} \tau - \tau_0^2 = 0. \)
9. Update \( \gamma_{k+2} := \frac{\gamma_{k+1}}{1 + \beta_k \gamma_{k+1}} \), and \( \beta_{k+2} := (1 - \tau_{k+1}) \beta_{k+1} \).

**End for.**

We note that since \( \tau_0 = 1 \), Step 3 shows that \( \bar{y}^0 = \bar{y}_0^* \), and while Step 6 leads to \( \bar{x}^1 = \bar{x}_1^* \). The main steps of Algorithm 2 are Steps 4, 5, and 7, where we need to solve the subproblem (2.6), and to update two dual steps, respectively. The first dual step requires the proximal operator \( \text{prox}_{\rho g^*} \) of \( g^* \), while the second dual step computes \( \bar{y}_{k+1}^* = \text{prox}_{\rho g^*} (A \bar{x}^{k+1}; \bar{y}) \).
When $g(\cdot) := \delta_{\{c\}}(\cdot)$ the indicator of $\{c\}$ in the constrained problem (1.3), we have

$$y_k^* (Ax; \hat{y}) = \nabla b \gamma (\beta_k (Ax - c), \hat{y})$$

and $y_{k+1}^* := y_k^* + \gamma_{k+1} (Ax_{\gamma_k+1} \hat{y}; \hat{x} - c).$

The first dual step only require one matrix-vector multiplication $Ax$. Clearly, by Step 5 it follows that $Ax_{k+1} = 1 - \gamma_k (Ax_{k+1} - c) + \gamma_{k+1} \nabla b (\hat{y}_{k+1}, \hat{y})$, which is equivalent to $Ax_{k+1} - c = \beta_k \nabla b \gamma (\hat{y}_{k+1}, \hat{y})$. Hence, $Ax_{k+1} - c = 1 - \gamma_k \beta_k \nabla b \gamma (\hat{y}_{k+1}, \hat{y}) + \gamma_{k+1} (\hat{y}_{k+1} - \hat{y})$ due to Step 5. Finally, we can derive an update rule for $y_k^*$ as

$$y_{k+1}^* = \nabla b_{\gamma} \left((1 - \gamma_k) \beta_k \nabla b \gamma (\hat{y}_{k+1}, \hat{y}) + \frac{\gamma_{k+1}}{\gamma_k+1} (\hat{y}_{k+1} - \hat{y})\right).$$

Consequently, each iteration of Algorithm 2 requires one solution of the primal step at Step 4, one matrix-vector multiplication $Ax$ and the adjoint $A^T y$.

4.3. Convergence analysis. The following theorem shows the convergence of Algorithm 2.

For the constrained setting (1.3), we still have the lower bound on $f(x^*) - f^*$ as in Theorem 3.2 i.e. $-\|y^*\|_2 \|A x_k - c\|_{\gamma_k} \leq f(x_k) - f^*$ for any $x_k \in X$ and $y^* \in Y^*$.

**Theorem 4.2.** Let $b_x$ be chosen such that $L_{b_x} = 1$, and $\{\hat{w}_k\}$ be the sequence generated by Algorithm 2 for solving (1.1), where $\gamma_1 > 0$ is given.

If $\gamma_1 = 1$, then the following convergence bound holds

$$P(x_k) - P^* \leq \frac{2 \gamma_1}{k+1} b_x (x^*; \hat{x}) + \frac{L_{b_x} y^*}{\gamma_1} \gamma_k \gamma_{k+1} (\|y^*\|_2 + \|y^* - \hat{y}\|_2) \frac{1}{2}.$$ 

where $y^* (x_k) \in \partial g (A x_k)$. If $g(\cdot) := \delta_{\{c\}}(\cdot)$, then the following bounds for (1.3) hold:

$$\begin{align*}
    f(x_k) - f^* &\geq -\|y^*\|_2 \|A x_k - c\|_{\gamma_k}, \\
    f(x_k) - f^* &\leq \frac{2 \gamma_1}{k+1} b_x (x^*; \hat{x}) + \frac{L_{b_x} y^*}{\gamma_1} \gamma_k \gamma_{k+1} (\|y^*\|_2 + \|y^* - \hat{y}\|_2) \frac{1}{2}.
\end{align*}$$

**Proof.** Since $S_{\beta_1}(x) \leq G_{\beta_1}(w; \hat{w}) + \gamma_1 b_x (x^*, \hat{x})$, using Lemma 4.1 we can show that $S_{\beta_1}(x_k) \leq G_{\beta_1}(\hat{w}; \hat{x}) + \gamma_1 b_x (x^*, \hat{x})$. Similar to the proof of Theorem 3.2 we obtain the bound (4.4). The bounds in (4.5) are the consequences of Lemma 2.1 using $\beta_k \leq \frac{\beta^*_k}{\gamma_k}$, $\gamma_k \leq \frac{\gamma_k}{\gamma_k + 1}$ and $\gamma_{k+1} \leq \frac{\gamma_{k+1}}{\gamma_{k+1} + 1} \leq 2 \gamma_k$.

Similar to Theorem 3.2 we can simplify the bound (4.5) to obtain a simple bound as in (3.4), where we omit the details here. The choice of $\gamma_1$ and $\beta_1$ in Theorem 4.2 also trades off the primal objective residual and the primal feasibility gap.

4.4. The choice of the smoother. For this algorithm, one needs to choose a norm $\|\cdot\|_X$ and a smoother $p_X$ which is strongly convex in the norm $\|\cdot\|_S$. One possibility is to choose $\|\cdot\|_S$ in order to have a simple formula for $\hat{x}_{k+1} = x_k^*(\hat{y}; \hat{x})$. A classical choice is a diagonal $S$ and $p_X(\cdot) = \frac{1}{2} \|\cdot\|_S$ for a given $\hat{x} \in \mathbb{R}^n$.

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

Another possibility is to choose $S = A$ and $p_X(\cdot) = \frac{1}{2} \|\cdot\|_S^2$. In that case, the computation of $\hat{x}_{k+1}$ may require an iterative sub-solver but we are allowed to take $\hat{x} = x^*$. Indeed, as $A x^* = c$, we have that for all $x$, $b_{X} (x, x^*) = \frac{1}{2} \|x - x^*\|_2^2 = \frac{1}{2} (A x - c)^T (A x - c)$. Hence, we can consider $x^*$ 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 (AL) 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 (1.3). The inexact variant of this algorithm can be found in our early technical report [53, Section 5.3].

The augmented Lagrangian smoother. We choose here $p_X = \|\cdot\|_X^2 = \|\cdot\|_A^2$, $p_Y = \|\cdot\|_Y^2 = \|\cdot\|_T^2$ and $x^* = x^*$ and $b(x, \hat{x}) := (1/2)\|A(x-x^*)\|_{2, c}^2 = (1/2)\|Ax-c\|_{2, c}^2$. This is indeed the augmented term for the Lagrange function of (1.3). Note that even though $\hat{x}$ is unknown, $b(x, \hat{x})$ can be computed easily using the equality $Ax^* = c$.

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

\[
\begin{align*}
\hat{y}^k & := (1 - \tau_k)\hat{y}^k + \tau_k y_{\gamma_0}^*(A\hat{x}^k; \hat{y}), \\
\hat{x}_{\gamma_0}^k & := \arg\min_{x \in X} \left\{ f(x) + \langle \hat{y}^k, A^* - b \rangle + \frac{\gamma_0}{2} \|Ax - c\|_{2, c}^2 \right\}, \\
\hat{y}^{k+1} & := \hat{y}^k + \gamma_0 (A\hat{x}_{\gamma_0}^k (\hat{y}^k) - c), \\
\hat{x}^{k+1} & := (1 - \tau_k)\hat{x}^k + \tau_k \hat{x}_{\gamma_0}^k (\hat{y}^k),
\end{align*}
\]

(ASALGARD)

where $\tau_k \in (0, 1)$, $\gamma_0 > 0$ is the penalty (or the primal smoothness) parameter, and $\beta_k$ is the dual smoothness parameter. As a result, this method is called 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 (ASALGARD) so that it requires only one matrix-vector multiplication $Ax$. Consequently, the complexity-per-iteration of ASALGARD remains essentially the same as the standard augmented Lagrangian method [17].

The update rule for parameters. In our augmented Lagrangian method, we only need to update $\tau_k$ and $\beta_k$ 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 $t_k := t_{k-1}^k$, we can derive

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

(5.1)

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 (5.2). In general, solving this subproblem remains challenging due to the non-separability of the quadratic term $\|Ax - b\|_T^2$. 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 [40].

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

PROPOSITION 5.1. Let $\{\hat{w}^k\}$ be the sequence generated by Algorithm 3. Then, we have

\[
\begin{align*}
\frac{-8L_{b, y}\|y^\ast - \hat{y}\|_Y \gamma_0 (k+2)^2}{\gamma_0 (k+2)^2} \leq f(\hat{x}^k) - f^* \leq \frac{8L_{b, y}\|y^\ast - \hat{y}\|_Y \gamma_0 (k+2)^2}{\gamma_0 (k+2)^2}, \\
\|A\hat{x}^k - b\|_Y \gamma_0 (k+2)^2 \leq \frac{8L_{b, y}\|y^\ast - \hat{y}\|_Y \gamma_0 (k+2)^2}{\gamma_0 (k+2)^2}.
\end{align*}
\]

(5.3)

As a consequence, the worst-case iteration-complexity of Algorithm 3 to achieve an $\epsilon$-primal solution $\hat{x}^k$ for (1.3) is $\mathcal{O} \left( \sqrt{\frac{8L_{b, y}\|y^\ast - \hat{y}\|_Y \gamma_0 (k+2)^2}{\gamma_0 (k+2)^2}} \right)$.
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 $(\bar{x}^0, \bar{y}^0) \in X \times Y$.

For $k = 0$ to $k_{\text{max}}$, perform:
3. Update $y_{\beta_k}^*(\bar{x}^k; \bar{y}) := \nabla b_y(\beta_k^{-1}(Ax^k - c), \bar{y})$.
4. Update
   \[ \bar{x}^k_0(\bar{y}^k) := \arg\min_{x \in X} \{ f(x) + (\bar{y}^k, Ax - c) + \frac{\gamma_0}{2} \| Ax - c \|^2_{3, *} \}. \]  
   (5.2)
5. Update $\bar{y}^{k+1} := \bar{y}^k + \gamma_0(\bar{A}\bar{x}^k_0(\bar{y}^k) - c)$ and $\bar{x}^{k+1} := (1 - t_k^{-1})\bar{x}^k + t_k^{-1}\bar{x}^k_0(\bar{y}^k)$.
6. Update $t_{k+1} := 0.5(1 + \sqrt{1 + 4t_k^2})$ and $\beta_{k+2} := (t_{k+1} - 1)t_{k+1}\beta_{k+1}$.

End for

The estimate \[5.3\] guides us to choose a large value for $\gamma_0$ such that we obtain better convergence bounds. However, if $\gamma_0$ is too large, then the complexity of solving the subproblem \[5.2\] increases commensurately. In practice, $\gamma_0$ is often updated using a heuristic strategy \[5.2\]. In general settings, since the solution $\bar{x}^k_{k+1}$ computed by \[5.2\] requires to solve 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 \[1.1\] is strongly convex with the convexity parameter $\mu_f > 0$, then it is well-known \[44\] 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
   \[ x^*(u) := \arg\max_{x \in X} \{ (u, x) - f(x) \}. \]  
   (5.4)

In addition, if $f_A^*(\cdot) := f^*(-A^T(\cdot))$, then $\nabla f_A$ is Lipschitz continuous with $L_{f_A^*} := \frac{L_A}{\mu_f}$.

The primal-dual update scheme. In this subsection, we only illustrate the modification of \textbf{ADSGARD} to solve the strongly convex primal problem \[1.1\] as
\[
\begin{align*}
\bar{y}^k & := (1 - \tau_k)\bar{y}^k + \tau_k y_{\beta_k}^*(A\bar{x}^k; \bar{y}) \\
\bar{x}^{k+1} & := (1 - \tau_k)\bar{x}^k + \tau_k x^*(-A^T\bar{y}^k) \\
\bar{y}^{k+1} & := \text{prox}_{\mu_f x^*}(\bar{y}^k + \mu_f A x^*(-A^T\bar{y}^k)).
\end{align*}
\]  
(ADSGARD$_\mu$)

We note that we no longer have the dual smoothness parameter $\gamma_k$. Hence, the conditions \[1.1\] of Lemma 4.1 reduce to $\beta_{k+1} \geq (1 - \tau_k)\beta_k$ and $(1 - \tau_k)\beta_k \geq L_{f_A^*}^\top t_k^2$. From these conditions we can derive the update rule for $\tau_k$ and $\beta_k$ as in Algorithm \[3\] which is
\[ t_{k+1} := \frac{1}{2}(1 + \sqrt{1 + 4t_k^2}), \quad \beta_{k+1} := \frac{(t_k - 1)}{t_k} \beta_k \quad \text{and} \quad \tau_k := t_k^{-1}. \]  
(5.5)

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 \textbf{ADSGARD}_$_\mu$, whose proof is in Appendix \[A.6\].

\textsc{Proposition 5.2.} Suppose that the objective $f$ of the constrained convex problem \[1.3\] is strongly convex with the convexity parameter $\mu_f > 0$. Let $\{\bar{w}^k\}$ be generated by \textbf{ADSGARD}_$_\mu$ using the update rule \[5.5\]. Then
\[
\begin{align*}
-\frac{8L_{f_A^*}L_{f_A}L_{A^\top} \| y^* \| \| y^* - \bar{y} \| y^*}{\mu_f (k+2)^2} & \leq f(\bar{x}^k) - f^* \leq \frac{8L_{f_A^*}L_{A} \| y^* \| \| y^* - \bar{y} \| \| y^* - \bar{y} \|_2 + 4b_y(y^*; \bar{y})}{\mu_f (k+2)^2}, \\
\| Ax^k - b \| & \leq \frac{8L_{f_A^*}L_{A} \| y^* - \bar{y} \| y^*}{\mu_f (k+2)^2}. 
\end{align*}
\]  
(5.6)
This result shows that \( \text{ADSGARD}_m \) has the \( O(1/k^2) \) convergence rate w.r.t. to the objective residual and the feasibility gap. We note that in both Propositions 5.1 and 5.2, the bounds only depend on the quantities in the dual space and \( L_A \).

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:

\[
f^* := \min_{x \in X} \{ f(x) : Ax - b \in K \},
\]

where \( f, A \) and \( b \) are defined as in (1.3), and \( K \) is a nonempty, closed and convex set in \( \mathbb{R}^m \).

If \( K \) is bounded, then a simple way to process (5.7) is using a slack variable \( r \in K \) such that \( r := Ax - b \) and \( z := (x, r) \) as a new variable. Then we can transform (5.7) into (1.3) with respect to the new variable \( z \). The primal subproblem corresponding to \( r \) is defined as \( \min \{ \langle -y, r \rangle : r \in K \} \), which is equivalent to the support function \( s_K(y) := \sup \{ \langle y, r \rangle : r \in K \} \) of \( K \). Consequently, the dual function becomes \( \hat{g}(y) := g(y) - s_K(y) \), where \( g(y) := \min \{ f(x) + \langle Ax - b, y \rangle : x \in X \} \).

Now, we can apply the algorithms presented in the previous sections to obtain an approximate solution \( \hat{z}^k := (\hat{x}^k, \hat{r}^k) \) with a convergence guarantee on \( f(\hat{x}^k) - f^*, \| A\hat{x}^k - \hat{r}^k - b \|_m \), \( \hat{x}^k \in X \) and \( \hat{r}^k \in K \) as in Theorem 3.2 or Theorem 4.2.

If \( K \) is a cone (e.g., \( K := \mathbb{R}^+_m \), \( K \) is a second order cone \( \mathcal{L}^+_m \), or \( K \) is a semidefinite cone \( \mathcal{S}^+_m \)), then with the choice \( p_y(\cdot) := (1/2)\| \cdot \|^2 \), we can substitute the smoothed function \( f_\beta \) in (2.7) by the following one

\[
\hat{g}_\beta(Ax, \hat{y}) := \max \{ \langle Ax - c, y \rangle - (\beta/2)\| y - \hat{y} \|^2 : y \in -K^* \},
\]

where \( K^* \) is the dual cone of \( K \), which is defined as \( K^* := \{ z : \langle z, x \rangle \geq 0, x \in K \} \). With this definition, we use the smoothed gap function \( G_{\gamma, \beta} \) as \( G_{\gamma, \beta}(v; \hat{u}) := P_{\beta}(x; \hat{y}) - D_\gamma(x; \hat{x}) \), where \( D_\gamma(x; \hat{x}) := \min \{ f(x) + \langle Ax - c, y \rangle + \gamma b_X(x, \hat{x}) : x \in X \} \) is the smooth dual function defined as before, and \( P_{\beta}(x; \hat{y}) := f(x) + \hat{g}_\beta(Ax, \hat{y}) \).

In principle, we can apply one of the two previous schemes to solve (5.7). Let us demonstrate the \( \text{ADSGARD} \) for this case. Since \( K \) is a cone, we remain using the original scheme (ADSGARD) with the following changes:

\[
\left\{ \begin{array}{ll}
y_{\beta_k}(A\hat{x}^k; \hat{y}) & := \text{proj}_{-K^*} \left( \hat{y} + \beta^{-1}_k(A\hat{x}^k - c) \right), \\
\hat{y}^{k+1} & := \text{proj}_{-K^*} \left( \hat{y}^k + \beta^k \frac{\hat{d}^k}{\tau^k} \left( A\hat{x}^k_{\gamma_{k+1}}(\hat{y}^k) - c \right) \right),
\end{array} \right.
\]

where \( \text{proj}_{-K^*} \) is the projection onto the cone \(-K^*\). In this case, we still have the convergence guarantee as in Theorem 4.2 for the objective residual \( f(\hat{x}^k) - f^* \) and the primal feasibility gap \( \text{dist}(A\hat{x}^k - c, K) \). We note that if \( K \) is a self-dual conic cone, then \( K^* = K \). Hence, \( y^*_{\beta_k}(A\hat{x}^k; \hat{y}) \) and \( \hat{y}^{k+1} \) can be either efficiently computed or a closed form.

5.4. Restarting. Similar to other accelerated gradient algorithms in [25, 40, 51], restarting ASGARD and ADSGARD may lead to a faster algorithm in practice. We discuss in this subsection how to restart these two algorithms using a fixed iteration restarting strategy [40].

If we consider \( \text{ASGARD} \) then, when a restart takes place, we perform the following:

\[
\left\{ \begin{array}{ll}
\hat{x}^{k+1} & \leftarrow \hat{x}^{k+1}, \\
\hat{y} & \leftarrow y_{\beta_{k+1}}(A\hat{x}^k; \hat{y}), \\
\beta_{k+1} & \leftarrow \beta_1, \\
\tau_{k+1} & \leftarrow 1.
\end{array} \right.
\]

Restarting the primal variable at \( \hat{x}^{k+1} \) is classical, see, e.g., [40]. For the dual center point \( \hat{y} \), we suggest to restart it at the last dual variable computed. Indeed, by (2.11), we know that the distance between \( y^*_{\beta_k}(A\hat{x}^k; \hat{y}) \) and the optimal solution \( y^* \) will remain bounded. Hence, in the
favorable cases, we will benefit from a smaller distance between the new center point and $y^*$, while in the unfavorable cases, restarting should not affect too much the convergence.

For ADSGARD, we suggest to restart it as follows:

$$
\begin{align*}
\hat{y}^{k+1} &\leftarrow \hat{y}^{k+1}, \\
\bar{y}^{k+1} &\leftarrow \bar{y}^{k+1}, \\
\hat{x} &\leftarrow x^{\ast}_{\hat{y}^{k+1}}(\bar{y}^{k+1}; \hat{x}), \\
\beta_{k+1} &\leftarrow \beta_1, \\
\gamma_{k+1} &\leftarrow \gamma_1, \\
\tau_{k+1} &\leftarrow 1.
\end{align*}
$$

(5.10)

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 state-of-the-arts. While other numerical experiments can be found in our technical reports [53], we instead focus some extreme cases where existing methods may encounter arbitrarily slow convergence rate due to lack of theory, while our methods exhibits the $O(1/k)$ rate as predicted by the theory.

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

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

(6.1)

The second inequality is repeated $d - 1$ times, which makes the problem degenerate. Yet, qualification conditions hold since this is a linear program. This fits into our framework with $f(x) := 2x_n + \delta_{\{x_n \geq 0\}}(x_n)$. $Ax := [\sum_{k=1}^{n-1} x_k; x_n - \sum_{k=1}^{n-1} x_k; \cdots; x_n - \sum_{k=1}^{n-1} x_k]$, $c := (1, 0, \ldots, 0)^\top \in \mathbb{R}^d$ 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 chose $n = 10$ and $d = 200$. We implement both ASGARD and ADSGARD and their restarting variants. In Figure 6.1, we compare our methods against the Chambolle-Pock method [13]. We can see that the Chambolle-Pock method struggles with the degeneracy while ASGARD still exhibits a $O(1/k)$ sublinear rate of convergence.

In Figure 6.2, we compare methods requiring the resolution of a nontrivial optimization problem at each iteration. In this case, the inversion of a rank deficient linear system, we thus compare (ASALGARD) with and without restart against ADMM [9]. For ADMM, we selected the step-size parameter by sweeping from small values to large values and choosing the one that gives the fastest performance. Here also, our algorithm resists to the degeneracy and restarting improves the performance.

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

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

(6.2)

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 $i = 1, \cdots, N$. 

14
It is trivial to show that the dual problem of (6.2) is the following generalization of convex feasibility problem in convex optimization:

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

(6.3)

Clearly, when $A_i = I$ the identity matrix, (6.3) becomes the well-known convex feasibility problem. When $A_i = I$ for some $i \in \{1, \ldots, N\}$ and $A_i = A$, otherwise, (6.3) becomes the multiple-set split feasibility problem as considered in the literature. Assume that (6.3) has solution and $N \geq 2$. Hence, (6.2) and (6.3) satisfy Assumption A.1.

Our aim is to apply Algorithm 1 and Algorithm 2 to solve the primal problem (6.2), and compare them with the most state-of-the-art ADMM algorithm with multiple blocks [21]. Clearly, with nonorthogonal $A_i$, the primal subproblem of computing $x_i$ in the parallel-ADMM scheme [21] does not have a closed form solution, we typically 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 (6.2). With the use of Euclidean distance for our smoother, Algorithm 1 and Algorithm 2 can solve the primal subproblem (2.6) in $x_i$ with a closed form solution, which just require the projection onto $\mathcal{X}_i$.

The first experiment is for $N = 2$. We choose $\mathcal{X}_1 := \{ y \in \mathbb{R}^n : \epsilon y_1 - \sum_{j=2}^n y_j \leq 1 \}$ and $\mathcal{X}_2 := \{ y \in \mathbb{R}^n : \sum_{j=2}^n y_j \leq -1 \}$ to be two half-planes, where $\epsilon > 0$ is fixed. The constant $\epsilon$ represents the angle between these half-planes. It is well-known [5] 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 \( \epsilon \), we observe the convergence speed of ADMM, while our algorithms seem not to depend on \( \epsilon \). Figure 6.3 shows the convergence rate on the absolute feasibility \( \| \sum_{i=1}^{N} x_i \|_2 \) of three algorithms for \( n = 10,000 \). Since the objective value is always zero, we omit its plot here.

![Comparison of Algorithm 1, Algorithm 2 and ADMM with different values of \( \epsilon \).](image1)

**FIG. 6.3.** Comparison of Algorithm 1, Algorithm 2 and ADMM with different values of \( \epsilon \) (left). Comparison of Algorithm 3, Algorithm 4 and their restarted variant (restarting after every 100 iterations) (right). The number of variables is \( Nn = 20,000 \).

The theoretical version of Algorithm 1 and Algorithm 2 exhibits a convergence rate slightly better than \( O(1/k) \) and is independent of \( \epsilon \), while ADMM can be arbitrarily slow as \( \epsilon \) 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 \( 10^{-5} \) accuracy and then make slow progress to achieve the \( 10^{-6} \) accuracy. We also notice that the averaging sequence of ADMM converges at the \( O(1/k) \) rate but it remains far away from our theoretical rate in Algorithm 1 and Algorithm 2.

If we combine these two algorithms with our restarting strategy, both algorithms converge after 102 iterations. We see that Algorithm 1 performs very similar to Algorithm 2. We can also observe that the performance of our algorithms depends on \( L_A \) and initial points, but it is relatively independent of the geometric structure of problems as opposed to the ADMM scheme for solving (6.2).

Now we extend to the case \( N = 3 \) and \( N = 4 \), where we add two more sets \( \mathcal{X}_3 \) and \( \mathcal{X}_4 \). We choose \( \mathcal{X}_3 := \left\{ y \in \mathbb{R}^n : 0.5 \epsilon y_1 - \sum_{j=2}^{n} y_j = 1 \right\} \) to be a hyperplane in \( \mathbb{R}^n \), and \( \mathcal{X}_4 := \left\{ y \in \mathbb{R}^n : -y_1 + \sum_{j=3}^{n} y_j \leq 1 \right\} \) to be a half-plane in \( \mathbb{R}^n \). We test our algorithms and the multiblock-ADMM method in [21] and the results are plotted in Figure 6.4 for the case \( n = 10,000 \). In both cases, ADMM still make a slow progress as \( \epsilon \) is decreasing and \( N \) is increasing. Algorithm 1 and Algorithm 2 seem to scale slightly to \( N \). We note that since \( A_i = I \) for \( i = 1, \ldots, N \). The complexity-per-iterations of three algorithms in our experiment is essentially the same, and is not necessary to report.

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

**Problem structure assumptions.** Our approach requires the convexity and the existence of primal and dual solutions of (1.1). We argue that such assumptions are mild for (1.1) and (1.2) and can be verified a priori. We emphasize that existing primal-dual methods including decomposition [50], splitting, and alternating direction methods 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 (1.1) [19, 18, 20, 13, 47, 48].
Convergence characterization. We characterize the $O(1/k)$-convergence rate of the objective residual $P(x^k) - P^*$ for (1.1) without any smoothness or strong convexity-type assumption on $f$ or $g$. According to [44], this convergence is optimal for the class of nonsmooth convex problems (1.1). Our convergence rate does not depend on the smoothness parameter $\gamma$ as in [3, 44], but still requires the boundedness of $\partial g(\cdot)$. Unfortunately, we do now have the convergence of the iterates.

For the constrained setting (1.3), we can characterize the $O(1/k)$-convergence rate on both the primal objective residual $f(x^k) - f^*$ and the feasibility gap $\|Az^k - c\|$ separately. We note that this rate has been often achieved via a Lipschitz gradient assumption on $f$ as seen in [19, 18, 20, 13, 47, 48]. Otherwise, without additional assumptions, we can obtain such a rate only in joint primal-dual variables $(x^k, y^k)$ as in [13, 28, 29]. For multiple objective terms, [21] considered different types of parallel-ADMM algorithms, but the convergence characterization at the end is $o(1/\sqrt{k})$ on the feasibility.

Specifically, we develop two different types of algorithms, one with a weighted averaging scheme and without any averaging at all in the primal space. In addition, our bounds on (1.3) only depend on the distance between the initial points and an optimal solution instead of both the primal and the dual prox-diameters as in [13, 28, 29], which are applicable to problems with unbounded domain.

Decomposition methods. Our basic methods broadly fall under the class of decomposition methods in contrast to alternating direction methods, such as ADMM. We note that the convergence rate guarantee in existing ADMM methods apply to the reformulated problems rather than the original one. Some authors directly developed ADMM for solving (1.1) [20]. However, unless other additional assumptions are imposed, ADMM is failed to converge with more than two blocks as shown in [13, 21]. Other authors have attempted to provide conditions for the convergence of multiblock ADMM, e.g., [25]. These conditions are relatively generic and do not provide a clear recipe on how to choose the parameters properly. One fundamental disadvantage of ADMM is that it requires to process the linear operators in each subproblem, which is often nontrivial in many applications, while our methods do not require any augmented term and can avoid this drawback.

Smoothing and smoothness parameter updates. We use differentiable smoothing functions in contrast to Nesterov’s smoothing technique in [44]. We propose explicit rules to update the smoothness parameters simultaneously. We emphasize that this is one of the key contributions of this paper. In Nesterov’s smoothing methods [44], the smoothness parameter is set a priori proportional to the ratio $\frac{\alpha}{\bar{D}_X}$ between the accuracy and the prox-diameter $D_X$ of $X$. It is clear that this parameter is often very small which leads to large Lipschitz constant in the smoothed dual
To the best of our knowledge, we propose here the first adaptive primal-dual algorithms for function. Hence, the algorithm with fixed smoothness parameter often has a poor performance. While we also provide a weighted averaging scheme, we alternatively derive a method without any averaging in the primal for solving (1.1). The non-averaging schemes are important since averaging may destroy key structures, such as the sparsity or low-rankness in sparse optimization. Our weighted averaging scheme has increasing weight at the later iterates compared to non-weighted averaging schemes [13, 28, 29, 50]. As indicated in [18, 20], weighted averaging schemes have better performance guarantee than non-weighted 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 based on augmented Lagrangian methods such as alternating direction methods (e.g., AMA, ADMM and their variants) [9, 31, 32, 50]. Bregman and other splitting methods [1, 17, 23, 26, 39, 37, 38], and using variational inequality framework [13, 29, 27]. 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 use a more general variational inequality/monotone inclusion framework to study (1.1) and (1.2). Hence, the results achieved are distinct from our results.

Appendix A. Appendix: 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 A.1.** Let $h$ be a proper, closed and convex function defined on $Z$, and $h^*$ is its Fenchel conjugate. Let $b_Z$ be a Bregman distance as defined in (2.2) with a weighted norm. We define a smoothed approximation of $h$ as

$$h_\beta(z; \tilde{z}) := \max_{\hat{z} \in Z} \{ \langle z, \hat{z} \rangle - h^*(\hat{z}) - \beta b_Z(\hat{z}, \tilde{z}) \},$$  \hspace{1cm} (A.1)

where $\tilde{z} \in Z$ is fixed and $\beta > 0$ is a smoothness parameter. We also denote by $z^*_\beta(z; \tilde{z})$ the solution of the maximization problem in (A.1). Then, the following facts hold:

(a) We have a relation between the partial derivatives of $(z, \beta) \mapsto h_\beta(z; \tilde{z})$ as

$$\frac{\partial h_\beta(z; \tilde{z})}{\partial \beta}(\beta) = -b_Z(z^*_\beta(z; \tilde{z}), \tilde{z}) = -b_Z(\nabla h_\beta(z; \tilde{z}), \tilde{z}).$$

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

$$h_{\beta_{k+1}}(\tilde{z}; \tilde{z}) \leq h_{\beta_k}(\tilde{z}; \tilde{z}) - (\beta_k - \beta_{k+1}) \frac{\partial h_\beta(\tilde{z}; \tilde{z})}{\partial \beta}(\beta_{k+1})$$

$$= h_{\beta_k}(\tilde{z}; \tilde{z}) + (\beta_k - \beta_{k+1}) b_Z(\nabla h_{\beta_{k+1}}(\tilde{z}; \tilde{z}), \tilde{z}).$$  \hspace{1cm} (A.2)

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

$$h_\beta(\tilde{z}; \tilde{z}) \leq h_\beta(\tilde{z}; \hat{z}) + \langle \nabla h_\beta(\tilde{z}; \tilde{z}), \tilde{z} - \hat{z} \rangle + \frac{1}{2\beta} \| \tilde{z} - \hat{z} \|_{Z, \ast}^2.$$  \hspace{1cm} (A.3)

$$h_\beta(\hat{z}; \tilde{z}) + \langle \nabla h_\beta(\hat{z}; \tilde{z}), \hat{z} - \tilde{z} \rangle \leq h_\beta(\hat{z}; \tilde{z}) - \frac{\beta}{2} \| \nabla h_\beta(\hat{z}; \tilde{z}) - \nabla h_\beta(\tilde{z}; \tilde{z}) \|_{Z, \ast}^2.$$  \hspace{1cm} (A.4)

(d) Both functions $h$ and $h_\beta$ evaluated at different points $z, \tilde{z} \in Z$ satisfy

$$h_\beta(\tilde{z}; \tilde{z}) + \langle \nabla h_\beta(\tilde{z}; \tilde{z}), z - \tilde{z} \rangle \leq h(z) - \beta b_Z(\nabla h_\beta(\tilde{z}; \tilde{z}), \tilde{z}).$$  \hspace{1cm} (A.5)
We prove from item (a) to item (f) as follows.

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

$$
\begin{align*}
\hat{h}_\beta(\hat{z}; \hat{z}) + \langle \nabla \hat{h}_\beta(\hat{z}; \hat{z}), \hat{z} - z \rangle &= \langle \hat{z}, \hat{z}_\beta^* \rangle - h^*(\hat{z}_\beta^*) - \beta b_Z(\hat{z}_\beta^*, \hat{z}) + \langle \hat{z}_\beta^*, \hat{z} - \hat{z} \rangle \\
&= \langle \hat{z}, \hat{z}_\beta^* \rangle - h^*(\hat{z}_\beta^*) - \beta b_Z(\hat{z}_\beta^*, \hat{z}) \\
&\leq \max_{u \in Z} \{ (z, u) - h^*(u) \} - \beta b_Z(z_\beta^*, \hat{z}) = h(z) - \beta b_Z(\nabla h_\beta(\hat{z}; \hat{z}), \hat{z}).
\end{align*}
$$

(e) The classical 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 $z_{\beta,1}^* = z_{\beta}^*(z; \hat{z}_1)$ and $z_{\beta,2}^* := z_{\beta}^*(z; \hat{z}_2)$. Using the definition of $h_\beta$ in (A.1) and its optimality condition, we can derive

$$
\begin{align*}
\hat{h}_\beta(z; \hat{z}_2) &= \max_{z \in Z} \{ (z, z_\beta^*) - h^*(z_\beta^*) - \beta b_Z(z_\beta^*, \hat{z}_2) \} = \langle z, z_{\beta,2}^* \rangle - h^*(z_{\beta,2}^*) - \beta b_Z(z_{\beta,2}^*, \hat{z}_2) \\
&= \langle z, z_{\beta,2}^* \rangle - h^*(z_{\beta,2}^*) - \beta b_Z(z_{\beta,2}^*, \hat{z}_1) + \beta b_Z(z_{\beta,2}^*, \hat{z}_1) - \beta b_Z(z_{\beta,2}^*, \hat{z}_2) \\
&\leq \langle z, z_{\beta,1}^* \rangle - h^*(z_{\beta,1}^*) - \beta b_Z(z_{\beta,1}^*, \hat{z}_1) - \beta b_Z(z_{\beta,2}^*, \hat{z}_2) \\
&= h_\beta(z; \hat{z}_1) - \frac{\beta}{2} \| z_{\beta,1}^* - z_{\beta,2}^* \|^2 + \beta \left( b_Z(z_{\beta,2}^*, \hat{z}_1) - b_Z(z_{\beta,2}^*, \hat{z}_2) \right),
\end{align*}
$$

which proves (A.7).

A.2. The proof of Lemma 2.1 Key bounds for approximate solutions. We consider the smooth objective residual $S_\beta(x) := (f(x) + g_\beta(Ax; \hat{y})) - (f(x^*) + g(Ax^*))$. By using the definition of $g_\beta$, we can derive that

$$
\begin{align*}
g_\beta(Ax; \hat{y}) &= \max_{\hat{y} \in Y} \{ \langle Ax, \hat{y} \rangle - g^*(\hat{y}) - \beta b_Y(\hat{y}, \hat{y}) \} \\
&\geq \langle Ax, y^* \rangle - g^*(y^*) - \beta b_Y(y^*, \hat{y}) \\
&= \langle Ax - Ax^*, y^* \rangle + \langle Ax^*, y^* \rangle - g^*(y^*) - \beta b_Y(y^*, \hat{y}) \\
&= \langle A(x - x^*), y^* \rangle + g(Ax^*) - \beta b_Y(y^*, \hat{y}),
\end{align*}
$$

(A.8)
where the last line is the equality case in the Fenchel-Young inequality using the fact that $A^\top x^* \in \partial g^*(y^*)$. Similarly, we have
\[
f^*_\gamma(-A^\top y; \hat{x}) = \max_{\hat{x} \in \mathcal{X}} \{ -A^\top y, \hat{x} - f(\hat{x}) - \gamma b_x(\hat{x}, \hat{x}) \} \\
\geq \langle A^\top (y^* - y), x^* \rangle + f^*(A^\top y^*) - \gamma b_x(x^*, \hat{x}). \tag{A.9}\]
Combining (A.8), (A.9), the definition (2.10) of $G_{\gamma\beta}(\cdot; \hat{w})$, and the strong duality condition (2.3), we can show that
\[
G_{\gamma\beta}(w) := P_\beta(x) - D_\gamma(y) \\
= f(x) + g_\beta(Ax; \hat{y}) + f^*_\gamma(-A^\top y; \hat{x}) + g^*(y) \tag{2.3} \\
\geq S_{\beta}(x) + f^*_\gamma(-A^\top y; \hat{x}) + g^*(y) - f^*(A^\top y^*) - g^*(y^*) \tag{A.9} \\
\geq S_{\beta}(x) + \langle A^\top (y^* - y), x^* \rangle + g^*(y) - g^*(y^*) - \gamma b_x(x^*, \hat{x}) \\
\geq S_{\beta}(x) - \gamma b_x(x^*, \hat{x}), \tag{A.10}\]
where the last inequality holds because $g^*$ is convex and $Ax^* \in \partial g^*(y^*)$ due to (2.4). This proves the first inequality of (2.11).
Since $b_\gamma(\cdot; \hat{w})$ is 1-strongly convex w.r.t. the weighted-norm, using the optimality condition of the maximization problem in (2.7) at $y := y^*$, and $u := Ax$, we obtain
\[
g_\beta(Ax; \hat{y}) \geq \langle Ax, y^* \rangle - g^*(y^*) - \beta b_\gamma(y^*, \hat{y}) + \frac{\beta}{2} y_\gamma^2(Ax; \hat{y}) - y^2 \tag{A.11}\]
By (2.4), we have $-A^\top y^* \in \partial f(x^*)$. Using this and the convexity of $f$, we have $f(x) \geq f^* + \langle A(x - x^*), y^* \rangle$. Summing up the last inequality and (A.11), then using the definition of $S_{\beta}(x)$, we obtain
\[
\frac{\beta}{2} y_\gamma^2(Ax; \hat{y}) - y^2 \leq \beta b_\gamma(y^*, \hat{y}) + S_{\beta}(x) + g(Ax^*) + g^*(y^*) - \langle Ax^*, y^* \rangle \leq \beta b_\gamma(y^*, \hat{y}) + S_{\beta}(x),
\]
which implies the second estimate in (2.11), where the last inequality is due to the Fenchel-Young equality $g(Ax^*) + g^*(y^*) = \langle Ax^*, y^* \rangle$.

Now, we consider the choice $g(\cdot) := \delta_{\{c\}}(\cdot)$ in the constrained setting (1.3). Under Assumption A1, any $w^* := (x^*, y^*) \in \mathcal{W}_*$ is a saddle point of the Lagrange function $L(x, y) := f(x) + \langle Ax - c, y \rangle$, i.e., $L(x^*, y^*) \leq L(x^*, y^*) \leq L(x, y^*)$ for all $x \in \mathcal{X}$ and $y \in \mathbb{R}^m$. The dual function $D$ in (1.2) becomes $D(y) := -f^*(-A^\top y) - c^\top y = \min_x \{ f(x) + \langle Ax - c, y \rangle \}$. It leads to $D(y) \leq D^* = f^* \leq f(x) + \langle y^*, Ax - c \rangle$, and hence
\[
f(x) - D(y) \geq f(x) - f^* \geq \langle c - Ax, y^* \rangle \geq -\|y^*\|_\gamma \|Ax - c\|_{\gamma, *}, \tag{A.12}\]
for all $(x, y) \in \mathcal{W}$, which proves (2.12).
Finally, we prove (2.13). Indeed, using the definition of $g$ and $g_\beta$, and $Ax^* = c$, we can write
\[
f(x) - f(x^*) = f(x) + g_\beta(Ax; \hat{y}) - f(x^*) - g(Ax^*) - g_\beta(Ax; \hat{y}) + g(Ax^*) \\
= S_{\beta}(x) - g_\beta(Ax; \hat{y}) + g(Ax^*) \tag{A.8} \leq S_{\beta}(x) - \langle Ax - x^*, y^* \rangle + \beta b_\gamma(y^*, \hat{y}) \tag{A.10} \\
\leq G_{\gamma\beta}(w; \hat{w}) + \langle c - Ax, y^* \rangle + \beta b_\gamma(y^*, \hat{y}) + \gamma b_x(x^*, \hat{x}).
\]
We then use the lower bound inequality (2.12) to get
\[
\langle y^*, c - Ax \rangle \leq f(x) - f^* \leq S_{\beta}(x) - g_\beta(Ax; \hat{y}) + g(Ax^*) = S_{\beta}(x) - g_\beta(Ax; \hat{y}), \tag{A.13}\]
where \( g(Ax^*) = 0 \) due to the feasibility of \( x^* \), i.e., \( Ax^* = c \). Now, it is obvious that
\[
g_\beta(Ax; \dot{y}) := \sup_{\tilde{y} \in \mathcal{Y}} \{ \langle Ax - c, \tilde{y} \rangle - \beta b_\gamma(\tilde{y}, \dot{y}) \} \geq \langle Ax - c, y^* \rangle - \beta b_\gamma(y^*, \dot{y}).
\]
Hence, combining this estimate and (A.13), we obtain the first inequality in (2.13).

As \( \nabla b_\gamma(\cdot, \dot{y}) \) is \( L_{by} \)-Lipschitz continuous, \( b_\gamma(\tilde{y}, \dot{y}) = 0 \) and \( \nabla b_\gamma(\tilde{y}, \dot{y}) = 0 \), we have
\[
g_\beta(Ax; \dot{y}) = \sup_{\tilde{y} \in \mathcal{Y}} \{ \langle Ax - c, \tilde{y} \rangle - \beta b_\gamma(\tilde{y}, \dot{y}) \} \geq \sup_{\tilde{y} \in \mathcal{Y}} \left\{ \langle Ax - c, \tilde{y} \rangle - \frac{\beta L_{by}}{2} \| \tilde{y} - \dot{y} \|_Y^2 \right\} = \frac{1}{2\beta L_{by}} \| Ax - c \|_{Y, *}^2 + \langle \dot{y}, Ax - c \rangle.
\]
The last equality comes from the formula of the Fenchel conjugate of the squared norm. Combining this inequality with (A.13), we obtain (2.13).

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

**Lemma A.2.** Let us denote \( S_k := P_{\beta_k}(\hat{x}_k) - P^* = f(\hat{x}_k) + g_{\beta_k}(A\hat{x}_k; \dot{y}) - f(x^*) - g(Ax^*) \). If \( \tau_k \in (0, 1) \), then
\[
S_{k+1} + \frac{L_A \tau_k^2}{\beta_k+1} \| \hat{x}^{k+1} - x^* \|_X^2 \leq (1 - \tau_k) S_k + \frac{L_A \tau_k^2}{\beta_k+1} \| \hat{x}^k - x^* \|_X^2
+ (1 - \tau_k) [ (\beta_k - \beta_{k+1}) L_{by} - \beta_{k+1} \tau_k ] \| \nabla g_{\beta_{k+1}}(A\hat{x}_k; \dot{y}) - \dot{y} \|_Y^2.
\] (A.14)

**Proof.** Using Lemma A.1 with \( h := g \) and \( h_\beta := g_\beta \), \( \mathcal{Z} := \mathcal{Y} \) and \( z := Ax \), we can proceed as
\[
f(\hat{x}^{k+1}) + g_{\beta_{k+1}}(A\hat{x}^{k+1}; \dot{y}) \leq f(\hat{x}^{k+1}) + g_{\beta_{k+1}}(A\hat{x}^{k+1}; \dot{y}) + \langle \nabla g_{\beta_{k+1}}(A\hat{x}_k; \dot{y}), A\hat{x}^{k+1} - A\hat{x}_k \rangle
+ \frac{1}{2\beta_{k+1}} \| A\hat{x}^k - A\hat{x}^{k+1} \|_{Y, *}^2
\]
\[
\nabla g_\gamma = \frac{g_\gamma}{2\beta_{k+1}} \leq f(x^{k+1}) + g_{\beta_{k+1}}(A\hat{x}_k; \dot{y}) + \langle A^T y^*_{\beta_{k+1}}(A\hat{x}_k; \dot{y}), \hat{x}^{k+1} - \hat{x}_k \rangle
+ \frac{L_A}{2\beta_{k+1}} \| \hat{x}_k - \hat{x}^{k+1} \|_X^2
\]
def. of \( x_{k+1} \)
\[
\leq f(\hat{x}_k) + g_{\beta_{k+1}}(A\hat{x}_k; \dot{y}) + \langle A^T y^*_{\beta_{k+1}}(A\hat{x}_k; \dot{y}), x - \hat{x}_k \rangle
+ \frac{L_A}{2\beta_{k+1}} \| \hat{x}_k - x \|_X^2 - \frac{L_A}{2\beta_{k+1}} \| \hat{x}^{k+1} - x \|_X^2.
\] (A.15)
where the last inequality comes from the definition of $\tilde{x}^{k+1}$ by using its optimality condition and the functions value at $x \in \mathcal{X}$.

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

$$
x - \tilde{x}^k = (1 - \tau_k)\tilde{x}^k + \tau_k x^* - (1 - \tau_k)\tilde{x}^k - \tau_k \tilde{x}^k = \tau_k (x^* - \tilde{x}^k),
$$

$$
x - \tilde{x}^k = (1 - \tau_k)\tilde{x}^k + \tau_k x^* - (1 - \tau_k)\tilde{x}^k - \tau_k \tilde{x}^k = (1 - \tau_k)(\tilde{x}^k - \tilde{x}^k) + \tau_k (x^* - \tilde{x}^k),
$$

$$
x - \tilde{x}^{k+1} = (1 - \tau_k)\tilde{x}^k + \tau_k x^* - \tilde{x}^k - \tau_k (\tilde{x}^{k+1} - \tilde{x}^k) = \tau_k (x^* - \tilde{x}^{k+1}).
$$

Now, we plug these expressions into (A.15) and using the convexity of $f$, we can derive

$$
f(\tilde{x}^{k+1}) + g_{\beta_{k+1}}(A\tilde{x}^{k+1}; \hat{y}) \leq (1 - \tau_k)f(\tilde{x}^k) + \tau_k f(x^*) + g_{\beta_{k+1}}(A\tilde{x}^k; \hat{y}) + \tau_k \langle A^\top y_{\beta_{k+1}}(A\tilde{x}^k; \hat{y}), x^* - \tilde{x}^k \rangle + (1 - \tau_k)\langle A^\top y_{\beta_{k+1}}(A\tilde{x}^k; \hat{y}), \tilde{x}^k - \tilde{x}^k \rangle + \frac{\tilde{L}_A^2}{2\beta_{k+1}}\|\tilde{x}^k - x^*\|_Y^2 - \frac{\tilde{L}_A^2}{2\beta_{k+1}}\|\tilde{x}^{k+1} - x^*\|_Y^2.
$$

(\text{A.5})

$$
(1 - \tau_k)f(x^*) + \tau_k f(x^*) + \tau_k g(Ax^*) - \tau_k \beta_{k+1} b_Y (\nabla g_{\beta_{k+1}}(A\tilde{x}^k; \hat{y}), \hat{y}) + (1 - \tau_k)\beta_{k+1}^2 \|
abla g_{\beta_{k+1}}(A\tilde{x}^k; \hat{y}) - \nabla g_{\beta_{k+1}}(A\tilde{x}^k; \hat{y})\|^2_Y + \frac{\tilde{L}_A^2}{2\beta_{k+1}}\|\tilde{x}^k - x^*\|_2^2 - \frac{\tilde{L}_A^2}{2\beta_{k+1}}\|\tilde{x}^{k+1} - x^*\|_2^2
$$

(\text{A.6})

Finally, using the $L_{b_Y}$-Lipschitz continuity of $\nabla b_Y$ in the weighted-norm $\| \cdot \|_Y$ and the fact that $\nabla b_Y(\hat{y}, \hat{y}) = 0$, we obtain (A.14) from the last derivation. $\square$

\textbf{A.3.1. The proof of Lemma 3.1: Small smoothed primal optimality gap.} Let us denote $S_k := P_{\beta_{k+1}}(\tilde{x}^k; \hat{y}) - P^* = f(\tilde{x}^k) + g_{\beta_k}(A\tilde{x}^k; \hat{y}) - f(x^*) - g(Ax^*)$. Using (A.14) from Lemma A.2 we have

$$
S_{k+1} + \frac{\tilde{L}_A^2}{\beta_{k+1}}\|\tilde{x}^{k+1} - x^*\|_X^2 \leq (1 - \tau_k)S_k + \frac{\tilde{L}_A^2}{\beta_{k+1}}\|\tilde{x}^k - x^*\|_X^2
$$

$$
+ \frac{(1 - \tau_k)}{2}[(\beta_k - \beta_{k+1})L_{b_Y} - \beta_{k+1}\tau_k]\|
abla g_{\beta_{k+1}}(A\tilde{x}^k; \hat{y}) - \hat{y}\|^2_Y.
$$

(A.16)

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

$$
(\beta_k - \beta_{k+1})L_{b_Y} = \beta_{k+1}\tau_k \quad \text{and} \quad (1 - \tau_k)\frac{\beta_{k+1}}{\tau_k} = \frac{\beta_k}{\tau_{k-1}}.
$$

(A.17)
By eliminating $\beta_k$ and $\beta_{k+1}$ from these equalities, we obtain $\tau_k^2(1 + \tau_k/L_{by}) = \tau_{k-1}^2(1 - \tau_k)$. Hence, we can compute $\tau_k$ by solving the cubic equation

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

(A.18)

At the same time, we also obtain from (A.17) an update rule $\beta_{k+1} := \frac{\beta_k}{1 + \frac{\tau_k}{L_{by}}}$.

Now, we show that (A.18) has a unique positive solution $\tau_k \in (0, 1)$ for any $L_{by} \geq 1$ and $\tau_{k-1} \in (0, 1]$. We consider the cubic polynomial $p_3(\tau)$ defined by the left-hand side of (A.18). Clearly, for any $\tau > 0$, we have $p_3'(\tau) = 3\tau^2/L_{by} + 2\tau_k + \tau_{k-1}^2 > 0$. Hence, $p_3(\cdot)$ is monotonically increasing on $(0, +\infty)$. In addition, since $p_3(0) = -\tau_{k-1}^2 < 0$ and $p_3(1) = 1/L_{by} + 1 > 0$, the equation (A.18) has only one positive solution $\tau_k \in (0, 1)$.

Next, we show that $\tau_k \leq \frac{2}{k+2}$. Indeed, by (A.18) we have $p_3(\tau) \geq \tau^2 + \tau_{k-1}^2 - \tau_{k-1}^2 := p_2(\tau)$. Since the unique positive root of $p_2(\tau) = 0$ is $\tilde{\tau}_k := \frac{\tau_{k-1}}{2},$ we have $p_3(\tau) \geq p_2(\tilde{\tau}_k)$ for $\tau \geq \tilde{\tau}_k$. As $p_3(\tau)$ is monotonically increasing on $\mathbb{R}_+$, its positive solution $\tau_k$ much be in $(0, \tilde{\tau}_k]$. Hence, we have $\tau_k \leq \frac{\tau_{k-1}}{2} \left(\sqrt{\frac{2}{\tilde{\tau}_k}} + 4 - \tau_{k-1}\right)$. 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 $\tau_0$, 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 (A.17) we have

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

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

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

$$\beta_{k+1} = \frac{\beta_k}{1 + \tau_k/L_{by}} \leq \frac{\beta_k}{k + 1 + L_{by}} \leq \frac{l+1}{l+1 + L_{by}} \to 0 \text{ as } k \to \infty.$$ Clearly, if $L_{by} = 1$, then $\beta_{k+1} = \frac{\beta_k}{1 + \tau_k/L_{by}} \leq \frac{k+1}{k+2} \beta_k \leq \frac{2\beta_k}{k+2}$ by induction.

Finally, we upper bound the ratio $r_k^2/\beta_{k+1}$ using the second equality in (A.17) as

$$\frac{\tau_{k-1}^2}{\beta_{k+1}} = \frac{\tau_{k-1}^2}{\beta_k} (1 - \tau_k) \leq \frac{\tau_{k-1}^2}{\beta_k} \prod_{l=1}^k (1 - \tau_l) \leq \frac{\tau_{k-1}^2}{\beta_1} \prod_{l=1}^k (1 - \frac{1}{l+1}) = \frac{\tau_{k-1}^2}{\beta_1} \frac{k}{k+1}.$$ Using these relations into (A.16) we obtain

$$\frac{\beta_{k+1}}{\tau_k^2} \sum_{k} + \frac{L_A}{2} \|\bar{x} - x^*\|^2_{\gamma} \leq \frac{\beta_k}{\tau_{k-1}} \sum_{k} + \frac{L_A}{2} \|x - x^*\|^2_{\gamma} \leq \frac{\beta_0}{\tau_0} \sum_{k} + \frac{L_A}{2} \|x_0 - x^*\|^2_{\gamma},$$

we get (3.1) with noting that $\tau_0 = 1$, the bound on $\frac{\tau_k^2}{\beta_{k+1}}$ and $\sum_{k}$ := $P_{\beta_{k+1}}(\bar{x}) - P^*$. □

A.4. The proof of Lemma 4.1: Gap reduction in ADSGARD. For simplicity of notation, we denote by $f_k^*(y) := f_{\gamma_k}^*(-A^T y; \hat{x})$ defined by (2.9), $\bar{y}_k^* := y_{\gamma_k}^*(A\bar{x}; \hat{y})$, $\hat{x}_{k+1}^* := x_{\gamma_k}^*(\hat{y}; \hat{\tilde{x}})$ and $\tilde{x}_{k+1}^* := x_{\gamma_k}^*(\hat{y}; \hat{x})$. By (A.3), $\nabla f_k^*$ is Lipschitz continuous with the Lipschitz constant $L_{\nabla f_k^*} := \gamma^{-1}$ and thus $\nabla f_k^*$ is Lipschitz continuous with the Lipschitz constant $\gamma_{k+1}^{-1} L_A$. 23
First, using the optimality condition for problem (2.9), we obtain
\[
f(\hat{x}^k) + (\beta_k b y, y) + \beta_k b y, y + g^*(y) \leq P_{\beta_k}(\hat{x}^k; y) - (\beta_k/2)\|y - \hat{y}_k\|^2_y. \tag{A.19}
\]
Second, using the definition of \(f'_{\gamma}(\cdot; \hat{x})\) in (2.5), we can show that
\[
(\hat{x}_{k+1}, y) + f(\hat{x}_{k+1}) = \gamma_{k+1}b x(\hat{x}_{k+1}, \hat{x}) - f'_{\gamma}(\hat{y}^k) + (\hat{x}_{k+1}, y - \hat{y}^k) \tag{A.20}
\]
Third, using (A.2) for \(f^*_{\gamma}(\cdot; \hat{x})\) and the co-coercivity (A.4) of \(f'_{\gamma}(\cdot; \hat{x})\), we can derive
\[
-D\gamma_k(\hat{y}^k; \hat{x}) = f^*_{\gamma_k}(-A^T \hat{y}^k; \hat{x}) + g^*(\hat{y}^k) \tag{A.21}
\]
Then, by the definition of \(\hat{y}^{k+1}\), we can write
\[
D\gamma_{k+1}(\hat{y}^{k+1}; \hat{x}) = -g^*(\hat{y}^{k+1}) - f'_{\gamma}(\hat{y}^{k+1}) + \gamma_{k+1}b x(\hat{x}_{k+1}, \hat{x}) \tag{A.22}
\]
Using these relations, the definition of \(\hat{x}^{k+1}\) and the convexity of \(f\), we have
\[
P_{\beta_k}(\hat{x}^{k+1}; y) = f(\hat{x}^{k+1}) + \max_{y \in Y} \{ (A\hat{x}^{k+1}, y) - g^*(y) - \beta_k b y, y \} \tag{A.13}
\]
\[
\leq \max_{y \in Y} \left\{ (1 - \tau_k) \left[ f(\hat{x}) + (A\hat{x}, y) - \beta_k b y, y + g^*(y) \right] + \tau_k \left[ (A\hat{x}_{k+1}, y) + f(\hat{x}_{k+1}) + g^*(y) \right] \right\} \tag{A.19} + \tag{A.20}
\]
\[
(1 - \tau_k)P_{\beta_k}(\hat{x}^k; \hat{y}) - \gamma_{k+1}b x(\hat{x}_{k+1}, \hat{x}) \tag{A.21}
\]
\[
- \min_{y \in Y} \left\{ \tau_k f_k(\hat{y}^k) + \tau_k(\nabla f_k(\hat{y}^k), y - \hat{y}^k) + \frac{(1 - \tau_k)\beta_k}{2} \|y - \hat{y}_k\|^2_y + \tau_k g^*(y) \right\} \tag{A.22}
\]
Let us define an auxiliary term \( T_k \) as
\[
T_k := (1 - \tau_k)\frac{r_{k+1}}{2} \|x^*_k\|^2 - (1 - \tau_k)(\gamma_k - \gamma_{k+1})b_X(x^*_k, \hat{x}) + \tau_k \gamma_{k+1}b_X(\hat{x}^*_{k+1}, \hat{x})
\]
and let us consider the change of variable \( u := (1 - \tau_k)\hat{y}^k + \tau_k y \) for \( y \in \mathcal{Y} \). Then \( u \in \mathcal{Y} \) and \( u - \hat{y}^k = \tau_k(y - \hat{y}^k) \) we have
\[
P_{\beta_{k+1}}(\hat{x}^*_{k+1}; \hat{y}^k) \leq (1 - \tau_k)G_{\gamma_0}\beta_k(\hat{w}^k; \hat{w}) - T_k - \min_{u \in \mathcal{Y}} \left\{ f_k^*(\hat{y}^k) + (\nabla f_k^*(\hat{y}^k), u - \hat{y}^k) + \frac{(1 - \tau_k)\beta_k}{2\tau_k} \|u - \hat{y}^k\|_Y^2 + g^*(u) \right\} \leq \frac{A.22 + 1}{1 - \tau_k} G_{\gamma_0}\beta_k(\hat{w}^k; \hat{w}) + D_{\gamma_{k+1}}(\hat{y}^*_{k+1}; \hat{x}) - T_k,
\]
Finally, we estimate \( T_k \) in (A.23) using the strong convexity of \( b_X(\cdot, \hat{x}) \) as follows:
\[
2T_k \geq (1 - \tau_k)\gamma_{k+1} \|x_{k+1} - \hat{x}^*_{k+1}\|^2 + \tau_k \gamma_{k+1} \|x_{k+1} - \hat{x}\|^2 - (1 - \tau_k)(\gamma_k - \gamma_{k+1})L_{b_X} \|x^*_{k+1} - \hat{x}\|^2 \geq (1 - \tau_k)[\gamma_k\gamma_{k+1} - (\gamma_k - \gamma_{k+1})L_{b_X}] \|x_{k+1}^* - \hat{x}\|^2 \geq 0.
\]
Substituting (A.25) into (A.24), we get \( G_{\gamma_{k+1}}\beta_{k+1}(\hat{w}^{k+1}; \hat{w}) \leq (1 - \tau_k)G_{\gamma_0}\beta_k(\hat{w}^k; \hat{w}) \).

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
\[
G_{\gamma_0}\beta_1(\hat{w}^1; \hat{w}) \leq -\gamma_1 b_X(\bar{x}^1, \bar{x}) + \frac{L_A}{2\gamma_1} \|\bar{y}_0^* - \bar{y}^*_0\|^2 - \beta_1 b_Y(\bar{y}^*_1, \hat{y})
\]
As \( \beta_1 \gamma_1 \geq L_A \) and \( \bar{y}^*_0 := \hat{y} \), we get \( G_{\gamma_0}\beta_1(\hat{w}^1; \hat{w}) \leq 0 \).

Next, we set the equality in three conditions of (4.14) to get \( \gamma_{k+1} = \gamma_k(1 + \tau_k/L_{b_X})^{-1}, \beta_{k+1} = (1 - \tau_k)\beta_k \) and \( (1 - \tau_k)\beta_{\gamma_{k+1}} = \tau_k^2 L_A \). In particular, \( \gamma_0, \beta_0, \gamma_1, \beta_1, \gamma_2, \beta_2, \gamma_3, \beta_3, \ldots \) and \( \beta_{\gamma_1}, \beta_{\gamma_2}, \beta_{\gamma_3}, \ldots \) are the unique positive solutions of the cubic equation \( p_3(\tau) := \tau^2 + \tau + \tau^2 \cdot \tau - \tau^2 = 0 \). Hence, similar to the proof of Lemma 3.1, we can show that \( \tau_k \in (0, 1) \) is the unique positive solution of the cubic equation \( p_3(\tau) := \tau^2 + \tau + \tau^2 \cdot \tau - \tau^2 = 0 \). In addition, \( \frac{A.1}{k+1} \leq \tau_k \leq \frac{A.1}{k+2} \) for \( k \geq 1 \) and \( \tau_0 = 1 \).

If \( L_{b_X} = 1 \), then \( \gamma_{k+1} = \frac{\gamma_k}{1 + \gamma_k/L_{b_X}} \leq \frac{\gamma_k}{1 + 1} \leq \frac{\gamma_k}{2} \). Similarly, \( \beta_{k+1} = (1 - \tau_k)\beta_k \leq \frac{\beta_k}{k+1} \).

Finally, we note that \( \beta_{\gamma_1} = \frac{\tau L_A}{\gamma_0} \geq \frac{\tau}{L_A} \geq \frac{k+2}{2\gamma_1} \geq \frac{\tau}{\gamma_0} \).

**A.5. The proof of Proposition 5.1: The accelerated augmented Lagrangian method.**

First of all with the choice of norms associated to the Lagrangian smoother, we have
\[
\hat{L}_A := \|Ax\|^2 = \max_{x \in \mathcal{X}} \left\{ \frac{\|Ax\|_Y^2}{\|x\|_X^2} \right\} = \max_{x \in \mathcal{X}} \left\{ \frac{\|Ax\|_Y^2}{\|x\|_X^2} \right\} = 1.
\]
Secondly, note that the conclusions of Lemma 4.1 are valid for any semi-norm. In particular, if we choose \( \beta_1 \gamma_0 \geq \hat{L}_A = 1 \),
\[
\gamma_{k+1} = \gamma_0 \geq \frac{\gamma_0}{1 + \gamma_0/L_{b_X}}, \quad \beta_{k+1} = (1 - \tau_k)\beta_k, \quad \text{and} \quad \frac{\hat{L}_A}{\gamma_0} = \frac{(1 - \tau_k)\beta_k}{\gamma_k}
\]
then \( G_{\gamma_0, \beta_{k+1}}(\hat{w}^{k+1}, \hat{w}) \leq (1 - \tau_k)G_{\gamma_0, \beta_k}(\hat{w}^k, \hat{w}) \leq 0 \).
Eliminating $\beta_{k+1}$ and $\beta_k$ in the equalities, we get
\[
\frac{\tau_{k+1}^2}{1 - \tau_{k+1}} = \tau_k^2.
\]
One can easily check by induction that $\beta_k = \beta_1 \prod_{i=1}^{k} (1 - \tau_i) = \beta_1 \sqrt{\frac{\tau_k}{\gamma_0}} = \frac{\tau_k}{\gamma_0}$ and $\tau_k \leq \frac{2}{k+2}$. We then conclude using Lemma 2.1 and the fact that $b_\chi(x^*, \hat{x}) = 0$:
\[
S_{\beta_k}(\hat{x}^k) \leq G_{\gamma_0, \beta_k}(\bar{w}^k, \bar{w}) + 0 \leq 0,
\]
and
\[
f(\hat{x}^k) - f^* \leq S_{\beta_k}(\hat{x}^k) - (y^*, A\hat{x}^k - c) + \beta_k b_{\gamma}(y^*; \hat{y}) \\
\leq \|y^*\|\|A\hat{x}^k - c\|_{\gamma, \gamma} + \beta_k b_{\gamma}(y^*; \hat{y}) \leq \frac{8L_{b_{\gamma}}\|y^*\|\|y^* - \hat{y}\|_\gamma + 4b_{\gamma}(y^*; \hat{y})}{\gamma_0(k+2)^2},
\]
and
\[
f(\hat{x}^k) - f^* \geq -\|y^*\|\|Ax - c\|_{\gamma, \gamma} \geq -\frac{8L_{b_{\gamma}}\|y^*\|\|y^* - \hat{y}\|_\gamma}{\gamma_0(k+2)^2}.
\]

The proposition is proved. \qed

A.6. The proof of Proposition 5.2. The strongly convex objective case. The proof follows the line of Lemma 4.1. We only need to replace the Lipschitz continuity coefficient $\frac{L_{\gamma}A}{\gamma_0}$ by $L_{J_{k+1}} = \frac{L_{\gamma}A}{\mu_j}$ in (A.22) and replace all other occurrences of $\gamma_{k+1}$ by 0. Under a choice of parameters satisfying (5.5), we obtain the gap reduction condition $G_{0, \beta_k+1}(\bar{w}^{k+1}; \bar{w}) \leq (1 - \tau_k)G_{0, \beta_k}(\bar{w}^k; \bar{w}) \leq 0$ as in Lemma 4.1. We can also check by induction that $\beta_k \leq \frac{4}{(k+2)^2} \frac{L_{\gamma}A}{\mu_j}$. We obtain the conclusion of Proposition 5.2 by using Lemma 2.1. \qed

Appendix B. This work is supported in part by the NSF-grant No. DMS-1619884, USA; and the European Commission under the grants MIRG-268398 and ERC Future Proof, and by the Swiss Science Foundation under the grants SNF 200021-132548, SNF 200021-146750 and SNF CRSII2-147633.

REFERENCES

[1] H. Bauschke and P. Combettes, Convex analysis and monotone operators theory in Hilbert spaces, Springer-Verlag, 2011.
[2] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.
[3] A. Beck and M. Teboulle, Smoothing and first order methods: A unified framework, SIAM J. Optim., 22 (2012), pp. 557–580.
[4] A. Beck and M. Teboulle, A fast dual proximal gradient algorithm for convex minimization and applications, Oper. Res. Letter, 42 (2014), pp. 1–6.
[5] A. Ben-Tal and A. Nemirovski, Lectures on modern convex optimization: Analysis, algorithms, and engineering applications, vol. 3 of MPS/SIAM Series on Optimization, SIAM, 2001.
[6] D. Bertsekas and J. N. Tsitsiklis, Parallel and distributed computation: Numerical methods, Prentice Hall, 1989.
[7] D. P. Bertsekas, Constrained Optimization and Lagrange Multiplier Methods, Athena Scientific, 1996.
[8] R. I. Boţ and C. Hendrich, A variable smoothing algorithm for solving convex optimization problems, TOP, 23 (2012), pp. 124–150.
[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends in Machine Learning, 3 (2011), pp. 1–122.
[10] S. Boyd and L. Vandenberghe, Convex Optimization, University Press, Cambridge, 2004.
[11] 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.

[12] V. Cevher, S. Becker, and M. Schmidt, Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics, IEEE Signal Processing Magazine, 31 (2014), pp. 32–43.

[13] A. Chambolle and T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, J. Math. Imaging Vis., 40 (2011), pp. 120–145.

[14] V. Chandrasekaranm, B. Recht, P. A. Parrilo, and A. S. Willsky, The convex geometry of linear inverse problems, Foundations of Computational Mathematics, 12 (2012), pp. 805–849.

[15] G. 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.

[16] G. Chen and M. Teboulle, A proximal-based decomposition method for convex minimization problems, Math. Program., 64 (1994), pp. 81–101.

[17] 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.

[18] D. Davis, Convergence rate analysis of primal-dual splitting schemes, UCLA CAM report 14-63, (2014).

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

[20] D. Davis and W. Yin, Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions, UCLA CAM report 14-58, (2014).

[21] W. Deng, M.-J. Lai, Z. Peng, and W. Yin, Parallel multi-block ADMM with o(1/k) convergence, arXiv preprint arXiv:1312.3040, (2013).

[22] 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.

[23] J. E. Esser, Primal-dual algorithm for convex models and applications to image restoration, registration and nonlocal inpainting, PhD Thesis, University of California, Los Angeles, Los Angeles, USA, 2010.

[24] F. Facchinei and J.-S. Pang, Finite-dimensional variational inequalities and complementarity problems, vol. 1–2, Springer-Verlag, 2003.

[25] P. Giselsson and S. Boyd, Monotonicity and Restart in Fast Gradient Methods, in IEEE Conference on Decision and Control, Los Angeles, USA, December 2014, CDC.

[26] T. Goldstein, E. Esser, and R. Baraniuk, Adaptive primal-dual hybrid gradient methods for saddle point problems, Tech. Report., (2013), pp. 1–26. http://arxiv.org/pdf/1305.0546v1.pdf.

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

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

[29] 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.

[30] M. Hong and Z.-Q. Luo, On the linear convergence of the alternating direction method of multipliers, arXiv preprint arXiv:1208.3922, (2012).

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

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

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

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

[35] T. Lin, S. Ma, and S.-Z. Zhang, On the sublinear convergence rate of multi-block ADMM, Journal of the Operations Research Society of China, 3 (2015), pp. 251–274.

[36] M. B. McCoy, V. Cevher, Q. Tran-Dinh, A. Asaee, and L. Baldassarre, Convexity in source separation: Models, geometry, and algorithms, IEEE Signal Processing Magazine, 31 (2014), pp. 87–95.

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

[38] 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.

[39] 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.

[40] 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.

[41] Y. Nesterov, A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$, Doklady AN SSSR, 269 (1983), pp. 543–547.
[42] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87 of Applied Optimization, Kluwer Academic Publishers, 2004.

[43] Y. Nesterov, *Excessive gap technique in nonsmooth convex minimization*, SIAM J. Optim., 16 (2005), pp. 127–152.

[44] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 235–249.

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

[46] B. O’Donoghue and E. Candès, *Adaptive Restart for Accelerated Gradient Schemes*, Found. Comput. Math., 15 (2015), pp. 715–732. [http://www-stat.stanford.edu/~candes/publications.html](http://www-stat.stanford.edu/~candes/publications.html).

[47] H. Ouyang, N. He, L. Q. Tran, and A. Gray, *Stochastic alternating direction method of multipliers*, JMLR W&CP, 28 (2013), pp. 80–88.

[48] 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.

[49] R. T. Rockafellar, *Convex Analysis*, vol. 28 of Princeton Mathematics Series, Princeton University Press, 1970.

[50] 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.

[51] W. Su, S. Boyd, and E. Candès, *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.

[52] Q. Tran-Dinh and V. Cevher, *Constrained convex minimization via model-based excessive gap*, in Proc. the Neural Information Processing Systems Foundation conference (NIPS2014), vol. 27, Montreal, Canada, December 2014, pp. 721–729.

[53] Q. Tran-Dinh and V. Cevher, *A primal-dual algorithmic framework for constrained convex minimization*, Tech. Report., LIONS, (2014), pp. 1–54.

[54] Q. Tran-Dinh and V. Cevher, *Splitting the smoothed primal-dual gap: Optimal alternating direction methods*, Tech. Report. (LIONS, EPFL), (2015), [http://arxiv.org/abs/1507.03734](http://arxiv.org/abs/1507.03734).

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

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

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