Noname manuscript No.
(will be inserted by the editor)

An optimal first-order primal-dual gap reduction framework for constrained convex optimization

Quoc Tran-Dinh *and* Volkan Cevher

Received: date / Accepted: date

Abstract We introduce an analysis framework for constructing *optimal* first-order primal-dual methods for the prototypical constrained convex optimization template. While this class of methods offers scalability advantages in obtaining numerical solutions, they have the disadvantage of producing sequences that are only approximately feasible to the problem constraints. As a result, it is theoretically challenging to compare the efficiency of different methods. To this end, we rigorously prove in the worst-case that the convergence of primal objective residual in first-order primal-dual algorithms must compete with their constraint feasibility convergence, and mathematically summarize this fundamental trade-off. We then provide a heuristic-free analysis recipe for constructing optimal first-order primal-dual algorithms that can obtain a desirable trade-off between the primal objective residual and feasibility gap and whose iteration convergence rates cannot be improved. Our technique obtains a smoothed estimate of the primal-dual gap and drives the smoothness parameters to zero while simultaneously minimizing the smoothed gap using problem first-order oracles.

Keywords: Model-based gap reduction technique; first-order primal-dual methods; augmented Lagrangian; smoothing techniques; separable convex minimization; parallel and distributed computation.

1 Introduction

We propose a new primal-dual analysis framework for constructing *optimal* first-order primal-dual methods in order to obtain numerical solutions to the following constrained convex optimization template:

$$f^\star := \min_{x\in\mathbb{R}^n} \left\{ f(x) : Ax - b \in \mathcal{K},\ x \in \mathcal{X} \right\}, \tag{1}$$

Quoc Tran-Dinh
Department of Statistics and Operations Research
The University of North Carolina at Chapel Hill (UNC), USA
E-mail: quoc.trandinh@epfl.ch

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

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper, closed and convex function; $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{K} \subset \mathbb{R}^m$ are nonempty, closed and convex sets; and $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are given. We assume that the domain $\mathcal{X}$ is "simple" so that it is easy to project.

The template (1) provides a unified formulation for a broad set of applications in various disciplines, see, e.g., [9,11,12,14,26,39,55]. Clearly, unconstrained problems can also be reformulated as (1) via splitting variables [10]. Moreover, it covers standard convex optimization subclasses, such as conic programming, monotropic programming, and geometric programming, as specific instances [7,8,10].

There are several reasons for our emphasis on first-order primal-dual methods for (1), with the most obvious one being their scalability. Coupled with recent demand for low-to-medium accuracy solutions in applications, these methods indeed provide a critical trade-off between the complexity-per-iteration and the iteration-convergence rate along with the ability to distribute and decentralize computation.

Unfortunately, the newfound popularity of primal-dual optimization has lead to an explosion in the number of different algorithmic variants, each of which requires different set of assumptions on problem settings or methods, such as strong convexity, Lipschitz gradient, or penalty parameter tuning. 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.

To this end, we address the following key question in this paper: "Can we construct heuristic-free, optimal first-order primal-dual methods?" The concept of an optimal algorithm in the setting of (1) has been elusive since virtually all methods produce primal sequences that are infeasible in the constraints of the form $Ax - b \in \mathcal{K}$. To overcome this challenge, we mathematically characterize the best rates on the primal objective residual and the constraint feasibility gap of algorithmic iterates and illustrate how they compete with each other, while requiring only a mild set of assumptions on the template (1). We then provide an analysis recipe for constructing optimal first-order primal-dual algorithms in a heuristic-free fashion whose convergence rates cannot be improved.

1.1 The role of the primal-dual gap function

It is natural to expect the constraints to slow down a minimization process for the primal optimality and not the constraint feasibility, and vice versa. To mathematically understand the basic issue, we need to study the primal-dual gap function where the two quantities are entangled. For notational ease, let us consider here a special case of (1) where $\mathcal{K} = 0$. The primal-dual gap function G is then given by

$$G(w) := \underbrace{\max\big\{f(x)+\langle Ax-b,\hat{y}\rangle : \hat{y}\in\mathbb{R}^m\big\}}_{\bar{f}(x)} - \underbrace{\min\big\{f(\hat{x})+\langle A\hat{x}-b,y\rangle : \hat{x}\in\mathcal{X}\big\}}_{g(y)}, \quad (2)$$

where $w := (x, y)$ is the concatenated primal-dual variables; $\bar{f}$ is the extended function of f; and g is the Lagrange dual function associated with (1).

The primal-dual gap function G is convex. Moreover, under strong duality (cf., Lemma 1), we have $G(w^\star) = 0$ if and only if $w^\star = (x^\star, y^\star)$ is a primal-dual solution of (1). Since the gap function G is generally nonsmooth but has a max-structure, we can obtain a smoothed estimate using two smoothing functions p_x and p_y as

$$G_{\gamma\beta}(w) := \max_{\hat{y}\in\mathbb{R}^m}\big\{f(x)+\langle Ax-b,\hat{y}\rangle-\beta p_y(\hat{y})\big\} - \min_{\hat{x}\in\mathcal{X}}\big\{f(\hat{x})+\langle A\hat{x}-b,y\rangle-\gamma p_x(\hat{x})\big\}, \quad (3)$$

where γ and β in $\mathbb{R}_{++}$ are two corresponding smoothness parameters. Note that if we choose $p_x(\hat{x}) = \frac{1}{2}\|\hat{x}\|_2^2$, then the dual solution can be computed by the proximal operator of $f + \delta_{\mathcal{X}}$, where $\delta_{\mathcal{X}}$ is the indicator function of $\mathcal{X}$.

With the smoothed gap setup in (3), we show (*cf.*, Section 3) that there exists a primal sequence $\{\bar{x}^k\} \subset \mathcal{X}$ and a dual sequence $\{\bar{y}^k\} \subset \mathbb{R}^m$ such that the following inequalities for the smoothed gap, objective, and the feasibility hold

$$\begin{aligned} G_{\gamma_{k+1}\beta_{k+1}}(\bar{w}^{k+1}) &\leq (1-\tau_k) G_{\gamma_k\beta_k}(\bar{w}^k) + \psi_k, \\ f(\bar{x}^k) - f^\star \leq \quad f(\bar{x}^k) - g(\bar{y}^k) \quad &= \mathcal{O}\left(\gamma_k + G_{\gamma_k\beta_k}(\bar{w}^k)\right), \\ \|A\bar{x}^k - b\|_2 \quad &= \mathcal{O}\left(\beta_k + \sqrt{\beta_k(\gamma_k + G_{\gamma_k\beta_k}(\bar{w}^k))}\right), \end{aligned} \tag{4}$$

for *any* choices of β_k and γ_k as long as $\gamma_k\beta_k = \mathcal{O}\left(\tau_k^2\right)$, where $\tau_k \in (0,1)$ and $\psi_k = \mathcal{O}(\tau_k)$. Against intuition, it appears possible to obtain arbitrarily fast rates via the choice of the rate-parameters β_k and γ_k under a given gap reduction model.

Naturally, there exists lower complexity bounds for minimizing sequences, depending on the chosen optimization oracles [31,33]. In the smoothed gap setting, we rely only on first-order oracles, i.e., computational primitives based on matrix-vector products involving A and A^T and the proximal operators of f and $f+\delta_{\mathcal{X}}$ As a result, without any further assumption on (1), it holds that $f(\bar{x}^k) - g(\bar{y}^k) = \Omega\left(\frac{1}{k}\right)$, which implies that $\tau_k = \Omega\left(\frac{1}{k}\right)$, $\gamma_k = \Omega\left(\frac{1}{k}\right)$, and $\beta_k = \Omega\left(\frac{1}{k}\right)$: *cf.*, [34,35].

1.2 Towards optimal first-order primal-dual methods: Model-based gap reduction. We say that a first-order primal-dual algorithm is *optimal* if its primal objective residual and feasibility gap convergence rates satisfy $f(\bar{x}^k) - f^\star = \mathcal{O}\left(\frac{1}{k}\right)$ and $\operatorname{dist}\left(A\bar{x}^k - b, \mathcal{K}\right) = \mathcal{O}\left(\frac{1}{k}\right)$, where $\operatorname{dist}(\cdot, \mathcal{K})$ measures the Euclidean distance to the set $\mathcal{K}$. Then, it remains to show that we can indeed construct optimal first-order convex optimization methods. For this purpose, we introduce a model-based gap reduction technique where the rate of parameters γ_k and β_k in (4) plays an interpretable role in minimizing the smoothed gap, following the model in (4).

To construct algorithms, we exploit the obvious correspondence between the duality gap function G and its smoothed estimate $G_{\gamma\beta}$: The first max-term presents an approximation to the primal objective f, and the second min-term provides an approximation to the dual objective g. Depending on the choice of p_x and p_y, we obtain different smoothed approximations and hence, we can develop different algorithms for solving (1). Regarding the choice of smoothing functions, two approaches stand out in the literature: (*i*) proximity smoothing, and (*ii*) barrier smoothing [6,19,20,24,27,28,34,36,51,53,54,58]. In this work, we demonstrate our results using the proximity smoothing technique [4,28,34,35,54].

With the smoothed gap setup in (3), we generate a primal-dual sequence $\{\bar{w}^k\}$ in $\mathcal{X} \times \mathbb{R}^m$ satisfying (4) with first-order oracles such that $\{G_{\gamma_k\beta_k}(\bar{w}^k)\}$ converges to zero, while simultaneously decreasing the product $\{\gamma_k\beta_k\}$ to zero. Among various strategies, we focus on the model-based gap reduction condition in (4) as well as its monotone version where $\psi_k \leq 0$ for all $k \geq 0$. We then show that first-order oracle information from (1) is sufficient to obtain $\tau_k = \mathcal{O}\left(\frac{1}{k}\right)$, and hence, the constructed algorithms are optimal in the sense of black-box models [31,33].

Surprisingly, any attempt to trade-off between the rates of convergences for the primal optimality and the feasibility worsens the overall convergence. Intriguingly, we will show that there still exists a practical trade-off since it must hold that

$$\gamma_k \beta_k = \Omega(\tau_k^2). \tag{5}$$

In the monotone gap-reduction model, in the light of (4), the condition (5) shows how the primal objective residual of the iterates competes with their primal feasibility gap by trading-off the values of the smoothness parameters.

1.3 Scalable methods for (1) and their limitations.
Scalable numerical approaches for solving (1) are mainly based on penalty, augmented Lagrangian, and other primal-dual (splitting) methods, depending on how they process the linear inclusion constraint $Ax - b \in \mathcal{K}$. In the sequel, we focus on the special case $\mathcal{K} = 0$ without loss of generality while reviewing the related work.

With the penalty methods, we can obtain low- or medium-accuracy solutions when we augment the objective f with a simple penalty function, such as

$$\min_{x\in\mathbb{R}^n}\Big\{f(x) + (\rho/2)\|Ax - b\|_2^2 : x \in \mathcal{X}\Big\}, \tag{6}$$

where $\rho > 0$ is the penalty parameter. While penalty methods are widely popular in large-scale applications, their empirical performance is quite sensitive to the choice of the penalty parameter [39].

While penalty and augmented Lagrangian methods have a fundamental difficulty in choosing the penalty parameter, their variants such as primal-dual splitting, AMA and ADMM methods enhance our computational capabilities and numerical robustness since we can apply (accelerated) proximal gradient methods or can distribute the computation: *cf.*, [2,15,25,44,46]. The scalability of these numerical convex optimization algorithms typically rely on two key structures:

Structure 1: ***Decomposability.*** The constrained convex optimization problem (1) is said to be *N-decomposable* if f and $\mathcal{X}$ can be represented as follows:

$$f(x) := \sum_{i=1}^{N} f_i(x_i), \quad \text{and} \quad \mathcal{X} := \prod_{i=1}^{N} \mathcal{X}_i, \tag{7}$$

where $x_i \in \mathbb{R}^{n_i}$, $\mathcal{X}_i \in \mathbb{R}^{n_i}$, $f_i : \mathbb{R}^{n_i} \to \mathbb{R}\cup\{+\infty\}$ is proper, closed and convex for $i = 1, \cdots, N$, and $\sum_{i=1}^{N} n_i = n$. Decomposability immediately supports parallel and distributed implementations in synchronous hardware architectures. This structure arises naturally in linear programming, network optimization, multi-stage models and distributed systems and machine learning [9,10].

Structure 2: ***Proximal tractability.*** Unconstrained problems can still pose significant difficulties in numerical optimization when they include non-smooth terms. However, many non-smooth problems (e.g., of the form (6)) can be solved nearly as efficiently as smooth problems, provided that the computation of the proximal operator is *tractable* [3,42,45]. By tractable proximal operator, we mean that the following strongly convex problem can be solved "efficiently" (e.g., by a closed form solution or by polynomial algorithms) for a given convex function h:

$$\mathrm{prox}_h(x) := \mathrm{argmin}\left\{h(z) + (1/2)\|z - x\|_2^2 : z \in \mathrm{dom}\,(h)\right\}. \tag{8}$$

It has been shown that many smooth and non-smooth functions support tractable proximal operators [12,14,26,16,55]. Clearly, decomposability also proves useful in the computation of (8). In our problem (1), we use the proximal operator of $h(\cdot) \leftarrow f(\cdot) + \delta_{\mathcal{X}}(\cdot)$, where $\delta_{\mathcal{X}}$ is the indicator function of $\mathcal{X}$.

On the basis of these structures, we can design algorithms featuring a full spectrum of (nearly) dimension-independent, global convergence rates for composite

convex minimization problems with well-understood analytical complexities [3,33, 38,37,52]. Unfortunately, several scalable first-order methods for (1) invariably feature one or both of the following two limitations which blocks their full impact.
Limitation 1: ***Non-ideal convergence characterizations.*** Ideally, the convergence characterization of an algorithm for solving (1) must establish rates both on the primal objective residual $f(\bar{x}^k) - f^\star$ and the feasibility gap $\|A\bar{x}^k - b\|$ of its linear constraints separately for its iterates $\bar{x}^k \in \mathcal{X}$. The constraint feasibility is critical so that the primal convergence rate has any significance. Rates on the gap function associated with the optimality condition of (1) concerning the joint primal-dual variables (x, y) are not necessarily meaningful at intermediate iterates since (1) is a constrained problem and $f(\bar{x}^k) - f^\star$ can easily be negative at all times as compared to the unconstrained setting where we trivially have $f(\bar{x}^k) - f^\star \geq 0$.

The convergence results of several existing methods are far from ideal. Most algorithms have guarantees in the ergodic sense (i.e., on the averaged history of iterates) [13,22,23,41,47,56] with non-optimal rates, which diminishes the practical performance; they rely on special function properties to improve convergence rates on the function and feasibility [40,41], which reduces the scope of their applicability; they provide rates on dual functions [21], or a weighted primal residual and feasibility score [47], which does not necessarily imply convergence on the absolute value of the primal residual or the feasibility; or they obtain convergence rate on the gap function value sequence composed both the primal and dual variables via variational inequality and gap function characterizations [13,22,23], where the rate is scaled by a diameter parameter which is not necessary bounded.[1]

Limitation 2: ***Computational inflexibility.*** Recent theoretical developments customize algorithms to exploit special function classes for scalability. When the model parameters are known a priori, this strategy is sensible. Unfortunately, specialized algorithms often require knowledge of function class parameters even if they are not known, and hence do not address the full scope of (1) (e.g., with self-concordant functions or fully non-smooth decompositions). Moreover, they often have complicated algorithmic implementations with backtracking steps to compute some of these parameters, which create computational bottlenecks.

1.4 Our contributions

To this end, the main contributions of this paper can be summarized as follows:

(a) We identify optimal rates of convergence for the objective residual and the feasibility gap in first-order primal-dual methods.
(b) 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. 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.
(c) We illustrate the new techniques enable us to exploit additional structures, including augmented Lagrangian smoothing, strongly convex or Lipschitz continuous gradient of the objectives.
(d) We show the flexibility of our framework applying to different constrained settings including conic programs.

[1] We refer to the standard ADMM (see, e.g., [10]) and not the parallel ADMM variant or multi-block ADMM, which can have convergence guarantees given additional assumptions.

Let us emphasize some key aspects of this work in detail. First, our characterization is radically different from existing results such as in [5,13,17,22,23,41,47] 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 cannot be improved. Third, our augmented Lagrangian algorithm generates simultaneously both the primal-dual sequence compared to existing augmented Lagrangian algorithms, while maintains its $\mathcal{O}\left(\frac{1}{k^2}\right)$-optimal convergence rate both on the objective residual and on the feasibility gap. Forth, we also describe how to adapt known structures on the objective and constraint components, such as strong convexity, Lipschitz gradient objectives and component full-column ranks. Fifth, this work significantly expands on our earlier conference work [48] not only with new methods but also by demonstrating the impact of warm-start. Finally, our follow up work [50] also demonstrates how our analysis framework and uncertainty principles extend to cover alternating direction optimization methods.
Remark 1 For the clarity, we focus on (1) when $\mathcal{K} = 0$ until Section 7. We also keep only the short proofs in the main text and detail the rest to the appendix.

1.5 Paper organization

Next section recalls preliminary concepts for convex analysis, and introduce a mixed-variational inequality formulation of (1). In Section 3, we propose a smoothing technique with proximity functions for (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 4 presents the first primal-dual algorithmic framework using accelerated (proximal-) gradient schemes for solving (1) and its convergence theory. Sections 5 and 6 provides the second primal-dual algorithmic framework using averaging sequences for solving (1) and its convergence theory. Section 7 specifies different instances of our algorithmic framework for (1) under other common optimization structures and removes the assumption $\mathcal{K} = 0$.

2 Preliminaries

This section recalls some basic notation, the primal-dual formulation for (1), and a variational inequality characterization of the optimality condition of (1).

2.1 Notation

Given a proper, closed, and convex function f, we use $\mathrm{dom}\,(f)$ and $\partial f(x)$ to denote its domain and its subdifferential at x. If f is differentiable, then we use $\nabla f(x)$ for its gradient at x. We call the function f smooth if its gradient ∇f exists at any point in $\mathrm{dom}\,(f)$ and ∇f is continuous in $\mathrm{dom}\,(f)$.

We denote by $f^*(s) := \sup\{\langle s, x\rangle - f(x) : x \in \mathrm{dom}\,(f)\}$, the Fenchel conjugate of f. For any $x \in \mathbb{R}^n$, we define $\|x\|$ as the norm of x, and $\|s\|_* := \max\{\langle s, x\rangle : \|x\| \leq 1\}$ as the dual norm of s. For a given set $\mathcal{X}$, $\delta_{\mathcal{X}}(x) := 0$ if $x \in \mathcal{X}$ and $\delta_{\mathcal{X}}(x) := +\infty$, otherwise, denotes the indicator function of $\mathcal{X}$. We use $\|x\|_2$ for the Euclidean norm. For simplicity of our presentation, we directly work with the Euclidean norm or the weighted Euclidean norm throughout.

For a smooth function f, we say that f is L_f-Lipschitz gradient if for any $x, \tilde{x} \in \mathrm{dom}\,(f)$, we have $\|\nabla f(x) - \nabla f(\tilde{x})\|_* \leq L_f\|x - \tilde{x}\|$, where $L(f) := L_f \in [0, \infty)$. We denote by $\mathcal{F}_L^{1,1}$ 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.

2.2 Optimality condition

Primal-dual formulation: Let us define the Lagrange function associated with the linear constraint $Ax - b = 0$ as $\mathcal{L}(x, y) := f(x) + \langle y, Ax - b\rangle$. The min-max problem associated with the Lagrange function $\mathcal{L}$ is defined as follows:

$$g^\star = \max_{y\in\mathbb{R}^m} g(y) = \max_{y\in\mathbb{R}^m} \min_{x\in\mathcal{X}} \mathcal{L}(x, y) \le \min_{x\in\mathcal{X}} \bar{f}(x) = f^\star, \tag{9}$$

where g is the dual function and $\bar{f}$ is the extension of f over $\mathbb{R}^n$

$$g(y) := \inf\{\mathcal{L}(x, y) := f(x) + \langle Ax - b, y\rangle : x \in \mathcal{X}\}, \tag{10}$$

$$\bar{f}(x) := \sup_{y\in\mathbb{R}^m} \mathcal{L}(x, y) = \begin{cases} f(x) & \text{if } Ax = b; \\ +\infty & \text{otherwise.} \end{cases} \tag{11}$$

Let us denote by $\mathcal{X}^\star$ the optimal solution set of (1). If $\mathcal{X}^\star$ is nonempty, then the optimal value $f^\star$ of (7) is finite. We define the Lagrange dual problem of (1) as

$$g^\star := \max\left\{g(y) : y \in \mathbb{R}^m\right\}. \tag{12}$$

In this case, we refer to (1) and (12) as the primal-dual problems. We note that g is proper if $\{x \in \mathcal{X} : Ax = b\} \cap \mathrm{dom}\,(f) \neq \emptyset$. We denote by $\mathrm{dom}\,(g)$ its domain.

Under the decomposability structure (7), we can write the dual function as

$$g(y) := \sum_{i=1}^N g^i(y), \quad \text{with} \quad g^i(y) := \inf\{f_i(x_i) + \langle A_i x_i - b_i, y\rangle : x_i \in \mathcal{X}_i\}, \tag{13}$$

which implies that we can compute g *in parallel*.

Optimality condition: We can write the optimality condition or Karush-Kuhn-Tucker (KKT) condition of (1) and (12) as follows:

$$\begin{cases} 0 \in \partial f(x^\star) + A^T y^\star + \mathcal{N}_{\mathcal{X}}(x^\star) \equiv \partial_x \mathcal{L}(x^\star, y^\star) + \mathcal{N}_{\mathcal{X}}(x^\star) \\ 0 = Ax^\star - b \qquad\qquad\qquad\qquad\quad \equiv \partial_y \mathcal{L}(x^\star, y^\star), \end{cases} \tag{14}$$

where $\mathcal{N}_{\mathcal{X}}(x^\star)$ is the normal cone of $\mathcal{X}$ at $x^\star$. Any point $w^\star := (x^\star, y^\star)$ satisfying (14) is called a KKT point of (1). We denote by $\mathcal{W}^\star$ the set of KKT points. Then $\mathcal{W}^\star = \mathcal{X}^\star \times \mathcal{Y}^\star$, where $\mathcal{X}^\star$ is the set of stationary points $x^\star$, and $\mathcal{Y}^\star$ is the set of corresponding multipliers $y^\star$.

Fundamental assumptions: In order to relate the primal problem (1) and its dual (12), we need to introduce a few assumptions, which are standard in the literature.

Assumption A. 1 *The constraint domain $\mathcal{X}$ and the solution set $\mathcal{X}^\star$ of (1) are nonempty. The function f is proper, closed, and convex. In addition, either $\mathcal{X}$ is a polytope or the following Slater condition holds*

$$\left\{x \in \mathbb{R}^n : Ax - b = 0\right\} \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset, \tag{15}$$

where $\mathrm{ri}(\mathcal{X})$ *is the relative interior of* $\mathcal{X}$.

Under Assumption A.1, the dual problem (12) is feasible. Moreover, its solution set $\mathcal{Y}^\star$ is nonempty and bounded. The KKT condition (14) is necessary and sufficient for $w^\star = (x^\star, y^\star)$ to be an optimal solution of (1) and (12).

Remark 2 Throughout this paper, we assume that Assumption A.1 holds.

Approximate solutions: For any optimal solution $x^\star \in \mathcal{X}^\star$ of (1), we have $f(x^\star) - f^\star = 0$, $x^\star \in \mathcal{X}$ and $Ax^\star - b \in \mathcal{K}$. Our goal in this paper is to design primal-dual algorithms that produce an approximation $x^\star_\varepsilon$ to $x^\star \in \mathcal{X}^\star$ in the following sense:

Definition 1 Given a target accuracy $\varepsilon \geq 0$, a point $x^\star_\varepsilon \in \mathcal{X}$ is said to be an ε-solution of (1) if $f(x^\star_\varepsilon) - f^\star \leq \varepsilon$ and $\mathrm{dist}(Ax^\star_\varepsilon - b, \mathcal{K}) \leq \varepsilon$.

For clarity and without loss of generality, we will work with $\mathcal{K} = 0$ in the sequel and then illustrate the algorithmic changes for the general case. As a result, the approximate feasibility takes the form $\|Ax^\star_\varepsilon - b\| \leq \varepsilon$. Moreover, we assume in Definition 1 that $x^\star_\varepsilon \in \mathcal{X}$, i.e., $x^\star_\varepsilon$ is exactly feasible to $\mathcal{X}$. This requirement is reasonable in practice since $\mathcal{X}$ is usually "simple," where the projection onto $\mathcal{X}$ can be computed efficiently, e.g., when $\mathcal{X}$ is a box, a simplex or a cone constraint.

Note that we can also use different accuracy levels for the absolute primal objective residual $f(x^\star_\varepsilon) - f^\star \leq \varepsilon_1$ and the primal feasibility gap $\|Ax^\star_\varepsilon - b\| \leq \varepsilon_2$. Note also that $f(x) - f^\star \geq -\|y^\star\|_*\|Ax - b\|$ for any $y^\star \in \mathcal{Y}^\star$ and $x \in \mathcal{X}$, which guarantees the lower bound of the objective residual $f(x) - f^\star$ (*cf.*, Lemma 3) .

2.3 Mixed-variational inequality formulation and its gap function

Mixed-variational inequality: Let $w := (x, y) \equiv (x^T, y^T)^T \in \mathbb{R}^n \times \mathbb{R}^m$ be the primal-dual variable, $\mathcal{W} := \mathcal{X} \times \mathbb{R}^m$ be the primal-dual domain, and $F(w) := [A^T y, b - Ax] \equiv \left((A^T y)^T, (b - Ax)^T\right)^T$ be the partial KKT mapping. Then, (14) can be reformulated into the following *mixed-variational inequality* (MVIP) [18]:

$$f(x) - f(x^\star) + \left\langle F(w^\star), w - w^\star \right\rangle \geq 0, \quad \forall w \in \mathcal{W}. \tag{16}$$

Finding a point $w^\star \in \mathcal{W}$ such that (16) holds is equivalent to solving the primal-dual problems (1)-(12).

Gap function: If we define the bifunction $\mathcal{B}(w, \tilde{w}) := f(x) - f(\tilde{x}) + \langle F(w), w - \tilde{w}\rangle = f(x) - f(\tilde{x}) - \langle A\tilde{x} - b, y\rangle + \langle Ax - b, \tilde{y}\rangle$, then $\mathcal{B}(w, w) = 0$ for all $w \in \mathcal{W}$. Let

$$G(w) := \max\left\{\mathcal{B}(w, \tilde{w}) : \tilde{w} \in \mathcal{W}\right\} \equiv \max_{\tilde{y} \in \mathbb{R}^m} \mathcal{L}(x, \tilde{y}) - \min_{\tilde{x} \in \mathcal{X}} \mathcal{L}(\tilde{x}, y) \equiv \bar{f}(x) - g(y), \tag{17}$$

be the Auslender gap function of (16) [1]. Then, the following result is standard in convex optimization due to the weak and the strong duality theory.

Lemma 1 *The gap function G defined by* (17) *is nonnegative on $\mathcal{W}$, i.e., $G(w) \geq 0$ for all $w \in \mathcal{W}$. Moreover, $G(w^\star) = 0$ if and only if $w^\star = (x^\star, y^\star) \in \mathcal{W}^\star$ is a primal-dual solution of* (1) *and* (12).

Clearly, the non-negativity of G is due to the weak duality theorem, and the second condition holds due to the strong duality theorem. In general, the gap function G is nonconvex and nonsmooth [43]. Fortunately, in the setting (1), G is convex, but is possibly nonsmooth and may take infinite values.

3 Smoothing the gap function via proximity functions

In this section, we provide a smooth approximation of the gap function G and prove key properties of this approximation.

3.1 Proximity functions and Bregman distances

Proximity functions: Given a nonempty, closed and convex set $\mathcal{Z}$, a continuous, and strongly convex function p with the parameter $\mu_p > 0$ is called a *proximity* function (or prox-function) of $\mathcal{Z}$ if $\mathcal{Z} \subseteq \mathrm{dom}\,(p)$. We also denote

$$\bar{z}^c := \mathrm{argmin}\,\{p(z) : z \in \mathrm{dom}\,(p)\}, \tag{18}$$

as the prox-center of p. Without loss of generality, we assume that $\mu_p = 1$ and $p(\bar{z}^c) = 0$. For example, $p_{\mathcal{Z}}(z) := (1/2)\|z\|_2^2$ is a simple prox-function in $\mathbb{R}^{n_z}$.

Given a prox-function p, if $p \in \mathcal{F}_L^{1,1}$, then its conjugate p^* is strongly convex with the convexity parameter $\mu_{p^*} = L_p^{-1}$ due to the Baillon-Haddad's theorem [2]. We denote by $\bar{s}^c$ the prox-center of p^*. It is clear that $p^*(\bar{s}^c) = -p(0)$. Moreover, $p^* \in \mathcal{F}_L^{1,1}$ with the Lipschitz constant $L_{p^*} = 1$.

Bregman distances: Given a smooth prox-function p defined on $\mathcal{Z}$ with the convexity parameter $\mu_p = 1$ and $p(\bar{z}^c) = 0$, we define the following Bregman distance

$$d_p(u, v) := p(u) - p(v) - \langle \nabla p(v), u - v\rangle, \quad \forall u, v \in \mathcal{Z}. \tag{19}$$

Clearly, $d_p(u, \bar{z}^c) = p(u)$, and $d_p(u, v) \geq \frac{1}{2}\|u - v\|^2$ for any $u, v \in \mathcal{Z}$. In addition, if p is Lipschitz continuous with the Lipschitz constant $L_p \geq 1$, then $d_p(u, v) \leq \frac{L_p}{2}\|u - v\|^2$ for any $u, v \in \mathbb{R}^q$. If $p(u) := \frac{1}{2}\|u\|_2^2$, then d_p becomes the standard Euclidean distance.

3.2 Smoothed primal-dual gap function

The gap function G defined in (17) is *convex* but generally *nonsmooth*. This subsection introduces a smoothed primal-dual gap function that approximates G. For this purpose, we first discuss smoothing strategies.

Smoothing functions: Let $\mathcal{I}_N := \{1, \cdots, N\}$ be the index set of components corresponding to the structure (7). We first decompose $\mathcal{I}_N$ into two subsets

$$\mathcal{I}_1 \subseteq \mathcal{I}_A := \big\{i \in \mathcal{I}_N : \lambda_{\min}(A_i^T A_i) \geq \sigma_i^2 > 0\big\}, \quad \text{and} \quad \bar{\mathcal{I}}_1 := \mathcal{I}_N \backslash \mathcal{I}_1, \tag{20}$$

where $\lambda_{\min}(\cdot)$ is the smallest eigenvalue. We allow $\mathcal{I}_1$ to be empty.

For each $i \in \mathcal{I}_N$, we choose a prox-function $p_i \in \mathcal{F}_L^{1,1}$, which is 1-strongly convex and its gradient is Lipschitz continuous with $L_{p_i} > 0$ on its corresponding domain. We define the prox-function of $\mathcal{X}$ and the constant $\bar{L}_i$, respectively as

$$p_x(Ax) := \sum_{i\in\mathcal{I}_1} p_i(x_i) + \sum_{i\in\bar{\mathcal{I}}_1} p_i(A_i x_i) \quad \text{and} \quad \bar{L}_i := \begin{cases} \bar{L}_{A_i} & i \in \mathcal{I}_1 \\ 1 & i \in \bar{\mathcal{I}}_1, \end{cases} \tag{21}$$

where $\bar{L}_{A_i} := \lambda_{\max}(A_i^T A_i)$ is the largest eigenvalue of $A_i^T A_i$. Clearly, we have

$$\begin{aligned} \frac{1}{2}\sum_{i\in\mathcal{I}_1} \|x_i - \bar{x}_i^c\|^2 + \frac{1}{2}\sum_{i\in\bar{\mathcal{I}}_1} \|A_i x_i - \bar{u}_i^c\|^2 &\leq p_x(Ax) \\ &\leq \frac{1}{2}\sum_{i\in\mathcal{I}_1} L_{p_i}\|x_i - \bar{x}_i^c\|^2 + \frac{1}{2}\sum_{i\in\bar{\mathcal{I}}_1} L_{p_i}\|A_i x_i - \bar{u}_i^c\|^2, \end{aligned}$$

where $\bar{x}_i^c$ is the prox-center of p_i for $i \in \mathcal{I}_1$ and $\bar{u}_i^c$ is the prox-center of p_i for $i \in \bar{\mathcal{I}}_1$. Here, we implicitly use the fact that $\nabla p_i(\bar{z}_i^c) = 0$, where $\bar{z}_i^c$ is either $\bar{x}_i^c$ or $\bar{u}_i^c$. Otherwise, we can replace p_i by the corresponding Bregman distance $d_{p_i}(\cdot, \bar{z}_i^c)$. We also choose $p_y \in \mathcal{F}_L^{1,1}$ a prox-function defined on $\mathbb{R}^m$ for the dual problem. For given two positive smoothness parameters γ and β, we consider the function

$$p_{\gamma\beta}(w) := \gamma p_x(Ax) + \beta p_y(y). \tag{22}$$

We call $p_{\gamma\beta}$ a *smoother*, or a *regularizer*, for the gap function G.

A smoothed approximation of the primal-dual gap function: For a given $p_{\gamma\beta}$ as defined by (22), we consider an approximation of f and g, respectively as follows:

$$\begin{aligned} g_\gamma(\bar{y}) &:= \min_{x\in\mathcal{X}}\left\{f(x) + \langle Ax - b, \bar{y}\rangle + \gamma p_x(Ax)\right\},\\ f_\beta(\bar{x}) &:= \max_{y\in\mathbb{R}^m}\left\{f(\bar{x}) + \langle A\bar{x} - b, y\rangle - \beta p_y(y)\right\}. \end{aligned} \tag{23}$$

The smoothed primal-dual gap (or the smoothed gap) $G_{\gamma\beta}$ is then defined as

$$G_{\gamma\beta}(\bar{w}) := f_\beta(\bar{x}) - g_\gamma(\bar{y}). \tag{24}$$

Loosely speaking, $G_{\gamma\beta}$ approaches G as γ and β approach to zero.

Evaluating the smoothed gap function: To evaluate f_β and g_γ, we need to solve the following two convex subproblems w.r.t. $\bar{x}$ and $\bar{y}$, respectively

$$\begin{cases} x^*_\gamma(\bar{y}) \in \arg\min_{x\in\mathcal{X}}\left\{f(x) + \langle \bar{y}, Ax - b\rangle + \gamma p_x(Ax)\right\},\\ y^*_\beta(\bar{x}) := \arg\max_{y\in\mathbb{R}^m}\left\{\langle A\bar{x} - b, y\rangle - \beta p_y(y)\right\} = \nabla p^*_y\big(\beta^{-1}(A\bar{x} - b)\big). \end{cases} \tag{25}$$

We denote by $w^*_{\gamma\beta}(\bar{w}) := (x^*_\gamma(\bar{y}), y^*_\beta(\bar{x})) \in \mathcal{W}$. While $y^*_\beta(\bar{x})$ can be computed explicitly as in (25), the computation of $x^*_\gamma(\bar{y})$ can be split into N subproblems due to the decomposability of f and $\mathcal{X}$ in (7), i.e.,

$$x^*_{\gamma,i}(\bar{y}) \in \begin{cases} \arg\min_{x_i\in\mathcal{X}_i}\left\{f_i(x_i) + \langle \bar{y}, A_ix_i - b_i\rangle + \gamma p_i(x_i)\right\}, & i \in \mathcal{I}_1,\\ \arg\min_{x_i\in\mathcal{X}_i}\left\{f_i(x_i) + \langle \bar{y}, A_ix_i - b_i\rangle + \gamma p_i(A_ix_i)\right\}, & i \in \bar{\mathcal{I}}_1. \end{cases} \tag{26}$$

As a result, g_γ becomes $g_\gamma(\bar{y}) := \sum_{i=1}^N g^i_\gamma(\bar{y})$ with

$$g^i_\gamma(\bar{y}) := \begin{cases} \min_{x_i\in\mathcal{X}_i}\left\{f_i(x_i) + \langle \bar{y}, A_ix_i - b_i\rangle + \gamma p_i(x_i)\right\}, & i \in \mathcal{I}_1,\\ \min_{x_i\in\mathcal{X}_i}\left\{f_i(x_i) + \langle \bar{y}, A_ix_i - b_i\rangle + \gamma p_i(A_ix_i)\right\}, & i \in \bar{\mathcal{I}}_1. \end{cases} \tag{27}$$

Alternatively, the function f_β defined by (23) can be computed explicitly as

$$f_\beta(\bar{x}) = f(\bar{x}) + \beta p^*_y\big(\beta^{-1}(A\bar{x} - b)\big) := f(\bar{x}) + \bar{p}_\beta(\bar{x}). \tag{28}$$

In practice, we prefer to choose p_y such that the computation of $y^*_\beta(\bar{x})$ in (25) is cheap. For example, if we select $p_y(y) := (1/2)\|y\|_2^2$, then $p_y \in \mathcal{F}_L^{1,1}$ with $L_{p_y} = 1$, and $p^*_y(v) = (1/2)\|v\|^2_* = (1/2)\|v\|_2^2$. Moreover, $y^*_\beta(\bar{x})$ computed by (25) reduces to $y^*_\beta(\bar{x}) := \beta^{-1}(A\bar{x} - b)$.

The diameter of domain: We define the following diameter of the domain $\mathcal{X}$

$$D_{\mathcal{X}} := \sum_{i=1}^N D_{\mathcal{X}_i}, \text{ where } D_{\mathcal{X}_i} := \begin{cases} \sup\left\{p_i(x_i) : x_i \in \mathcal{X}_i \cap \mathrm{dom}\,(f_i)\right\} & i \in \mathcal{I}_1\\ \sup\left\{p_i(A_ix_i) : x_i \in \mathcal{X}_i \cap \mathrm{dom}\,(f_i)\right\} & i \in \bar{\mathcal{I}}_1. \end{cases} \tag{29}$$

For designing algorithms, we summarize our technical assumptions as follows:

Assumption A. 2 *For each $i \in \mathcal{I}_N$, the prox-function p_i is 1-strongly convex and smooth, and its gradient is Lipschitz continuous with the Lipschitz constant $L_{p_i} \geq 1$. The prox-function p_y is also 1-strongly convex and smooth, and its gradient is Lipschitz continuous with the Lipschitz constant $L_{p_y} \geq 1$. For each $i \in \mathcal{I}_N$, $D_{\mathcal{X}_i}$ defined by (29) is bounded, i.e., $D_{\mathcal{X}_i} \in [0, +\infty)$.*

Our default choice of p_i as well as p_y is the quadratic prox-function $p(\cdot) := \frac{1}{2}\|\cdot\|^2$, with $L_p = 1$. Clearly, if $\mathcal{X}_i$ or $\mathrm{dom}\,(f_i)$ is bounded, then $D_{\mathcal{X}_i}$ is also bounded. Hence, if $D_{\mathcal{X}_i}$ is bounded for all $i \in \mathcal{I}_N$, then $D_{\mathcal{X}}$ is also bounded. We also assume that Assumption A.2 holds in the sequel.

Key properties of f_β and g_γ: The smoothed gap components g^i_γ defined by (27), and f_β defined by (28) satisfy the following properties (*cf.*, Appendix A.1.2).

Lemma 2 *For $i \in \mathcal{I}_N$, the function $g^i_\gamma(\cdot)$ defined by (26) is concave and smooth on $\mathbb{R}^m$. Its gradient is given by $\nabla g^i_\gamma(\cdot) := A_i x^*_{\gamma,i}(\cdot) - b_i$, which is Lipschitz continuous with the Lipschitz constant $L_{g^i_\gamma} := \gamma^{-1}\bar{L}_i$, where $\bar{L}_i$ is defined by (21). Consequently, for all $y, \bar{y} \in \mathbb{R}^m$, the following estimates hold*

$$\begin{aligned} 0 &\le g^i_\gamma(\bar{y}) + \langle \nabla g^i_\gamma(\bar{y}), y - \bar{y}\rangle - g^i_\gamma(y) \le \tfrac{L_{g^i_\gamma}}{2}\|y - \bar{y}\|^2,\\ g^i_\gamma(y) &\le g^i_\gamma(\bar{y}) + \langle \nabla g^i_\gamma(\bar{y}), y - \bar{y}\rangle - \tfrac{1}{2L_{g^i_\gamma}}\|\nabla g^i_\gamma(y) - \nabla g^i_\gamma(\bar{y})\|^2_*. \end{aligned} \tag{30}$$

For $\bar{y} \in \mathbb{R}^m$ and $\gamma > 0$ and g^i defined by (13), g^i_γ satisfies the following estimate

$$g^i_\gamma(\bar{y}) - \gamma D_{\mathcal{X}_i} \le g^i(\bar{y}) \le g^i_\gamma(\bar{y}). \tag{31}$$

For a fixed $\bar{y} \in \mathbb{R}^m$, the function $g^i_\gamma(\bar{y})$ is nondecreasing, concave and differentiable in $\mathbb{R}_{++}$ w.r.t. γ. Moreover, for γ and $\bar{\gamma}$ in $\mathbb{R}_{++}$, we have

$$g^i_\gamma(\bar{y}) \le g^i_{\bar{\gamma}}(\bar{y}) + (\gamma - \bar{\gamma})\bar{p}^*_i, \quad \text{where} \quad \bar{p}^*_i := \begin{cases} p_i(x^*_{\bar{\gamma},i}(\bar{y})) & \text{if } i \in \mathcal{I}_1,\\ p_i(A_i x^*_{\bar{\gamma},i}(\bar{y})) & \text{if } i \notin \mathcal{I}_1, \end{cases} \tag{32}$$

and $x^*_{\bar{\gamma},i}(\bar{y})$ is defined by (26).

Consequently, $g_\gamma := \sum_{i=1}^N g^i_\gamma$ is also concave and smooth. Its gradient $\nabla g_\gamma(y) = Ax^*_\gamma(y) - b$ is Lipschitz continuous with the Lipschitz constant $L_{g_\gamma} := \gamma^{-1}\bar{L}_g$, where $\bar{L}_g := \sum_{i=1}^N \bar{L}_i$. Moreover, the estimates (30), (31) and (32) also hold for g_γ.

Alternatively, let f_β and $\bar{p}_\beta$ be defined by (28). Then, $\bar{p}_\beta$ is convex and smooth, its gradient is Lipschitz continuous with the Lipschitz constant $L_{\bar{p}_\beta} := \frac{\|A\|^2}{\beta}$. Moreover, we have

$$\begin{aligned} f_\beta(x) &\ge f_{\bar{\beta}}(x) + (\bar{\beta} - \beta)p_y(y^*_{\bar{\beta}}(x)),\\ \bar{p}_\beta(x) &\ge \bar{p}_\beta(\hat{x}) + \langle \nabla \bar{p}_\beta(\hat{x}), x - \hat{x}\rangle + \tfrac{1}{2L_{p_y}\beta}\|A(x - \hat{x})\|^2, \end{aligned} \tag{33}$$

for $\beta, \bar{\beta} \in \mathbb{R}_{++}$ and $x, \hat{x} \in \mathcal{X}$.

Since g_γ defined by (23) is concave and smooth, and its gradient is Lipschitz continuous, we can in principle apply the accelerated gradient scheme in [33] to solve the following smoothed dual problem

$$g^\star_\gamma := \max\left\{g_\gamma(y) : y \in \mathbb{R}^m\right\}. \tag{34}$$

Then, we can obtain the $\mathcal{O}\big(\frac{1}{\varepsilon}\big)$-worst-case complexity in terms of the dual objective residual $g(\bar{y}^k) - g^\star$ as in [35]. We can also use an averaging scheme to recover the primal solution as in [29,57]. However, as a disadvantage, such schemes fix a priori the smoothness parameter γ at $\gamma := \mathcal{O}\Big(\frac{\varepsilon}{D_{\mathcal{U}}}\Big)$, which is too restrictive.

3.3 An estimate for the objective residual and primal feasibility gap

The following lemma provides a fundamental estimate on the bounds of $f(\bar{x}^k) - f^\star$ and $\|A\bar{x}^k - b\|$. For clarity of exposition, we move this proof to Appendix A.1.3.

Lemma 3 *Let g_γ and f_β and $G_{\gamma\beta}$ defined by* (23) *and* (24)*, respectively. Then, for any $y^\star \in \mathcal{Y}^\star$ and $x \in \mathcal{X}$, one has*

$$-\|y^\star\|_*\|Ax - b\| \le f(x) - f^\star \le f(x) - g(y). \tag{35}$$

Let $\{\bar{w}^k\}$ be a primal-dual sequence in $\mathcal{W}$, and $\{(\gamma_k, \beta_k)\}$ be a smoothness parameter sequence in $\mathbb{R}^2_{++}$. Then, we have

$$\begin{cases} f(\bar{x}^k) - g(\bar{y}^k) \le S_k := G_{\gamma_k\beta_k}(\bar{w}^k) + \gamma_k D_{\mathcal{X}} + \beta_k p_y(0), \\ \|A\bar{x}^k - b\| \le \beta_k\Big[\bar{c}_\star + \sqrt{\bar{c}_\star^2 + \big(2L_{p_y}\beta_k^{-1}S_k - \|\bar{s}_c\|^2\big)}\Big], \end{cases} \tag{36}$$

where $\bar{c}_\star := \|L_{p_y}y^\star - \bar{s}_c\|_*$, provided that $\bar{c}_\star + 2L_{p_y}\beta_k^{-1}S_k - \|\bar{s}_c\|^2 \ge 0$.

In particular, if we choose $p_y(y) := \frac{1}{2}\|y\|_2^2$, then

$$\begin{cases} f(\bar{x}^k) - g(\bar{y}^k) \le \bar{G}_k + \gamma_k D_{\mathcal{X}} \\ \|A\bar{x}^k - b\|_2 \le 2\beta_k D_{\mathcal{Y}^\star} + \sqrt{2\beta_k\big(\bar{G}_k + \gamma_k D_{\mathcal{X}}\big)}, \end{cases} \tag{37}$$

where $\bar{G}_k := G_{\gamma_k\beta_k}(\bar{w}^k)$ and $D_{\mathcal{Y}^\star} := \min\{\|y^\star\|_2 : y^\star \in \mathcal{Y}^\star\}$ is the norm of minimum norm dual solutions.

The estimates (35), (36) and (37) are independent of optimization methods using to construct $\{\bar{w}^k\}$. However, their convergence guarantee depends on the smoothness parameters γ_k and β_k. Hence, the convergence rate of the objective residual $f(\bar{x}^k) - f^\star$ and feasibility gap $\|A\bar{x}^k - b\|$ depends on the rate of $\{(\gamma_k, \beta_k)\}$.

3.4 Descent models for the smoothed gap function

Our goal is to generate a primal-dual sequence $\{\bar{w}^k\} \subseteq \mathcal{W}$ and a smoothness parameter sequence $\{(\gamma_k, \beta_k)\} \subseteq \mathbb{R}^2_{++}$ so that $\{G_{\gamma_k\beta_k}(\bar{w}^k)\}$ converges to 0^+, where $G_{\gamma_k\beta_k}(\cdot)$ is defined by (24). To achieve this goal, we propose to use the following model imposing on $G_{\gamma\beta}$ in order to design algorithms in the next sections:

Definition 2 The primal-dual sequence $\{\bar{w}^k\} \subseteq \mathcal{W}$ and the smoothness parameter sequence $\{(\gamma_k, \beta_k)\} \subseteq \mathbb{R}^2_{++}$ are said to satisfy the *model-based gap reduction* (MGR) *condition* on $G_{\gamma\beta}$ if the following inequality is satisfied

$$G_{\gamma_{k+1}\beta_{k+1}}(\bar{w}^{k+1}) \le (1 - \tau_k)G_{\gamma_k\beta_k}(\bar{w}^k) + \psi_k, \tag{38}$$

where $\tau_k \in (0, 1)$, $\sum_{i=0}^{\infty}\tau_k = \infty$, and $\lim_{k\to\infty}\psi_k = 0$.

In this definition, we have not specified convergence properties of the parameter sequences $\{\tau_k\}$ and $\{\psi_k\}$. However, as mentioned previously, we can obtain a monotone or a non-monotone model by appropriately choosing the augmented term ψ_k. If we chose $\psi_k \le 0$, then we obtain a monotone model, while if $\psi_k > 0$, we deal with a nonmonotone model.

With the monotone model, i.e., $\psi_k \le 0$, by induction, we can derive from (38) that $G_{\gamma_k\beta_k}(\bar{w}^k) \le \omega_k G_{\gamma_0\beta_0}(\bar{w}^0)$, where $\omega_k := \prod_{i=0}^{k-1}(1-\tau_i)$. Hence, the convergence rate of $\{G_{\gamma_k\beta_k}(\bar{w}^k)\}$ is upper bounded by the convergence rate of $\{\omega_k\}$. If we

update the parameters γ_k and β_k in an alternating fashion, then the monotone model covers Nesterov's excessive gap technique in [34] as a special case.

With the non-monotone model, one can upper bound ψ_k by $\psi_k \le (1-\tau_k)R_k - R_{k+1}$ for a given sequence $\{R_k\}$. Then, we can write this model as $G_{\gamma_{k+1}\beta_{k+1}}(\bar{w}^{k+1}) - R_{k+1} \le (1-\tau_k)\big[G_{\gamma_k\beta_k}(\bar{w}^k) - R_k\big]$. Hence, $G_{\gamma_k\beta_k}(\bar{w}^k) \le R_k + \omega_k\big[G_{\gamma_0\beta_0}(\bar{w}^0) - R_0\big]$. Clearly, the convergence rate of $\{G_{\gamma_k\beta_k}(\bar{w}^k)\}$ is upper bounded by the convergence rate of the two sequences $\{\omega_k\}$ and $\{R_k\}$. In particular, if $G_{\gamma_0\beta_0}(\bar{w}^0) \le R_0$, then $G_{\gamma_k\beta_k}(\bar{w}^k) \le R_k$, which shows that $G_{\gamma_k\beta_k}(\bar{w}^k)$ is upper bounded by R_k.

Several primal-dual methods can be developed to maintain (38). In the next sections, we demonstrate three primal-dual schemes based on our MGR technique.

4 The accelerated primal-dual gap reduction algorithm

Our goal is to design a new scheme for updating the primal-dual sequence $\{\bar{w}^k\}$ and the parameter sequence $\{(\gamma_k, \beta_k)\}$ that maintain the MGR condition (38).

4.1 The accelerated smoothed-gap reduction scheme

Our new scheme builds upon Nesterov's acceleration idea [32,33]. At each iteration, we apply an accelerated (proximal-)gradient step to minimize f_β, while it maximizes g_γ. Since $f_\beta(\cdot) = f(\cdot) + \beta p_y^*(\beta^{-1}(A\cdot - b))$ is nonsmooth, we use the proximal-operator of $f_{\mathcal{X}} := f + \delta_{\mathcal{X}}$ to generate a proximal-gradient step. Alternatively, since g_γ is smooth and has Lipschitz gradient, we can use its gradient. As a key feature, we must update the parameters γ_k and β_k simultaneously at each iteration, which is different from existing methods.

Let $\bar{w}^k := (\bar{x}^k, \bar{y}^k) \in \mathcal{W}$ and $\tilde{w}^k := (\tilde{x}^k, \tilde{y}^k) \in \mathcal{W}$ be given. The *Accelerated Smoothed GAp ReDuction* (ASGARD) scheme generates a new primal-dual point $\bar{w}^{k+1} := (\bar{x}^{k+1}, \bar{y}^{k+1})$ as

$$\begin{cases} \hat{w}^k &:= (1-\tau_k)\bar{w}^k + \tau_k\tilde{w}^k, \\ \bar{x}^{k+1} &:= \operatorname{prox}_{\rho_{k+1}f_{\mathcal{X}}}\left(\hat{x}^k - \rho_{k+1}A^T y^*_{\beta_{k+1}}(\hat{x}^k)\right), \\ \bar{y}^{k+1} &:= \hat{y}^k + \lambda_{k+1}\big(Ax^*_{\gamma_{k+1}}(\hat{y}^k) - b\big), \\ \tilde{x}^{k+1} &:= \tilde{x}^k - \tau_k^{-1}\left(2 - \rho_{k+1}\beta_{k+1}^{-1}\bar{L}_A\right)\left(\hat{x}^k - \bar{x}^{k+1}\right), \\ \tilde{y}^{k+1} &:= \tilde{y}^k - \tau_k^{-1}\left(2 - \lambda_{k+1}\gamma_{k+1}^{-1}\bar{L}_g\right)\left(\hat{y}^k - \bar{y}^{k+1}\right), \end{cases} \tag{ASGARD}$$

where $\tau_k \in (0,1]$, $\rho_{k+1} > 0$ and $\lambda_{k+1} > 0$ are two step-sizes which will be determined in the sequel such that $\rho_{k+1}\beta_{k+1}^{-1}\bar{L}_A \in (0,2)$ and $\lambda_{k+1}\gamma_{k+1}^{-1}\bar{L}_g \in (0,2)$, respectively. The constants $\bar{L}_g := \sum_{i=1}^N \bar{L}_i$ and $\bar{L}_A := \|A\|^2$ are defined as before.

The ASGARD scheme requires one solution $x^*_{\gamma_{k+1}}(\hat{y}^k)$ of the primal subproblem in (25), and one dual solution $y^*_{\beta_{k+1}}(\bar{x}^k)$ at the second line of (25). In addition, it requires a proximal step of $f_{\mathcal{X}}$. Computing $x^*_{\gamma_{k+1}}(\hat{y}^k)$ as well as this proximal step can be implemented *in parallel* when the decomposition structure (7) is available.

The following lemma shows that $\bar{w}^{k+1}$ updated by (ASGARD) maintains the MGR condition (38), whose proof can be found in Appendix A.2.1.

Lemma 4 *Let $\bar{w}^{k+1} := (\bar{x}^{k+1}, \bar{y}^{k+1})$ be updated by* (ASGARD). *Let $\hat{c}$ and $\bar{L}_g$ be the two constants defined by*

$$\hat{c}_1 := \max\left\{\max_{i\in\mathcal{I}_1}\left\{\frac{L_{p_i}\bar{L}_{A_i}}{\sigma_i^2}\right\}, \max_{i\notin\mathcal{I}_1}\{L_{p_i}\}\right\}, \text{ and } \bar{L}_g := \sum_{i=1}^{N}\bar{L}_i, \tag{39}$$

where σ_i^2 is given in (20), *and $\bar{L}_{A_i}$ and $\bar{L}_i$ are defined by* (21). *Then, if $\tau_k \in (0,1]$, $\beta_k > 0$ and $\gamma_k > 0$ are chosen such that*

$$\left(1 + \frac{\tau_k}{L_{p_y}^2}\right)\beta_{k+1} \geq \beta_k \quad \text{and} \quad \left(1 + \frac{\tau_k}{\hat{c}_1}\right)\gamma_{k+1} \geq \gamma_k, \tag{40}$$

then $\bar{w}^{k+1} \in \mathcal{W}$ and satisfies the MGR condition (38) *with*

$$\psi_k := \frac{\tau_k^2}{2u_{k+1}}\left\{\left[\|\tilde{y}^k - y^\star\|^2 - \|\tilde{y}^{k+1} - y^\star\|^2\right] + \left[\|\tilde{x}^k - x^\star\|^2 - \|\tilde{x}^{k+1} - x^\star\|^2\right]\right\}, \tag{41}$$

where $u_{k+1} := 2\rho_{k+1} - \rho_{k+1}^2\beta_{k+1}^{-1}\bar{L}_A = 2\lambda_{k+1} - \lambda_{k+1}^2\gamma_{k+1}^{-1}\bar{L}_g > 0$.

In addition, if $(1-\tau_k)\tau_{k-1}^2 u_{k+1} = \tau_k^2 u_k$, then

$$G_k(\bar{w}^k) \leq \frac{\tau_k^2}{(1-\tau_k)u_{k+1}}\left[\frac{(1-\tau_0)u_1}{\tau_0^2}G_0(\bar{w}^k) + \frac{1}{2}\|\tilde{y}^0 - y^\star\|^2 + \frac{1}{2}\|\tilde{x}^0 - x^\star\|^2\right]. \tag{42}$$

4.2 Updating the smoothing and gap reduction parameters

Next, using Lemma 4, we can develop the rules for updating τ_k, β_k, γ_k and u_k so that the conditions in (40) and $(1-\tau_k)\tau_{k-1}^2\gamma_{k+1} = \tau_k^2\gamma_k$ hold. One way of updating these parameters is presented in the following lemma, whose proof can be found in Appendix A.2.2.

Lemma 5 *Suppose that $p_y(\cdot)$ and $p_i(\cdot)$ are chosen such that $L_{p_y} = L_{p_i} = 1$ for $i = 1, \cdots, N$. Suppose further that $\mathcal{I}_N$ is chosen by $\mathcal{I}_N = \emptyset$. Then, $\hat{c}_1$ defined by* (39) *satisfies $\hat{c}_1 = 1$. The parameters τ_k, γ_k, β_k and u_{k+1} updated by*

$$\tau_k := \frac{1}{k+1}, \quad \gamma_k := \frac{2\gamma_1}{k+1}, \quad \beta_k := \frac{2\beta_1}{k+1}, \quad \text{and} \quad u_{k+1} = \frac{u_1}{k+1}, \tag{43}$$

satisfy the conditions (40) *and $(1-\tau_k)\tau_{k-1}^2 u_{k+1} = \tau_k^2 u_k$ in Lemma 4, and $\tau_0 = 1$. The parameters ρ_k and λ_k respectively updated by*

$$\rho_k := \frac{2\beta_1}{\bar{L}_A(k+1)}\left[1 \pm \sqrt{1 - \frac{\bar{L}_A u_1(k+1)}{2\beta_1 k}}\right] \text{ and } \lambda_k := \frac{2\gamma_1}{\bar{L}_g(k+1)}\left[1 \pm \sqrt{1 - \frac{\bar{L}_g u_1(k+1)}{2\gamma_1 k}}\right], \tag{44}$$

satisfy $2 - \rho_k\beta_k^{-1}\bar{L}_A > 0$ and $2 - \lambda_k\gamma_k^{-1}\bar{L}_g > 0$, provided that $\bar{L}_A u_1 \leq \beta_1$ and $\bar{L}_g u_1 \leq \gamma_1$.

From Lemma 5, we can choose $\bar{L}_A u_1 = \beta_1$ and $\bar{L}_g u_1 = \lambda_1$. This leads to $\frac{\beta_1}{\gamma_1} = \frac{\bar{L}_A}{\bar{L}_g}$. Now, if we fix $\gamma_1 > 0$, then we can choose $\beta_1 := \frac{\bar{L}_A\gamma_1}{\bar{L}_g}$, and hence, $u_1 := \frac{\gamma_1}{\bar{L}_g}$. In addition, we can simplify the update rules for ρ_k and λ_k as $\rho_k := \frac{2\gamma_1}{\bar{L}_g(k+1)}\left[1 \pm \sqrt{\frac{k-1}{2k}}\right]$ and $\lambda_k := \frac{2\gamma_1}{\bar{L}_g(k+1)}\left[1 \pm \sqrt{\frac{k-1}{2k}}\right]$, respectively.

We note that, if we do not choose $\tau_0 = 1$, then the convergence guarantee for (ASGARD) depends on the value $G_{\gamma_0\beta_0}(\bar{w}^0)$. The values β_0 and γ_0 can arbitrarily be chosen such that they trade off the primal objective residual $f(\bar{x}^k) - f^\star$ and the feasibility gap $\|A\bar{x}^k - b\|$. The initial point $(\bar{x}^0, \bar{y}^0) \in \mathcal{W}$ can also arbitrarily be chosen, while setting $\tilde{x}^0 := \bar{x}^0$ and $\tilde{y}^0 := \bar{y}^0$.

We note that we limit the choice of the proximity functions L_{p_y} and L_{p_i} such that $L_{p_y} = L_{p_i} = 1$ for $i = 1, \cdots, N$. As a standard example, quadratic proximity functions satisfy this condition. In addition, we choose $\mathcal{I}_N = \emptyset$. This choice may increase the complexity-per-iteration of our method since it requires to handle the composite prox-functions $p_i(A_i(\cdot))$ instead of $p_i(\cdot)$, and may destroy the tractable proximity of $f_{\mathcal{X}_i}$. For general choices of p_y, p_i and $\mathcal{I}_N$, if we use ASGARD, then we can prove that (ASGARD) can achieve at least the $\mathcal{O}\left(\frac{\ln(k)}{k}\right)$-convergence rate.

4.3 The primal-dual algorithmic template

Similar to the accelerated scheme [3,32], we eliminate $(\tilde{x}^k, \tilde{y}^k)$ in (ASGARD) as

$$\begin{cases} \hat{x}^{k+1} := \bar{x}^{k+1} + \frac{(t_k\omega_k - 1)}{t_{k+1}}(\bar{x}^{k+1} - \hat{x}^k) + \frac{(t_k - 1)}{t_{k+1}}(\hat{x}^k - \bar{x}^k) \\ \hat{y}^{k+1} := \bar{y}^{k+1} + \frac{t_k\hat{\omega}_k - 1}{t_{k+1}}(\bar{y}^{k+1} - \hat{y}^k) + \frac{(t_k - 1)}{t_{k+1}}(\hat{y}^k - \bar{y}^k), \end{cases}$$

where $t_k := \frac{1}{\tau_k}$, $\omega_k := 2 - \rho_{k+1}\beta_{k+1}^{-1}\bar{L}_A > 0$ and $\hat{\omega}_k := 2 - \lambda_{k+1}\gamma_{k+1}^{-1}\bar{L}_g > 0$ due to Lemma 5.

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

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

Initialization:

1: Choose $\gamma_1 > 0$ and $\tau_0 := 1$. Set $\beta_1 := \frac{\bar{L}_A\gamma_1}{\bar{L}_g}$ and $u_1 := \frac{\gamma_1}{\bar{L}_g}$.

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

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

3: Update $\gamma_{k+1} := \frac{2\gamma_1}{k+2}$, $\beta_{k+1} := \frac{2\beta_1}{k+2}$ and $u_{k+1} := \frac{u_1}{k+1}$.

4: Compute $x^*_{\gamma_{k+1}}(\hat{y}^k)$ by (25) *in parallel*, and $y^*_{\beta_{k+1}}(\hat{x}^k)$ by (25).

5: Update $\rho_{k+1} := \frac{2\gamma_1}{\bar{L}_g(k+2)}\left[1 \pm \sqrt{\frac{k}{2(k+1)}}\right]$ and $\lambda_{k+1} := \frac{2\gamma_1}{\bar{L}_g(k+2)}\left[1 \pm \sqrt{\frac{k}{2(k+1)}}\right]$.

6: Update the primal step $\bar{x}^{k+1}$ *in parallel* using the prox of $f_{\mathcal{X}}$ as

$$\bar{x}^{k+1} := \mathrm{prox}_{\rho_{k+1}f_{\mathcal{X}}}\left(\hat{x}^k - \rho_{k+1}A^T y^*_{\beta_{k+1}}(\hat{x}^k)\right).$$

7: Update the dual step $\bar{y}^{k+1}$ as $\bar{y}^{k+1} := \hat{y}^k + \lambda_{k+1}\big(Ax^*_{\gamma_{k+1}}(\hat{y}^k) - b\big)$.

8: Compute $t_k := k + 1$, $t_{k+1} := k + 2$, $\omega_k := 2 - \rho_{k+1}\beta_{k+1}^{-1}\bar{L}_A$ and $\hat{\omega}_k := 2 - \lambda_{k+1}\gamma_{k+1}^{-1}\bar{L}_g$.

9: Update the primal and dual vectors

$$\begin{cases} \hat{x}^{k+1} := \bar{x}^{k+1} + \frac{(t_k\omega_k - 1)}{t_{k+1}}(\bar{x}^{k+1} - \hat{x}^k) + \frac{(t_k - 1)}{t_{k+1}}(\hat{x}^k - \bar{x}^k), \\ \hat{y}^{k+1} := \bar{y}^{k+1} + \frac{t_k\hat{\omega}_k - 1}{t_{k+1}}(\bar{y}^{k+1} - \hat{y}^k) + \frac{(t_k - 1)}{t_{k+1}}(\hat{y}^k - \bar{y}^k). \end{cases}$$

End for

The computationally heavy steps of Algorithm 1 are given by Steps 4, 6, and 7. At Step 4, we need to compute $x^*_{\gamma_{k+1}}(\hat{y}^k)$, which requires to solve the primal subproblem in (25) once. This computation can be implemented in parallel using the structure (7). In addition, $y^*_{\beta_{k+1}}(\hat{x}^k)$ at Step 4 needs a matrix-vector multiplication Ax. At Step 6, it requires one proximal-step on $f + \delta_{\mathcal{X}}$, which can also be implemented in parallel. For this step, we also need one adjoint matrix-vector multiplication $A^T y$. Step 7 demands only one matrix-vector multiplication Ax. We also note that the computation of $x^*_{\gamma_{k+1}}$ and $y^*_{\beta_{k+1}}$ at Step 4 is independent, which can be performed in parallel. Similarly, the update of $\bar{x}^{k+1}$ and $\bar{y}^{k+1}$ at Steps 6 and 7 can also be exchanged.

4.4 Convergence analysis

Under Assumption A.1, the dual solution set $\mathcal{Y}^\star$ of (12) is nonempty and bounded. Hence, $D_{\mathcal{Y}^\star}$ defined in Lemma 3 satisfies $D_{\mathcal{Y}^\star} \in [0, +\infty)$. The following theorem shows the convergence of Algorithm 1, which is proved in Appendix A.2.3.

Theorem 1 *Suppose that p_y and p_i are chosen such that $L_{p_y} = L_{p_i} = 1$ for $i = 1, \cdots, N$, and $\mathcal{I}_N = \emptyset$. Let $\{\bar{w}^k\}$ be the sequence generated by Algorithm 1. Then, we have*

$$\begin{cases} f(\bar{x}^k) - f^\star \le \frac{\bar{L}_g R_0^2}{2\gamma_1 k} + \frac{2\gamma_1}{(k+1)}\left(D_{\mathcal{X}} + \frac{\bar{L}_A}{\bar{L}_g} p_y(0)\right), \\ \|A\bar{x}^k - b\| \le \frac{2\gamma_1 \bar{L}_A}{\bar{L}_g(k+1)}\left[\bar{c}_\star + \sqrt{\bar{c}_\star^2 + 2L_{p_y}\left(\frac{\bar{L}_g^2 R_0^2}{\gamma_1^2 \bar{L}_A} + \frac{\bar{L}_g}{\bar{L}_A} D_{\mathcal{X}} + p_y(0)\right) - \|\bar{s}_c\|^2}\right], \end{cases} \tag{45}$$

where $\gamma_1 > 0$ is given, $\bar{c}_\star := \|L_{p_y} y^\star - \bar{s}^c\|_*$, and $R_0^2 := R_0(w^\star)^2 = \|\bar{y}^0 - y^\star\|^2 + \|\bar{x}^0 - x^\star\|^2$ and $D_{\mathcal{X}}$ is given in (29).

If we choose $p_y(\cdot) := \frac{1}{2}\|\cdot\|_2^2$, then we reach the following guarantees

$$\begin{cases} f(\bar{x}^k) - f^\star \le \frac{\bar{L}_g R_0^2}{2\gamma_1 k} + \frac{2\gamma_1 D_{\mathcal{X}}}{(k+1)}, \\ \|A\bar{x}^k - b\|_2 \le \frac{2\gamma_1 \bar{L}_A}{\bar{L}_g(k+1)}\left[2D_{\mathcal{Y}^\star} + \sqrt{2\left(\frac{\bar{L}_g^2 R_0^2}{\gamma_1^2 \bar{L}_A} + \frac{\bar{L}_g}{\bar{L}_A} D_{\mathcal{X}}\right)}\right]. \end{cases} \tag{46}$$

Note that $-D_{\mathcal{Y}^\star}\|A\bar{x}^k - b\| \le -\|y^\star\|_*\|A\bar{x}^k - b\| \le f(\bar{x}^k) - f^\star$ for any $\bar{x}^k \in \mathcal{X}$ and $y^\star \in \mathcal{Y}^\star$. As a consequence, then the worst-case iteration-complexity of Algorithm 1 to achieve an ε-primal solution $\bar{x}^k$ for (1) *in the sense of Definition 1 is $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$.*

The choice of γ_0 in Theorem 1 trades off between R_0^2 and $D_{\mathcal{X}}$ on the primal objective residual $f(\bar{x}^k) - f^\star$, while the choice of β_0 trades off the feasibility gap $\|A\bar{x}^k - b\|$. One limitation of Algorithm 1 is the presence of A_i in $p_i(A_i x_i)$ of the subproblem (53) when $i \notin \mathcal{I}_1$. In this case, if A_i is not orthogonal, then A_i may destroy the tractability of the proximal operator for $f_{\mathcal{X}_i} := f_i + \delta_{\mathcal{X}_i}$.

Remark 3 Since the proximal operator of $f_{\mathcal{X}}$ at Step 6 of Algorithm 1 can be computed *in parallel*, we can substitute it by the following proximal operator that takes into account the individual Lipschitz constant $\bar{L}_{A_i}$

$$\bar{x}_i^{k+1} := \mathrm{prox}_{\rho_{k+1} f_{\mathcal{X}_i}}\left(\hat{x}_i^k - \rho_{k+1} A_i^T y^*_{\beta_{k+1}}(\hat{x}^k)\right), \quad \forall i \in \mathcal{I}_N,$$

where $f_{\mathcal{X}_i}(\cdot) := f_i(\cdot) + \delta_{\mathcal{X}_i}(\cdot)$, with $\delta_{\mathcal{X}_i}$ being the indicator function of $\mathcal{X}_i$. In this case, the conclusions of Theorem 1 is preserved.

5 The accelerated dual smoothed gap reduction method

We develop a new primal-dual scheme that can remove one proximal operator computation in Algorithm 1 (i.e., Step 6) by means of averaging in the primal.

5.1 The accelerated dual gap reduction scheme

We assume that $\bar{w}^k := (\bar{x}^k, \bar{y}^k) \in \mathcal{W}$ is given. We derive below the update scheme for the new point $\bar{w}^{k+1} := (\bar{x}^{k+1}, \bar{y}^{k+1})$ from $\bar{w}^k$ such that the MGR condition (38) holds. This scheme includes two main steps: an accelerated gradient step on the smoothed dual function g_γ, and an averaging step to construct a primal point

$$\begin{cases} \hat{y}^k & := (1-\tau_k)\bar{y}^k + \tau_k y^*_{\beta_k}(\bar{x}^k), \\ \bar{y}^{k+1} & := \hat{y}^k + L^{-1}_{g_{\gamma_{k+1}}}\big(Ax^*_{\gamma_{k+1}}(\hat{y}^k) - b\big), \\ \bar{x}^{k+1} & := (1-\tau_k)\bar{x}^k + \tau_k x^*_{\gamma_{k+1}}(\hat{y}^k), \end{cases} \tag{ADSGARD}$$

where $\tau_k \in (0,1)$ and the parameters $\beta_k > 0$ and $\gamma_{k+1} > 0$ will be updated in the sequel. This scheme requires to solve one primal subproblem in (25) to compute $x^*_{\gamma_{k+1}}(\hat{y}^k)$, while it needs two dual steps of updating $y^*_{\beta_k}(\bar{x}^k)$ from the second line of (25), and $\bar{y}^{k+1}$. Since the accelerated step is applied to g_γ, we call this scheme the *Accelerated Dual Smoothed GAp ReDuction* (ADSGARD) scheme.

The following lemma shows that $\bar{w}^{k+1}$ updated by (ADSGARD) maintains (38), whose proof can also be found in Appendix A.3.1.

Lemma 6 *Let $\bar{w}^{k+1} := (\bar{x}^{k+1}, \bar{y}^{k+1})$ be updated by* (ADSGARD)*. Let $\bar{L}_g$ be defined by* (47)*, and $\hat{c}_2$ be the constant defined by*

$$\hat{c}_2 := \max\left\{\max_{i\in\mathcal{I}_1}\left\{\frac{L_{p_i}\bar{L}_{A_i}}{\sigma_i^2}\right\}, \max_{i\notin\mathcal{I}_1}\{L_{p_i}\}\right\}, \tag{47}$$

where σ_i^2 is given in (20)*, and $\bar{L}_{A_i}$ is defined by* (21)*. Then, if $\tau_k \in (0,1)$, $\beta_k > 0$ and $\gamma_{k+1} > 0$ are chosen such that*

$$\big(1+\hat{c}_2^{-1}\tau_k\big)\gamma_{k+1} \geq \gamma_k,\ \beta_{k+1} \geq (1-\tau_k)\beta_k,\ \text{and } (1-\tau_k)\gamma_{k+1}\beta_k \geq \bar{L}_g\tau_k^2, \tag{48}$$

then $\bar{w}^{k+1} \in \mathcal{W}$ and satisfies the gap reduction condition (38) *with*

$$\psi_k := -\frac{\gamma_{k+1}}{2\tau_k}\sum_{i=1}^N \frac{1}{\bar{L}_i}\|A_i(x^*_{\gamma_{k+1},i}(\hat{y}^k) - \bar{x}_i^c) - (1-\tau_k)A_i(x^*_{\gamma_{k+1},i}(\bar{y}^k) - \bar{x}_i^c)\|_*^2 \leq 0. \tag{49}$$

5.2 Updating the smoothing and gap reduction parameters

Our goal is to find update rules for $\{(\tau_k, \gamma_k, \beta_k)\}$ such that the condition (48) holds. One possibility is shown in the following lemma, which is proved in Appendix A.3.2.

Lemma 7 *Let us choose $\bar{c}_2 := \max\big\{\hat{c}_2, \frac{3}{2}\big\}$, where $\hat{c}_2$ is defined by* (47)*. Then, the parameters $(\tau_k, \gamma_k, \beta_k)$ updated by the following rules*

$$\tau_k = \frac{\bar{c}_2}{k+\bar{c}_2+1} \in (0,1),\ \gamma_k = \frac{(\bar{c}_2+1)\gamma_0}{k+\bar{c}_2+1}\ \text{and}\ \beta_k = \frac{\bar{c}_2^2\bar{L}_g(k+\bar{c}_2+2)}{\gamma_0(\bar{c}_2+1)(k+1)(k+\bar{c}_2+1)}, \tag{50}$$

satisfies the condition (48)*. Moreover, the convergence rate of $\{\tau_k\}$ is optimal. In addition, $\beta_k \leq \frac{\bar{c}_2^2\bar{L}_g}{\gamma_0(\bar{c}_2+1)k}$ and $\gamma_k\beta_k < \frac{\bar{c}_2^3\bar{L}_g}{(\bar{c}+1)k(k+\bar{c}_2+1)}$.*

From (48) we can derive the tightest condition $\gamma_k\beta_k = \frac{\hat{c}_2\bar{L}_g\tau_k^2}{(1-\tau_k)(\hat{c}_2+\tau_k)}$. Hence, the optimal convergence rate of $\{\gamma_k\beta_k\}$ is $\gamma_k\beta_k = \mathcal{O}\left(\frac{1}{k^2}\right)$. On the other hand, by Lemma 3, we see that $f(\bar{x}^k) - f^\star = \mathcal{O}(\gamma_k)$ and $\|A\bar{x}^k - b\| = \mathcal{O}(\beta_k)$. Since $\gamma_k\beta_k = \mathcal{O}\left(\frac{1}{k^2}\right)$, if we decrease the rate of γ_k, i.e., decrease the objective residual $f(\bar{x}^k)-f^\star$, then the rate of β_k is increased, i.e., we increase the rate of the feasibility gap $\|A\bar{x}^k-b\|$. The same augment can be applied to the case where γ_k is increasing.

5.3 Finding initial points for the monotone gap reduction model

In principle, we can start (ADSGARD) at any initial point $\bar{w}^0 \in \mathcal{W}$. However, the worst-case complexity bounds will depend on the value $G_{\gamma_0\beta_0}(\bar{w}^0)$. To simplify this worst-case complexity bound, we show how to construct a point $\bar{w}^0 \in \mathcal{W}$ such that condition (38) holds with $\psi_0 \le 0$. Let $\bar{y}^c := \nabla p_y^*(\mathbf{0}^m) \in \mathbb{R}^m$ be the prox-center of p_y. We compute the point $\bar{w}^0 := (\bar{x}^0, \bar{y}^0)$ based on the following scheme:

$$\left\{\begin{array}{l} \bar{x}^0 = x^*_{\gamma_0}(\bar{y}^c) := \mathrm{argmin}\left\{f(x) + \langle \bar{y}^c, Ax - b\rangle + \gamma_0 p_x(Ax) : x \in \mathcal{X}\right\}, \\ \bar{y}^0 = y^*_{\beta_0}(\bar{x}^0) := \nabla {p_y}^*\left(\beta_0^{-1}(A\bar{x}^0 - b)\right), \end{array}\right. \tag{51}$$

where $\gamma_0 > 0$ and $\beta_0 > 0$. The following lemma shows that $\bar{w}^0$ computed by (51) satisfies (38) with $\psi_0 \le 0$, whose proof can be found in Appendix A.3.3.

Lemma 8 *If $\bar{w}^0 := (\bar{x}^0, \bar{y}^0)$ is generated by (51), then it satisfies*

$$G_{\gamma_0\beta_0}(\bar{w}^0) \le -\gamma_0 p_x(A\bar{x}^0) - \left(2\beta_0^2\gamma_0 L_{p_y}\right)^{-1}\left(\gamma_0\beta - \bar{L}_g L_{p_y}\right)\|A\bar{x}^0 - b\|^2, \tag{52}$$

Hence, if γ_0 and β_0 are chosen such that $\gamma_0\beta_0 \ge L_{p_y}\bar{L}_g$, then $G_{\gamma_0\beta_0}(\bar{w}^0) \le 0$.

Finally, given $\gamma_0 > 0$, from (50) we have $\gamma_0\beta_0 = \frac{\bar{L}_g\bar{c}_2^2(\bar{c}_2+2)}{(\bar{c}_2+1)^2} > \frac{2}{3}\bar{c}_2\bar{L}_g$. Hence, if $\bar{c}_2 \ge \frac{3}{2}L_{p_y}$, then $\gamma_0\beta_0 \ge \bar{L}_g L_{p_y}$, which is the condition of Lemma 8. We note that $L_{p_y} \ge 1$. Hence, if to choose $\bar{c}_2 = \max\left\{\hat{c}_2, \frac{3}{2}L_{p_y}\right\}$, then both conditions in Lemma 7 and Lemma 8 are satisfied. In Algorithm 2 below, we choose this value for the constant $\bar{c}_2$.

5.4 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) as shown in Algorithm 2.

The main steps of Algorithm 2 are Steps 6, 8 and 9, where we need to solve the primal convex subproblem (53), to update two dual steps, respectively. While solving (53) can be implemented in a parallel or distributed fashion due to the decomposable structure of f and $\mathcal{X}$ as in (7), the dual steps only require matrix-vector multiplication Ax. Clearly, by Step 10, it follows that $A\bar{x}^{k+1} - b = (1 - \tau_k)(A\bar{x}^k - b) + \tau_k(A\hat{x}^*_{k+1} - b)$, and by Step 6, we have $\bar{y}^*_k = \nabla {p_y}^*\left(\beta_k^{-1}(A\bar{x}^k - b)\right)$, which is equivalent to $A\bar{x}^k - b = \beta_k\nabla p_y(\bar{y}^*_k)$. Hence, $A\bar{x}^{k+1} - b = (1-\tau_k)\beta_k\nabla p_y(\bar{y}^*_k) + \frac{\tau_k\bar{L}_g}{\gamma_{k+1}}(\bar{y}^{k+1} - \hat{y}^k)$ due to Step 9. Finally, we can derive

$$\bar{y}^*_{k+1} = \nabla p_y^*\left(\beta_{k+1}^{-1}\left((1-\tau_k)\beta_k\nabla p_y(\bar{y}^*_k) + \frac{\bar{L}_g\tau_k}{\gamma_{k+1}}(\bar{y}^{k+1} - \hat{y}^k)\right)\right). \tag{54}$$

Consequently, each iteration of Algorithm 2 requires one solution of the primal convex subproblem (53), one matrix-vector multiplication Ax and its adjoint A^Ty.

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

Initialization:

Choose $\gamma_0 > 0$ (e.g., $\gamma_0 := \sqrt{2\bar{L}_g}$), compute $\bar{c}_2 := \max\left\{\hat{c}_2, \frac{3}{2}L_{p_y}\right\}$.

Set $\beta_0 := \frac{\bar{L}_g\bar{c}_2^2(\bar{c}_2+2)}{\gamma_0(\bar{c}_2+1)^2}$ and $\bar{y}^c := \nabla p_y^*(\mathbf{0}^m)$.

Solve the following primal convex subproblem

$$\bar{x}^0 := \mathrm{argmin}\left\{f(x) + \langle A^T\bar{y}^c, x\rangle + \gamma_0 p_x(Ax) : x \in \mathcal{X}\right\}.$$

Compute $\bar{y}^0 := \nabla p_y{}^*\left(\beta_0^{-1}(A\bar{x}^0 - b)\right)$.

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

Update $\tau_k := \frac{\bar{c}_2}{k+\bar{c}_2+1}$, $\gamma_{k+1} := \frac{(\bar{c}_2+1)\gamma_0}{k+\bar{c}_2+2}$ and $\beta_k := \frac{\bar{c}_2^2\bar{L}_g(k+\bar{c}_2+2)}{\gamma_0(\bar{c}_2+1)(k+1)(k+\bar{c}_2+1)}$.

Compute $\bar{y}_k^* := \nabla p_y{}^*\left(\beta_k^{-1}(A\bar{x}^k - b)\right)$.

Update $\hat{y}^k := (1-\tau_k)\bar{y}^k + \tau_k\bar{y}_k^*$.

Solve the following convex subproblem

$$\hat{x}_{k+1}^* := \mathrm{argmin}\left\{f(x) + \langle A^T\hat{y}^k, x\rangle + \gamma_{k+1}p_x(Ax) : x \in \mathcal{X}\right\}. \tag{53}$$

Update the dual vector: $\bar{y}^{k+1} := \hat{y}^k + \frac{\gamma_{k+1}}{\bar{L}_g}\left(A\hat{x}_{k+1}^* - b\right)$.

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

End for

5.5 Convergence analysis

Let $D_{\mathcal{Y}^\star}$ be defined in Lemma 3. The following theorem shows the convergence of Algorithm 2, while the lower bound on $f(\bar{x}^k) - f^\star$ remains as in Theorem 1, i.e.: $-D_{\mathcal{Y}^\star}\|A\bar{x}^k - b\| \le -\|y^\star\|_*\|A\bar{x}^k - b\| \le f(\bar{x}^k) - f^\star$ for any $\bar{x}^k \in \mathcal{X}$ and $y^\star \in \mathcal{Y}^\star$.

Theorem 2 *Let $\{\bar{w}^k\}$ be the sequence generated by Algorithm 2. Then, the following convergence bounds hold*

$$\begin{cases} f(\bar{x}^k) - f^\star \le \frac{(\bar{c}_2+1)\gamma_0}{k+\bar{c}_2+1}D_{\mathcal{X}} + \frac{\bar{c}_2^2\bar{L}_g p_y(0)(k+\bar{c}_2+2)}{\gamma_0(\bar{c}_2+1)(k+\bar{c}_2+1)(k+1)}, \\ \|A\bar{x}^k - b\| \le \frac{\bar{c}_2^2\bar{L}_g(k+\bar{c}_2+2)}{\gamma_0(\bar{c}_2+1)(k+1)(k+\bar{c}_2+1)}\left[\|L_{p_y}y^\star - \bar{s}_c\|_*^2 + \sqrt{\|L_{p_y}y^\star - \bar{s}_c\|_*^2 + \bar{P}_*}\right], \end{cases} \tag{55}$$

where $\bar{P}_* := \frac{2L_{p_y}\gamma_0^2(\bar{c}_2+1)^2}{\bar{c}_2^2\bar{L}_g} + 2L_{p_y}p_y(0) - \|\bar{s}^c\|^2$, $D_{\mathcal{X}}$ is defined by (29) *and $y^\star \in \mathcal{Y}^\star$.*

If we choose $p_y(\cdot) := \frac{1}{2}\|\cdot\|_2^2$, then we have the following guarantees

$$\begin{cases} f(\bar{x}^k) - f^\star \le \frac{(\bar{c}_2+1)\gamma_0}{k+\bar{c}_2+1}D_{\mathcal{X}}, \\ \|A\bar{x}^k - b\|_2 \le \frac{\bar{c}_2^2\bar{L}_g(k+\bar{c}_2+2)}{\gamma_0(\bar{c}_2+1)(k+1)(k+\bar{c}_2+1)}\left[2D_{\mathcal{Y}^\star} + \frac{\sqrt{2}\gamma_0(\bar{c}_2+1)}{\bar{c}_2\sqrt{\bar{L}_g}}\right]. \end{cases} \tag{56}$$

As a consequence, if $\gamma_0 := \sqrt{\bar{L}_g}$, then the worst-case iteration-complexity of Algorithm 2 to achieve an ε-solution $\bar{x}^k$ for (1) *in the sense of Definition 1 is $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$.*

Proof Substituting the expression of γ_k and β_k from Lemma 7 into (36) of Lemma 3, and then using $\frac{\gamma_k}{\beta_k} = \frac{\gamma_0^2(\bar{c}_2+1)^2(k+1)}{\bar{c}_2^2\bar{L}_g(k+\bar{c}_2+2)} \le \frac{\gamma_0^2(\bar{c}_2+1)^2}{\bar{c}_2^2\bar{L}_g}$ we directly obtain (56).

With $p_y(y) := (1/2)\|y\|_2^2$, we substituting γ_k, β_k and $\gamma_k\beta_k$ from Lemma 7 into (37) of Lemma 3 to obtain the bounds (56). The remaining conclusions are the consequences of (56). □

The choice of γ_0 in Theorem 2 also trades off the primal objective residual and the primal feasibility gap. Indeed, the smaller γ_0 leads to the smaller $f(\bar{x}^k) - f^\star$. One limitation of Algorithm 2 is the presence of A_i in the composite prox-function $p_i(A_i(\cdot))$ of the subproblem (53). When A_i is not orthogonal, the operator A_i may destroy the tractability of the proximal operator for $f_{\mathcal{X}_i} := f_i + \delta_{\mathcal{X}_i}$.

6 The accelerated primal smoothed gap reduction method

Even if $f_{\mathcal{X}_i}$ has a tractable proximity operator when $i \notin \mathcal{I}_1$, the presence of A_i in p_i can require significant computation. As a result, the ADSGARD scheme may have a disadvantage. To overcome this drawback, we propose in this section as a symmetric variant of (ADSGARD) that relies on the acceleration of the primal. In contrast to (22), we use the following smoother

$$p_{\gamma\beta}(w) := \gamma p_x(x) + \beta p_y(y),$$

where $p_x(x) := \sum_{i\in\mathcal{I}} p_i(x_i)$, with p_i being the prox-function of $\mathcal{X}_i$ for all $i \in \mathcal{I}$.

Let $\bar{w}^k := (\bar{x}^k, \bar{y}^k) \in \mathcal{W}$ be given. We update the new point $\bar{w}^{k+1} := (\bar{x}^{k+1}, \bar{y}^{k+1})$ from $\bar{w}^k$ using the following scheme to maintain the MGR condition (38)

$$\left\{\begin{array}{ll} \hat{x}^k &:= (1-\tau_k)\bar{x}^k + \tau_k x^*_{\gamma_k}(\bar{y}^k)\\ \bar{x}^{k+1} &:= \mathrm{prox}_{\frac{\beta_{k+1}}{\bar{L}_A} f_{\mathcal{X}}}\left(\hat{x}^k - \frac{\beta_{k+1}}{\bar{L}_A}A^T y^*_{\beta_{k+1}}(\hat{x}^k)\right),\\ \bar{y}^{k+1} &:= (1-\tau_k)\bar{y}^k + \tau_k y^*_{\beta_{k+1}}(\hat{x}^k), \end{array}\right. \tag{APSGARD}$$

where $\tau_k \in (0,1)$ and the parameters $\beta_k > 0$ and $\gamma_{k+1} > 0$, which will be updated in the sequel. Since this scheme performs an accelerated proximal-gradient step on the primal term f_β of the smoothed gap $G_{\gamma\beta}$, we call this scheme the *Accelerated Primal Smoothed GAp ReDuction* (APSGARD) scheme.

We note that, the point $x^*_{\gamma_k}(\bar{y}^k) := (x^*_{\gamma_k,1}(\bar{y}^k), \cdots, x^*_{\gamma_k,N}(\bar{y}^k))$ at the first line of (APSGARD) is computed as follows:

$$x^*_{\gamma_k,i}(\bar{y}^k) := \mathop{\mathrm{argmin}}_{x_i}\Big\{f_i(x_i) + \langle \bar{y}^k, A_i x_i - b_i\rangle + \gamma_k p_i(x_i) : x_i \in \mathcal{X}_i\Big\}, \quad \forall i \in \mathcal{I}_N, \tag{57}$$

without using the composite prox-function $p_i(A_i(\cdot))$ for $i \notin \mathcal{I}_1$ as in (ASGARD) and (ADSGARD). Hence, we can exploit the tractable proximity of $f_{\mathcal{X}_i}$ for $i \notin \mathcal{I}_1$. However, the downside is that (APSGARD) now requires an additional proximal step of $f_{\mathcal{X}}$ at the second line of (APSGARD).

Similar to Lemma 6, we can show that $\{\bar{w}^k\}$ generated by (APSGARD) maintains the MGR condition (38) if the parameters τ_k, γ_k and β_k satisfy

$$\gamma_{k+1} \geq (1-\tau_k)\gamma_k, \quad \beta_{k+1}\left(1 + \frac{\tau_k}{L^2_{p_y}}\right) \geq \beta_k, \quad \text{and} \quad (1-\tau_k)\gamma_k\beta_{k+1} \geq \bar{L}_A\tau_k^2. \tag{58}$$

Under these conditions, we propose one update rule for τ_k, γ_k and β_k as follows:

$$\tau_k := \frac{\bar{c}_3}{k+\bar{c}_3+1}, \quad \beta_k := \frac{(\bar{c}_3+1)\beta_0}{k+\bar{c}_3+1}, \quad \text{and} \quad \gamma_k := \frac{\bar{L}_A\bar{c}_3^2(k+\bar{c}_3+2)}{\beta_0(\bar{c}_3+1)(k+1)(k+\bar{c}_3+1)}, \tag{59}$$

where $\bar{c}_3 := \max\left\{\frac{3}{2}L_{p_y}, L_{p_y}^2\right\}$ and $\beta_0 > 0$ is given. Clearly, the constant $\bar{c}_3$ does not depend on matrix A. If we choose $p_y(\cdot) := (1/2)\|\cdot\|_2^2$, then $\bar{c}_3 := \frac{3}{2}$.

Now, we summarize the convergence of Algorithm 2 using (APSGARD) as a substitute to (ADSGARD). The proof of this theorem can be found in Appendix A.4. Here, the lower bound on $f(\bar{x}^k) - f^\star$ remains as in Theorem 1.

Theorem 3 *Let $\{\bar{w}^k\}$ be the sequence generated by Algorithm 2 using the primal-dual scheme* (APSGARD) *and the update rules* (59) *with given $\beta_0 > 0$. Then, the following estimate holds*

$$\begin{cases} f(\bar{x}^k) - f^\star \le \frac{\bar{c}_3^2\bar{L}_A(k+\bar{c}_3+2)}{\beta_0(\bar{c}_3+1)(k+\bar{c}_3+1)(k+1)}D_{\mathcal{X}} + \frac{\beta_0(\bar{c}_3+1)p_y(0)}{k+\bar{c}_3+1}, \\ \|A\bar{x}^k - b\| \le \frac{\beta_0(\bar{c}_3+1)}{k+\bar{c}_3+1}\left[\bar{c}_* + \sqrt{\bar{c}_* + \frac{2L_{p_y}\bar{L}_A\bar{c}_3^2(\bar{c}_3+2)D_{\mathcal{X}}}{\beta_0^2(\bar{c}_3+1)^2} + (2L_{p_y}p_y(0) - \|\bar{s}_c\|^2)}\right], \end{cases} \tag{60}$$

where $\bar{c}_* := \|L_{p_y}y^\star - \bar{s}_c\|_*$, $D_{\mathcal{X}}$ is defined by (29) *and $y^\star \in \mathcal{Y}^\star$.*

If we choose $p_y(\cdot) := \frac{1}{2}\|\cdot\|_2^2$, then we have the following guarantees

$$\begin{cases} f(\bar{x}^k) - f^\star \le \frac{9\bar{L}_A(2k+7)D_{\mathcal{X}}}{10\beta_0(2k+5)(k+1)}, \\ \|A\bar{x}^k - b\|_2 \le \frac{5\beta_0}{2k+5}\left[2D_{\mathcal{Y}^\star} + \frac{3\sqrt{7\bar{L}_A D_{\mathcal{X}}}}{5\beta_0}\right]. \end{cases} \tag{61}$$

As a consequence, if $\beta_0 := \sqrt{\bar{L}_A}$, then the worst-case iteration-complexity of this algorithm to achieve an ε-solution $\bar{x}^k$ for (1) *in the sense of Definition 1 is $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$.*

We note that we still use the initial point $\bar{w}^0$ as in (51) for this variant. In Theorem 3, the value β_0 trades off between $f(\bar{x}^k) - f^\star$ and $\|A\bar{x}^k - b\|$ instead of γ_0 as in Theorem 2.

7 Special instances of the primal-dual gap reduction framework

This section specifies Algorithms 1 and 2 to solve (1) by further exploiting the its structures as well as using difference choices of the prox-function.

7.1 Accelerated augmented Lagrangian smoothed gap reduction method

When f and $\mathcal{X}$ are not decomposable, i.e., $N = 1$, we can modify Algorithms 1 and 2 to obtain a new augmented Lagrangian algorithm. For clarity of exposition, we only present Algorithm 2 using (ADSGARD) in the sequel. The inexact variant of this algorithm can be found in our early technical report [49, Section 5.3].

The augmented Lagrangian smoother: Under Assumption A.1, there exists a feasible point $\bar{x}^c \in \mathcal{X}$ such that $A\bar{x}^c = b$. We choose the prox-function p_x as $p_x(u) := (1/2)\|u-b\|_2^2$. In this case $p_x(Ax) = (1/2)\|A(x-\bar{x}^c)\|_2^2 = (1/2)\|Ax-b\|_2^2$, which is the augmented term in the augmented Lagrangian method. Alternatively, we choose $p_y(\cdot) := (1/2)\|\cdot\|_2^2$ for the dual smoother. It is well-known that the augmented Lagrangian method is simply the proximal-point method applying to (12).

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

$$\begin{cases} \hat{y}^k &:= (1-\tau_k)\bar{y}^k + \tau_k\beta_k^{-1}(A\bar{x}^k - b), \\ \hat{x}_{k+1}^* &:= \mathrm{argmin}_{x\in\mathcal{X}}\left\{f(x) + \langle\hat{y}^k, Ax-b\rangle + \frac{\gamma_0}{2}\|Ax-b\|_2^2\right\}, \\ \bar{y}^{k+1} &:= \hat{y}^k + \gamma_0(A\hat{x}_{k+1}^* - b), \\ \bar{x}^{k+1} &:= (1-\tau_k)\bar{x}^k + \tau_k\hat{x}_{k+1}^*, \end{cases} \tag{FALSGARD}$$

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

This scheme consists of two dual steps at lines 1 and 3. However, we can combine these steps as in (54) so that it requires only one matrix-vector multiplication Ax. Consequently, the complexity-per-iteration of (FALSGARD) remains essentially the same as the standard augmented Lagrangian method [8].

The initial point: Similar to (51), we can initialize (FALSGARD) via

$$\begin{cases} \bar{x}^0 := \mathrm{argmin}\left\{f(x) + (\gamma_0/2)\|Ax - b\|_2^2 : x \in \mathcal{X}\right\}, \\ \bar{y}^0 := \beta_0^{-1}\left(A\bar{x}^0 - b\right), \end{cases} \tag{62}$$

where β_0 is chosen such that $\gamma_0\beta_0 \geq 1$ and $\bar{y}^c := \nabla p_y^*(\mathbf{0}^m) = \mathbf{0}^m$. Clearly, with $\bar{w}^0 := (\bar{x}^0, \bar{y}^0)$ computed by this formula, we have $G_0(\bar{w}^0) \leq 0$ due to Lemma 8.

The update rule for parameters: In our augmented Lagrangian method, we can set $\gamma_k = \gamma_0 > 0$ to be constant, while updating τ_k and β_k such that the two last conditions (48) of Lemma 6 hold. Using these conditions we can derive an update rule for β_k and τ_k as follows:

$$\beta_{k+1} := (1 - \tau_k)\beta_k, \quad \text{and} \quad \tau_{k+1} := \frac{\tau_k}{2}\Big(\sqrt{\tau_k^2 + 4} - \tau_k\Big). \tag{63}$$

If we choose $\tau_0 := 0.5(\sqrt{5} - 1)$, we have $\beta_0\gamma_0 = \frac{\tau_0^2}{1-\tau_0} = 1$, which satisfies the condition in Lemma 8.

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

Algorithm 3 (*Fast Augmented Lagrangian Smoothed GAp ReDuction*)

Initialization:

1: Choose an initial value $\gamma_0 > 0$. Set $\tau_0 := 0.5(\sqrt{5} - 1)$ and $\beta_0 := \gamma_0^{-1}$.
2: Compute $\bar{w}^0 := (\bar{x}^0, \bar{y}^0)$ by (62).

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

3: Update $\beta_{k+1} := (1 - \tau_k)\beta_k$.
4: Update $\bar{w}^{k+1} := (\bar{x}^{k+1}, \bar{y}^{k+1})$ using (FALSGARD).
5: Update $\tau_{k+1} := 0.5\tau_k\Big(\sqrt{\tau_k^2 + 4} - \tau_k\Big)$.

End for

The main step of Algorithm 3 is the solution of the primal convex subproblem

$$\hat{x}_{k+1}^* := \mathrm{argmin}\Big\{f(x) + \langle \hat{y}^k, Ax - b\rangle + (\gamma_0/2)\|Ax - b\|_2^2 : x \in \mathcal{X}\Big\}. \tag{64}$$

In general, solving this subproblem remains challenging due to the non-separability of the quadratic term $\|Ax - b\|_2^2$. 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 [30].

Convergence guarantee: The following theorem shows the convergence of Algorithm 3, whose proof is moved to Appendix A.5.

Theorem 4 *Let $\{\bar{w}^k\}$ be the sequence generated by Algorithm 3. Then, we have*

$$\begin{cases} -\frac{1}{2}\|A\bar{x}^k-b\|_2^2 - D_{\mathcal{Y}^\star}\|A\bar{x}^k-b\|_2 \le f(\bar{x}^k) - f^\star \le 0, \\ \qquad\qquad\qquad\qquad \|A\bar{x}^k-b\|_2 \le \frac{8D_{\mathcal{Y}^\star}}{\gamma_0(k+1)^2}. \end{cases} \tag{65}$$

As a consequence, the worst-case iteration-complexity of Algorithm 3 to achieve an ε-primal solution $\bar{x}^k$ for (1) in the sense of Definition 1 is $\mathcal{O}\left(\frac{D_{\mathcal{Y}^\star}}{\sqrt{\gamma_0\varepsilon}}\right)$.

The estimate (65) guides us to choose a large value for γ_0 such that we obtain better convergence bounds. However, if γ_0 is too large, then the complexity of solving the subproblem (23) increases commensurately. In practice, γ_0 is often updated using a heuristic strategy [8,10]. The bound (65) shows that the sequence $\{f(\bar{x}^k)\}$ converges to $f^\star$ from below, which is different from unconstrained setting, where $f(\bar{x}^k) \ge f^\star$. In addition, this bound does not depend on the diameter of $\mathcal{X}$, which shows that $\mathcal{X}$ is not necessary to be bounded as in Assumption A.2. In general settings, since the solution $\hat{x}^*_{k+1}$ computed by (64) requires to solve a generic convex problem, it no longer has a closed form expression.

7.2 The strong convexity of the objective function

If the objective function f_i of (1) is strongly convex with the convexity parameter $\mu_{f_i} > 0$ for all $i \in \mathcal{I}_N$, then it is well-known that the dual function g defined by (10) is smooth. Its gradient is given by $\nabla g(y) := Ax^*(y) - b$ which is Lipschitz continuous with the Lipschitz constant $\hat{L}_g := \sum_{i=1}^N \mu_{f_i}^{-1}\|A_i\|^2$ (see [35]), where $x^*(y)$ is the unique solution of the following primal subproblem:

$$x^*(y) := \mathop{\mathrm{argmin}}_{x\in\mathcal{X}}\big\{f(x) + \langle A^Ty, x\rangle\big\} = \sum_{i=1}^N \mathop{\mathrm{arg\,min}}_{x_i\in\mathcal{X}_i}\big\{f_i(x_i) + \langle A_i^Ty, x_i\rangle\big\}. \tag{66}$$

When f and $\mathcal{X}$ are decomposable as (7) with $N \ge 2$, we compute $x^*(y)$ *in parallel*.

The primal-dual update scheme: Principally, we can modify scheme (ASGARD), (ADSGARD) or (APSGARD) to adapt this strongly convex structure. In this subsection, we only illustrate the modification of (ADSGARD) as follows:

$$\begin{cases} \hat{y}^k & := (1-\tau_k)\bar{y}^k + \tau_k y^*_{\beta_k}(\bar{x}^k) \\ \bar{x}^{k+1} & := (1-\tau_k)\bar{x}^k + \tau_k x^*(\hat{y}^k) \\ \bar{y}^{k+1} & := \hat{y}^k + \hat{L}_g^{-1}\big(Ax^*(\hat{y}^k) - b\big). \end{cases} \tag{ADSGARD$_\mu$}$$

We note that we no longer have the primal smoothness parameter γ_k. Hence, the conditions (48) of Lemma 6 reduce to $\beta_{k+1} \ge (1-\tau_k)\beta_k$ and $(1-\tau_k)\beta_k \ge \hat{L}_g\tau_k^2$. From these conditions can derive the update rule of τ_k and β_k as in Algorithm 3

$$\tau_{k+1} := 0.5\tau_k\big[(\tau_k^2+4)^{1/2} - \tau_k\big] \quad\text{and}\quad \beta_{k+1} := (1-\tau_k)\beta_k, \tag{67}$$

where $\tau_0 := 0.5(\sqrt{5}-1)$ and $\beta_0 := \hat{L}_g$.

We can also modify (51) to compute the initial point $\bar{w}^0 := (\bar{x}^0, \bar{y}^0)$ as

$$\bar{x}^0 := \mathop{\mathrm{argmin}}_{x\in\mathcal{X}}\big\{f(x) + \langle A^T\bar{y}^c, x\rangle\big\}, \quad\text{and}\quad \bar{y}^0 := \bar{y}^c + \nabla p_y^*\left(\hat{L}_g^{-1}(A\bar{x}^0 - b)\right), \tag{68}$$

where $\bar{y}^c := \nabla p_y^*(\mathbf{0}^m)$ is a the prox-center of p_y. Clearly, if we choose $p_y(\cdot) := (1/2)\|\cdot\|_2^2$, then $\bar{x}^0 := \mathrm{argmin}\{f(x) : x \in \mathcal{X}\}$ and $\bar{y}^0 := \hat{L}_g^{-1}(A\bar{x}^0 - b)$.

Convergence guarantee: The following corollary shows the convergence of the scheme (ADSGARD$_\mu$), whose proof is in Appendix A.6.

Corollary 1 *Suppose that the objective f_i of* (1) *is strongly convex with the convexity parameter $\mu_{f_i} > 0$ for all $i \in \mathcal{I}_N$. Let $\left\{\bar{w}^k\right\}$ be a sequence generated by* (ADSGARD$_\mu$) *using the initial point* (68) *and the update rule* (67)*. Then*

$$\begin{cases} -D_{\mathcal{Y}^\star}\|A\bar{x}^k - b\| \le f(\bar{x}^k) - f^\star \le 0, \\ \qquad\qquad \|A\bar{x}^k - b\| \;\le \frac{4\hat{L}_g D_{\mathcal{Y}^\star}}{(k+2)^2}, \\ \qquad\qquad \|\bar{x}^k - x^\star\| \;\le \frac{4D_{\mathcal{Y}^\star}}{(k+2)}\sqrt{\frac{\hat{L}_g}{\underline{\mu}_f}}, \end{cases} \tag{69}$$

where $\hat{L}_g := \sum_{i=1}^N \frac{\|A_i\|^2}{\mu_{f_i}}$, $\underline{\mu}_f := \min\left\{\mu_{f_i} : i \in \mathcal{I}_N\right\} > 0$, $D_{\mathcal{Y}^\star}$ is defined in Theorem 2, and $x^\star \in \mathcal{X}^\star$. As a consequence, the worst-case iteration-complexity to attain an ε-solution $\bar{x}^k$ of (1) *in the sense of Definition 1 is $\mathcal{O}\left(D_{\mathcal{Y}^\star}\sqrt{\frac{\hat{L}_g}{\varepsilon}}\right)$.*

We note that the bounds in Corollary 1 does not require the boundedness of $\mathcal{X}$ as assumed in Assumption A.2. In addition, $\left\{f(\bar{x}^k)\right\}$ converges to $f^\star$ from below.

7.3 The component-wise strong convexity of the objective function

Let us denote by $\mathcal{I}_s := \left\{i \in \mathcal{I}_N : \mu_{f_i} \equiv \mu(f_i) > 0\right\}$ the index subset of strongly convex objective components f_i. If there exists $i \in \mathcal{I}_N$ such that f_i is strongly convex, then $\mathcal{I}_s \neq \emptyset$. Hence, g_γ defined by (23) can be replaced by the following

$$g_\gamma(y) := \sum_{i\in\mathcal{I}_s} g^i(y) + \sum_{i\notin\mathcal{I}_s} g^i_\gamma(y), \tag{70}$$

where g^i is the dual component defined in (13) and g^i_γ is the smoothed dual component defined by (27). We again illustrate a modification of the scheme (ADSGARD) to adapt this structure. First, the Lipschitz constant $\bar{L}_i$ defined by (21) becomes $\bar{L}_i := \frac{\|A_i\|^2}{\mu_{f_i}}$ for $i \in \mathcal{I}_s$, the prox-diameter $D_{\mathcal{X}}$ defined by (29) is $D_{\mathcal{X}} := \sum_{i\notin\mathcal{I}_s} D_{\mathcal{X}_i}$. Second, the conditions for selecting parameters τ_k, γ_k and β_k remain the same as in (48), where $\hat{c}_2$ is replaced by

$$\hat{c}_4 := \max\left\{\max_{i\in\mathcal{I}_1\setminus\mathcal{I}_s}\left\{\frac{L_{p_i}\bar{L}_{A_i}}{\sigma_i^2}\right\}, \max_{i\in\bar{\mathcal{I}}_1\setminus\mathcal{I}_s}\left\{L_{p_i}\right\}\right\}. \tag{71}$$

Using these modifications, we obtain a new variant of Algorithm 2 to adapt this structure. Consequently, the conclusions of Theorem 2 are preserved.

7.4 The Lipschitz gradient continuity of the full objective

If the objective function f of (1) is smooth, and its gradient ∇f is Lipschitz continuous with the Lipschitz constant $L_f := L(f) > 0$, then we can modify the proximal step of Algorithm 1 and Algorithm 2 to further exploit this structure.

Instead of using the proximal step of $f_{\mathcal{X}} := f + \delta_{\mathcal{X}}$ as in the second line of (ASGARD) and (APSGARD), we use the following projected gradient step

$$\bar{x}^{k+1} = \mathrm{proj}_{\mathcal{X}}\left(\hat{x}^k - L^{-1}_{f_{\beta_{k+1}}}\left(\nabla f(\hat{x}^k) + A^T y^*_{\beta_{k+1}}(\hat{x}^k)\right)\right), \tag{72}$$

where $\text{proj}_{\mathcal{X}}$ is the projection onto $\mathcal{X}$ and $L_{f_{\beta_{k+1}}} := L_f + \beta_{k+1}^{-1}\bar{L}_A$ is the Lipschitz constant of $\nabla f_{\beta_{k+1}}(\cdot) = \nabla f(\cdot) + A^T y^*_{\beta_{k+1}}(\cdot)$. Clearly, the projected gradient step (72) is generally cheaper than the prox-step $\text{prox}_{f_{\mathcal{X}}}$ of $f_{\mathcal{X}}$, especially when $\mathcal{X}$ is simple (e.g., bound, box and cone constraints) or f is non-decomposable.

In this case, the last condition for updating parameters in (40) is replaced by

$$\frac{\tau_k^2}{(1-\tau_k)\tau_{k-1}^2} \leq \min\left\{\frac{\gamma_k}{\gamma_{k-1}}, \frac{L_f + \beta_{k-1}^{-1}\bar{L}_A}{L_f + \beta_k^{-1}\bar{L}_A}\right\}, \tag{73}$$

while the last condition in (58) becomes

$$(1-\tau_k)\gamma_k \geq (L_f + \beta_{k+1}^{-1}\bar{L}_A)\tau_k^2. \tag{74}$$

From these conditions, we can derive the update rule for parameters τ_k, γ_k and β_k, respectively for each case. We omit the derivation details in this section.

7.5 Extension to general cone constraints

The theory presented in the previous sections can be extended to solve the following general constrained convex optimization problem:

$$f^\star := \min_x \{f(x) : Ax - b \in \mathcal{K},\ x \in \mathcal{X}\}, \tag{75}$$

where f, $\mathcal{X}$, $\mathcal{K}$, A and b are defined as in (1).

If $\mathcal{K}$ is bounded, then a simple way to process (75) is using a slack variable $r \in \mathcal{K}$ such that $r := Ax - b$ and $z := (x, r)$ as a new variable. Then we can transform (75) into (1) with respect to the new variable z. The primal subproblem corresponding to r is defined as $\min\{\langle -y, r\rangle : r \in \mathcal{K}\}$, which is equivalent to the support function of $\mathcal{K}$, i.e., $s_{\mathcal{K}}(y) := \max\{\langle y, r\rangle : r \in \mathcal{K}\}$. Consequently, the dual function becomes $\tilde{g}(y) := g(y) - s_{\mathcal{K}}(y)$, where $g(y) := \min\{f(x) + \langle Ax - b, y\rangle : x \in \mathcal{X}\}$. Now, we can apply the algorithms presented in the previous sections to obtain an approximate solution $\bar{z}^k := (\bar{x}^k, \bar{r}^k)$ with a convergence guarantee on: $f(\bar{x}^k) - f^\star$, $\|A\bar{x}^k - \bar{r}^k - b\|$, $\bar{x}^k \in \mathcal{X}$ and $\bar{r}^k \in \mathcal{K}$ as in Theorems 1, 2 or 3.

If $\mathcal{K}$ is a cone (e.g., $\mathcal{K} := \mathbb{R}^m_+$, $\mathcal{K}$ is a second order cone $\mathcal{L}^m_+$, or $\mathcal{K}$ 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_β in (28) by the following one

$$\hat{f}_\beta(x) := f(x) + \max\left\{\langle Ax - b, y\rangle - (\beta/2)\|y\|^2 : y \in -\mathcal{K}^*\right\}, \tag{76}$$

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

In principle, we can apply one of three previous schemes to solve (75). Let us demonstrate the (ADSGARD) for this case. Since $\mathcal{K}$ is a cone, we remain using the original scheme (ADSGARD) with the following changes:

$$\begin{cases} y^*_{\beta_k}(\bar{x}^k) := \text{proj}_{-\mathcal{K}^*}\left(\beta_k^{-1}(A\bar{x}^k - b)\right), \\ \bar{y}^{k+1} \quad\quad := \text{proj}_{-\mathcal{K}^*}\left(\hat{y}^k + \frac{\gamma_{k+1}}{\bar{L}_g}\left(Ax^*_{\gamma_{k+1}}(\hat{y}^k) - b\right)\right), \end{cases}$$

where $\mathrm{proj}_{\mathcal{S}}$ is the projection onto the convex set $\mathcal{S}$. In this case, we remain having the convergence guarantee as in Theorem 2 for the objective residual $f(\bar{x}^k)-f^\star$ and the primal feasibility gap $\mathrm{dist}\left(A\bar{x}^k-b,\mathcal{K}\right)$. We note that if $\mathcal{K}$ is a self-dual conic cone, then $\mathcal{K}^*=\mathcal{K}$. Hence, $y^*_{\beta_k}(\bar{x}^k)$ and $\bar{y}^{k+1}$ can often be computed efficiently or in closed form.

A Appendix: The proof of theoretical results

This section provides the full proof of Lemmas and Theorems in the main text.

A.1 Two technical results

A.1.1 The proximal-gradient descent lemma

The following lemma has a similar proof as [3, Lemma 2.3], but using [33, (2.1.7)].

Lemma 9 *Let f and g be two proper, closed and convex functions, and $f\in\mathcal{F}_L^{1,1}$. Suppose that we apply the following proximal-gradient step with $\rho L_f\in(0,2)$*

$$\bar{x}^{k+1}:=prox_{\rho g}\left(\hat{x}^k-\rho\nabla f(\hat{x}^k)\right)$$

to solve the composite convex minimization problem

$$\min_{x\in\mathbb{R}^n}\left\{F(x):=f(x)+g(x)\right\}.$$

Then, the following estimate holds for any $x\in\mathrm{dom}\,(F)$

$$\begin{aligned}F(x)\geq\hat{\ell}_k(x)\geq F(x^{k+1})+\frac{1}{\rho}\langle\hat{x}^k-x^{k+1},x-\hat{x}^k\rangle+\left(\frac{1}{\rho}-\frac{L_f}{2}\right)\|\hat{x}^k-x^{k+1}\|^2\\+\frac{1}{2L_f}\|\nabla f(x)-\nabla f(\hat{x}^k)\|^2,\end{aligned}$$

where $\hat{\ell}_k(x):=f(\hat{x}^k)+\langle\nabla f(\hat{x}^k),x-\hat{x}^k\rangle+\frac{1}{2L_f}\|\nabla f(x)-\nabla f(\hat{x}^k)\|^2$.

A.1.2 The proof of Lemma 2: Properties of g_γ and f_β

For $i\in\mathcal{I}_1$, the proof can be found in [35]. We only prove for $i\notin\mathcal{I}_1$. For fixed $i\notin\mathcal{I}_1$, we consider $\mathcal{U}_i:=\{u\in\mathbb{R}^m:u=A_ix-b_i,\ x_i\in\mathcal{X}_i\}$ and the function

$$\hat{f}_i(u):=\inf_{x_i}\left\{f_i(x_i):u=A_ix_i-b_i,\ x_i\in\mathcal{X}_i\right\}.$$

Under Assumption A.1, it is clear that $\mathcal{U}_i$ is nonempty and convex in $\mathbb{R}^m$. The function $\hat{f}_i$ is also proper, closed and convex in $\mathbb{R}^m$, see [11, Example 3.17].

We have $p_i(A_ix_i)=p_i(u+b_i)\geq\frac{1}{2}\|A_i(x_i-\bar{x}_i^c)\|^2=\frac{1}{2}\|u-\bar{u}^c\|^2$, where $\bar{u}^c:=A_i\bar{x}_i^c-b_i$. Hence, $p_i(\cdot+b_i)$ is strongly convex on $\mathcal{U}_i$ with the convexity parameter $\mu_{p_i}:=1>0$. By definition of g_γ^i in (26), we can easily express g_γ^i as

$$g_\gamma^i(\bar{y}):=\min_{u\in\mathcal{U}_i}\left\{\hat{f}_i(u)+\langle\bar{y},u\rangle+\gamma p_i(u+b_i)\right\}.\tag{77}$$

This function is well-defined due to the strong convexity of p_i. Moreover, it is concave and smooth on $\mathbb{R}^m$. Its gradient ∇g_γ^i is given by $\nabla g_\gamma^i(y)=u_\gamma^*(y)=A_ix_{\gamma,i}^*(y)-b_i$, where $u_\gamma^*(y)$ is the unique solution of (77). The Lipschitz continuity of ∇g_γ^i with the Lipschitz constant $L_g^\gamma:=\gamma^{-1}\bar{L}_i\equiv\gamma^{-1}$ can be proved as in [35, Theorem 1]. Then, the inequalities (30) follows from this property as a direct consequence due to [33, Theorem 2.1.5].

To prove the second part of Lemma 2, it is sufficient to prove for $i\notin\mathcal{I}_1$. Otherwise, we substitute A_i by the identity matrix $\mathbb{I}_i$. We first define $\varphi_{\bar{y}}^i(u,\gamma):=$

$\hat{f}_i(u) + \langle \bar{y}, u\rangle + \gamma p_i(u + b_i)$. The function $\varphi^i_{\bar{y}}$ is strongly convex in u and linear in γ. Hence, $g^i_\gamma(\bar{y})$ defined by (77) is concave and smooth w.r.t. $\gamma > 0$. Moreover, $\frac{dg^i_\gamma(\bar{y})}{d\gamma}|_{\gamma=\bar{\gamma}} = p_i(u^*_{\bar{\gamma}}(\bar{y}) + b_i) \geq 0$, which shows that $g^i_\gamma(\bar{y})$ is nondecreasing, where $u^*_{\bar{\gamma}}(\bar{y})$ is the solution of (77). Using the concavity of $g^i_\gamma(\bar{y})$ w.r.t. γ, we obtain $g^i_\gamma(\bar{y}) \leq g^i_{\bar{\gamma}}(\bar{y}) + (\gamma - \bar{\gamma})p_i(u^*_{\bar{\gamma}}(\bar{y}) + b_i)$. Substituting the relation $u^*_\gamma(y) = A_i x^*_{\gamma,i}(y) - b_i$ into the last estimate, we obtain (32).

We note that since $g_\gamma = \sum_{i=1}^N g^i_\gamma$, the properties of g_γ follow from the ones of g^i_γ. We finally prove the properties of $\bar{p}_\beta$ and f_β. Based on $f_\beta = f + \bar{p}_\beta$ and $\bar{p}_\gamma(x) = \max\{\langle Ax - b, y\rangle - \beta p_y(y) : y \in \mathbb{R}^m\}$, the convexity and smoothness of $\bar{p}_\beta$ follows from [35]. In addition, $\nabla\bar{p}_\beta$ is Lipschitz continuous [35] with the Lipschitz constant $L_{\bar{p}_\beta} := \beta^{-1}\|A\|^2$. Moreover, since $\bar{p}_\beta(x) = \beta p^*_y(\beta^{-1}(A(\cdot) - b))$, the first inequality of (33) follows from the Lipschitz gradient continuity of p^*_y, while the second one is a consequence of (32) by substituting g^i_γ by $-\bar{p}_\beta$. □

A.1.3 The proof of Lemma 3: Key bounds for approximate solutions

Under Assumption A.1, any $(x^\star, y^\star) \in \mathcal{W}^\star$ is a saddle point of the Lagrange function $\mathcal{L}(x, y) := f(x) + \langle Ax - b, y\rangle$, i.e., $\mathcal{L}(x^\star, y) \leq \mathcal{L}(x^\star, y^\star) \leq \mathcal{L}(x, y^\star)$ for all $x \in \mathcal{X}$ and $y \in \mathbb{R}^m$. It leads to $g(y) \leq g^\star = f^\star \leq f(x) + \langle y^\star, Ax - b\rangle$, and hence

$$f(x) - g(y) \geq f(x) - f^\star \geq \langle b - Ax, y^\star\rangle \geq -\|y^\star\|_*\|Ax - b\|, \tag{78}$$

for all $(x, y) \in \mathcal{W}$, which proves (35). By the definition (10) of g and (23) of g_γ, using (31) we have

$$g_{\gamma_k}(\bar{y}^k) - \gamma_k D_{\mathcal{X}} \leq g(\bar{y}^k) \leq g_{\gamma_k}(\bar{y}^k). \tag{79}$$

Combining (78), (79), and the definition of f_β in (28) we obtain

$$\begin{aligned} -\|y^\star\|_*\|A\bar{x}^k - b\| \leq f(\bar{x}^k) - f^\star &\overset{(78)}{\leq} f(\bar{x}^k) - g(\bar{y}^k) \\ &\overset{(79)+(23)}{\leq} f_{\beta_k}(\bar{x}^k) - g_{\gamma_k}(\bar{y}^k) + \gamma_k D_{\mathcal{X}} - \beta_k p^*_y(\beta_k^{-1}(A\bar{x}^k - b)) \\ &\overset{(24)}{=} G_{\gamma_k\beta_k}(\bar{w}^k) + \gamma_k D_{\mathcal{X}} - \beta_k p^*_y\big(\beta_k^{-1}(A\bar{x}^k - b)\big). \end{aligned} \tag{80}$$

Since $p^*_y(\beta_k^{-1}(A\bar{x}^k - b)) \geq p^*_y(\bar{s}^c) = -p_y(0)$, the last inequality leads to

$$-\|y^\star\|_*\|A\bar{x}^k - b\| \leq f(\bar{x}^k) - f^\star \leq f(\bar{x}^k) - g(\bar{y}^k) \leq S_k.$$

which is indeed the first estimate of (36), where $S_k := G_{\gamma_k\beta_k}(\bar{w}^k) + \gamma_k D_{\mathcal{X}} + \beta_k p_y(0)$.

Next, by the $\frac{1}{L_{p_y}}$-strong convexity of p^*_y and $\nabla p^*_y(\bar{s}^c) = 0$, we have

$$\begin{aligned} \beta_k p^*\big(\beta_k^{-1}(A\bar{x}^k - b)\big) &\geq \beta_k p^*_y(\bar{s}^c) + \frac{\beta_k}{2L_{p_y}}\|\beta_k^{-1}(A\bar{x}^k - b) - \bar{s}^c\|^2 \\ &= -\beta_k p_y(0) + \frac{1}{2L_{p_y}\beta_k}\|A\bar{x}^k - b\|^2 - \frac{1}{L_{p_y}}\langle \bar{s}^c, A\bar{x}^k - b\rangle + \frac{\beta_k}{2L_{p_y}}\|\bar{s}^c\|^2. \end{aligned}$$

Combining this inequality, (78) and (80), we obtain

$$\begin{aligned} \langle y^\star, b - A\bar{x}^k\rangle \leq\ & G_{\gamma_k\beta_k}(\bar{w}^k) + \gamma_k D_{\mathcal{X}} + \beta_k p_y(0) \\ & - \frac{1}{2L_{p_y}\beta_k}\|A\bar{x}^k - b\|^2 - \frac{1}{L_{p_y}}\langle \bar{s}^c, A\bar{x}^k - b\rangle - \frac{\beta_k}{2L_{p_y}}\|\bar{s}^c\|^2. \end{aligned}$$

Rearranging this expression and using the Cauchy-Schwarz inequality, we obtain

$$-\|y^\star - L_{p_y}^{-1}\bar{s}^c\|_*\|A\bar{x}^k - b\| \le S_k - (2L_{b_y}\beta_k)^{-1}\|A\bar{x}^k - b\|^2 - \beta_k(2L_{p_y})^{-1}\|\bar{s}^c\|^2,$$

which leads to

$$\|A\bar{x}^k - b\|^2 - 2\beta_k\|L_{b_y}y^\star - \bar{s}^c\|_*\|A\bar{x}^k - b\| - \big(2L_{p_y}\beta_k S_k - \beta_k^2\|\bar{s}^c\|^2\big) \le 0.$$

Let $t := \|A\bar{x}^k - b\|$. We obtain from the last inequality the inequation $t^2 - 2\beta_k\|L_{p_y}y^\star - \bar{s}^c\|_* t - \big(2L_{p_y}\beta_k S_k - \beta_k^2\|\bar{s}^c\|^2\big) \le 0$. This inequation of t leads to

$$t := \|A\bar{x}^k - b\| \le \beta_k\Big[\|L_{p_y}y^\star - \bar{s}^c\|_* + \big[\|L_{p_y}y^\star - \bar{s}^c\|_*^2 + \big(2L_{p_y}\beta_k^{-1}S_k - \|\bar{s}^c\|^2\big)\big]^{1/2}\Big],$$

which is the second estimate of (36), provided that $\|L_{p_y}y^\star - \bar{s}^c\|_*^2 + 2L_{p_y}\beta_k^{-1}S_k - \|\bar{s}^c\|^2 \ge 0$.

If we choose $p_y(y) := (1/2)\|y\|_2^2$, then $\nabla p_y(y) = y$, $L_{p_y} = 1$ and $\bar{s}_c = 0$. In this case, the right-hand side S_k of (36) reduces to $S_k = G_{\gamma_k\beta_k}(\bar{w}^k) + \gamma_k D_{\mathcal{X}}$, which is in the first estimate of (37). The right-hand side of the second estimate of (36) reduces to $\beta_k\Big[\|y^\star\|_2 + \sqrt{\|y^\star\|_2^2 + 2\beta_k^{-1}S_k}\Big] \le 2\beta_k\|y^\star\|_2 + \sqrt{2\beta_k S_k}$, which is in the second estimate of (37). □

A.2 The convergence analysis of the ASGARD method

In this appendix, we provide the full proof of Lemmas 4 and 5, and Theorem 1.

A.2.1 The proof of Lemma 4: Maintaining the gap reduction condition

By (28), we have $\bar{p}_\beta(x) = \beta p_y^*(\beta^{-1}(Ax - b))$. Using the second estimate in (33) with $\hat{x} := \hat{x}^k$, we get

$$\bar{p}_\beta(x) \ge \bar{p}_\beta(\hat{x}^k) + \langle\nabla\bar{p}_\beta(\hat{x}^k), x - \hat{x}^k\rangle + (2\beta L_{p_y})^{-1}\|A(x - \hat{x}^k)\|^2. \tag{81}$$

By Lemma 9, $\bar{x}^{k+1}$ is obtained from $\hat{x}^k$ by applying one proximal-gradient step from (ASGARD) to minimize $f_\beta := f + \bar{p}_\beta$ with $\bar{p}_\beta \in \mathcal{F}_L^{1,1}$ and $L_{\bar{p}_\beta} := \frac{\bar{L}_A}{\beta}$, we have

$$\begin{aligned} f_\beta(\bar{x}^{k+1}) &\le \hat{\ell}_{f_\beta}^k(x) + \frac{1}{\rho}\langle\hat{x}^k - \bar{x}^{k+1}, \hat{x}^k - x\rangle - \left(\frac{1}{\rho} - \frac{\bar{L}_A}{2\beta}\right)\|\hat{x}^k - \bar{x}^{k+1}\|^2 - \frac{1}{2\beta L_{p_y}}\|A(x - \hat{x}^k)\|^2 \\ &\le f_\beta(x) + \frac{1}{\rho}\langle\hat{x}^k - \bar{x}^{k+1}, \hat{x}^k - x\rangle - \left(\frac{1}{\rho} - \frac{\bar{L}_A}{2\beta}\right)\|\hat{x}^k - \bar{x}^{k+1}\|^2 - \frac{1}{2\beta L_{p_y}}\|A(x - \hat{x}^k)\|^2, \end{aligned} \tag{82}$$

where $\hat{\ell}_{f_\beta}^k(x) := f(x) + \bar{p}_\beta(\hat{x}^k) + \langle\nabla\bar{p}_\beta(\hat{x}^k), x - \hat{x}^k\rangle + \frac{1}{2\beta\bar{L}_{p_y}}\|A(x - \hat{x}^k)\|^2$ for $x \in \mathcal{X}$.

Alternatively, since $\bar{y}^{k+1}$ is obtained from $\hat{y}^k$ by applying one the gradient ascent step at line 3 of (ASGARD) to $g_\gamma \in -\mathcal{F}_L^{1,1}$, using again Lemma 9 to get

$$\begin{aligned} -g_\gamma(\bar{y}^{k+1}) &\le -\hat{\ell}_{g_\gamma}^k(y) + \frac{1}{\lambda}\langle\hat{y}^k - \bar{y}^{k+1}, \hat{y}^k - y\rangle - \left(\frac{1}{\lambda} - \frac{\bar{L}_g}{2\gamma}\right)\|\hat{y}^k - \bar{y}^{k+1}\|^2 - \gamma\tilde{r}_k(y) \\ &\le -g_\gamma(y) + \frac{1}{\lambda}\langle\hat{y}^k - \bar{y}^{k+1}, \hat{y}^k - y\rangle - \left(\frac{1}{\lambda} - \frac{\bar{L}_g}{2\gamma}\right)\|\hat{y}^k - \bar{y}^{k+1}\|^2 - \gamma\tilde{r}_k(y), \end{aligned} \tag{83}$$

where $\tilde{r}_k$ and $\hat{\ell}_{g_\gamma}^k$ are defined as

$$\begin{aligned} \tilde{r}_k(y) &:= \sum_{i\in\mathcal{I}_1}\frac{1}{2\bar{L}_{A_i}}\|x_{\gamma,i}^*(y) - x_{\gamma,i}^*(\hat{y}^k)\|^2 + \sum_{i\notin\mathcal{I}_1}\frac{1}{2}\|A_i(x_{\gamma,i}^*(y) - x_{\gamma,i}^*(\hat{y}^k))\|^2, \\ \hat{\ell}_{g_\gamma}^k(y) &:= g_\gamma(\hat{y}^k) + \langle\nabla g_\gamma(\hat{y}^k), y - \hat{y}^k\rangle - \gamma\tilde{r}_k(y), \\ g_\gamma(\cdot) &= \textstyle\sum_{i=1}^N g_\gamma^i(\cdot), \\ \nabla g_\gamma^i(y) &= A_i x_{\gamma,i}^*(y) - b_i, \\ \bar{L}_g &= \textstyle\sum_{i=1}^N \bar{L}_i. \end{aligned}$$

Using the second inequality of (82) with $x := \bar{x}^k$ and of (83) with $y := \hat{y}^k$, respectively, then summing up the results and using $G_{\gamma\beta} = f_\beta - g_\gamma$ to obtain

$$\begin{aligned} G_{\gamma\beta}(\bar{w}^{k+1}) \le G_{\gamma\beta}(\bar{w}^k) + \frac{1}{\rho}\langle \hat{x}^k - \bar{x}^{k+1}, \hat{x}^k - \bar{x}^k\rangle - \left(\frac{1}{\rho} - \frac{\bar{L}_A}{2\beta}\right)\|\hat{x}^k - \bar{x}^{k+1}\|^2 \\ + \frac{1}{\lambda}\langle \hat{y}^k - \bar{y}^{k+1}, \hat{y}^k - \bar{y}^k\rangle - \left(\frac{1}{\lambda} - \frac{\bar{L}_g}{2\gamma}\right)\|\hat{y}^k - \bar{y}^{k+1}\|^2 - \left(\beta\tilde{q}_k^* + \gamma\tilde{r}_k^*\right), \end{aligned} \tag{84}$$

where $\tilde{q}_k^*$ and $\tilde{r}_k^*$ are defined respectively by

$$\begin{cases} \tilde{q}_k^* := \frac{1}{2\beta^2\bar{L}_{py}}\|A(\bar{x}^k - \hat{x}^k)\|^2, \\ \tilde{r}_k^* := \sum\limits_{i\in\mathcal{I}_1}\frac{1}{2\bar{L}_{A_i}}\|x_{\gamma,i}^*(\bar{y}^k) - x_{\gamma,i}^*(\hat{y}^k)\|^2 + \sum\limits_{i\notin\mathcal{I}_1}\frac{1}{2}\|A_i(x_{\gamma,i}^*(\bar{y}^k) - x_{\gamma,i}^*(\hat{y}^k))\|^2. \end{cases} \tag{85}$$

Similarly, summing up the first inequality of (82) and (83), we get

$$\begin{aligned} G_{\gamma\beta}(\bar{w}^{k+1}) \le \hat{H}_{\gamma\beta}(w) + \frac{1}{\rho}\langle \hat{x}^k - \bar{x}^{k+1}, \hat{x}^k - x\rangle - \left(\frac{1}{\rho} - \frac{\bar{L}_A}{2\beta}\right)\|\hat{x}^k - \bar{x}^{k+1}\|^2 \\ + \frac{1}{\lambda}\langle \hat{y}^k - \bar{y}^{k+1}, \hat{y}^k - y\rangle - \left(\frac{1}{\lambda} - \frac{\bar{L}_g}{2\gamma}\right)\|\hat{y}^k - \bar{y}^{k+1}\|^2, \end{aligned} \tag{86}$$

for any $w \in \mathcal{W}$, where $\hat{H}_{\gamma\beta}(\cdot)$ is defined as

$$\hat{H}_{\gamma\beta}(w) := \left[f(x) + \bar{p}_\beta(\hat{x}^k) + \langle \nabla\bar{p}_\beta(\hat{x}^k), x - \hat{x}^k\rangle\right] - \left[g_\gamma(\hat{y}^k) + \langle \nabla g_\gamma(\hat{y}^k), y - \hat{y}^k\rangle\right]. \tag{87}$$

Next, multiplying (84) by $1-\tau_k$ and (86) by $\tau_k \in (0,1]$ and summing up the results, we obtain

$$\begin{aligned} G_{\gamma\beta}(\bar{w}^{k+1}) \le (1-\tau_k)G_{\gamma\beta}(\bar{w}^k) + \tau_k\hat{H}_{\gamma\beta}(w) + \frac{1}{\rho}\langle \hat{x}^k - \bar{x}^{k+1}, \hat{x}^k - (1-\tau_k)\bar{x}^k - \tau_k x\rangle \\ - \left(\frac{1}{\rho} - \frac{\bar{L}_A}{2\beta}\right)\|\hat{x}^k - \bar{x}^{k+1}\|^2 + \frac{1}{\lambda}\langle \hat{y}^k - \bar{y}^{k+1}, \hat{y}^k - (1-\tau_k)\bar{y}^k - \tau_k y\rangle \\ - \left(\frac{1}{\lambda} - \frac{\bar{L}_g}{2\gamma}\right)\|\hat{y}^k - \bar{y}^{k+1}\|^2 - (1-\tau_k)\left(\beta\tilde{q}_k^* + \gamma\tilde{r}_k^*\right). \end{aligned} \tag{88}$$

Using the first line of (ASGARD), we have $\hat{y}^k - (1-\tau_k)\bar{y}^k = \tau_k\tilde{y}^k$ and $\hat{x}^k - (1-\tau_k)\bar{x}^k = \tau_k\tilde{x}^k$. Hence, we can rearrange (88) as

$$\begin{aligned} G_{\gamma\beta}(\bar{w}^{k+1}) \le (1-\tau_k)G_{\gamma\beta}(\bar{w}^k) + \tau_k\hat{H}_{\gamma\beta}(w) + \frac{\tau_k^2}{2u_\rho}\left[\|\tilde{x}^k - x\|^2 - \|\tilde{x}^{k+1} - x\|^2\right] + \\ + \frac{\tau_k^2}{2v_\lambda}\left[\|\tilde{y}^k - y\|^2 - \|\tilde{y}^{k+1} - y\|^2\right] - (1-\tau_k)\left(\beta\tilde{q}_k^* + \gamma\tilde{r}_k^*\right), \end{aligned} \tag{89}$$

where

$$\begin{aligned} \tilde{x}^{k+1} &:= \tilde{x}^k - \tfrac{(2-\rho\beta^{-1}\bar{L}_A)}{\tau_k}(\hat{x}^k - \bar{x}^{k+1}), \\ \tilde{y}^{k+1} &:= \tilde{y}^k - \tfrac{(2-\lambda\gamma^{-1}\bar{L}_g)}{\tau_k}(\hat{y}^k - \bar{y}^{k+1}), \\ u_\rho &:= 2\rho - \rho^2\beta^{-1}\bar{L}_A, \\ v_\lambda &:= 2\lambda - \lambda^2\gamma^{-1}\bar{L}_g. \end{aligned}$$

Here, the two first lines are exactly given by the fourth and fifth lines of (ASGARD). Alternatively, using the first estimate in (33) with $\bar{x} := \bar{x}^k$, $\beta := \beta_k$ and $\bar{\beta} := \beta_{k+1}$, and (32) with $\bar{y} := \bar{y}^k$, $\gamma := \gamma_k$ and $\bar{\gamma} := \gamma_{k+1}$, we get

$$\begin{cases} f_{\beta_{k+1}}(\bar{x}^k) & \leq f_{\beta_k}(\bar{x}^k) \quad +(\beta_k - \beta_{k+1})\bar{q}_k^*, \\ -g_{\gamma_{k+1}}(\bar{y}^k) & \leq -g_{\gamma_k}(\bar{y}^k) +(\gamma_k - \gamma_{k+1})\bar{r}_k^*, \end{cases} \tag{90}$$

where the quantities $\bar{q}_k^*$ and $\bar{r}_k^*$ are defined as

$$\bar{q}_k^* := p_y(y^*_{\beta_{k+1}}(\bar{x}^k)) \;\text{ and }\; \bar{r}_k^* := \sum_{i\in\mathcal{I}_1} p_i(x^*_{\gamma_{k+1}}(\bar{y}^k)) + \sum_{i\notin\mathcal{I}_1} p_i(A_i x^*_{\gamma_{k+1}}(\bar{y}^k)). \tag{91}$$

Now, we consider $\hat{H}_{\gamma\beta}(w)$ given by (87). Using the definition (28) of f_β and (11) of $\bar{f}$, we have

$$\begin{aligned} \tilde{\ell}_f^k(x) &:= f(x) + \bar{p}_\beta(\hat{x}^k) + \langle \nabla\bar{p}_\beta(\hat{x}^k), x - \hat{x}^k\rangle \\ &= f(x) + \langle A\hat{x}^k - b, y^*_\beta(\hat{x}^k)\rangle + \langle A^T y^*_\beta(\hat{x}^k), x - \hat{x}^k\rangle - \beta p_y(y^*_\beta(\hat{x}^k)) \\ &= f(x) + \langle Ax - b, y^*_\beta(\hat{x}^k)\rangle - \beta p_y(y^*_\beta(\hat{x}^k)) \\ &\leq f(x) + \max\left\{\langle Ax - b, y\rangle : y\in\mathbb{R}^n\right\} - \beta p_y(y^*_\beta(\hat{x}^k)) \\ &\overset{(11)}{=} \bar{f}(x) - \beta p_y(y^*_\beta(\hat{x}^k)). \end{aligned} \tag{92}$$

Alternatively, using the definition (23) of g_γ, and (10) of g, we also have

$$\begin{aligned} \tilde{\ell}_g^k(y) &:= g_\gamma(\hat{y}^k) + \langle \nabla g_\gamma(\hat{y}^k), y - \hat{y}^k\rangle \\ &= f(x^*_\gamma(\hat{y}^k)) + \langle Ax^*_\gamma(\hat{y}^k) - b, y\rangle + \gamma p_x(Ax^*_\gamma(\hat{y}^k)) \\ &\geq \min\left\{f(x) + \langle Ax - b, y\rangle : x\in\mathcal{X}\right\} + \gamma p_x(Ax^*_\gamma(\hat{y}^k)) \\ &= g(y) + \gamma p_x(Ax^*_\gamma(\hat{y}^k)). \end{aligned} \tag{93}$$

Combining (92), (93), and (17) with $G(w) := \bar{f}(x) - g(y)$, we can show that

$$H_{\gamma_{k+1}\beta_{k+1}}(w) \leq G(w) - \left(\beta_{k+1}\hat{q}_k^* + \gamma_{k+1}\hat{r}_k^*\right), \tag{94}$$

where the two quantities $\hat{q}_k^*$ and $\hat{r}_k^*$ are given by

$$\hat{q}_k^* := p_y(y^*_{\beta_{k+1}}(\hat{x}^k)) \quad\text{and}\quad \hat{r}_k^* := \sum_{i\in\mathcal{I}_1} p_i(x^*_{\gamma_{k+1}}(\hat{y}^k)) + \sum_{i\notin\mathcal{I}_1} p_i(A_i x^*_{\gamma_{k+1}}(\hat{y}^k)). \tag{95}$$

Using (89) with $\gamma := \gamma_{k+1}$, $\beta := \beta_{k+1}$ and $w = w^\star \in \mathcal{W}^\star$, and then combining the result with both (90) and (94) we achieve

$$\begin{aligned} G_{k+1}(\bar{w}^{k+1}) \leq (1-\tau_k)G_k(\bar{w}^k) + \tau_k G(w^\star) + \frac{\tau_k^2}{2u_{k+1}}\left[\|\tilde{y}^k - y^\star\|^2 - \|\tilde{y}^{k+1} - y^\star\|^2\right] \\ + \frac{\tau_k^2}{2v_{k+1}}\left[\|\tilde{x}^k - x^\star\|^2 - \|\tilde{x}^{k+1} - x^\star\|^2\right] - \mathcal{R}_k, \end{aligned} \tag{96}$$

where $G_k := G_{\gamma_k\beta_k}$, $u_{k+1} := u_\rho = 2\rho_{k+1} - \rho_{k+1}^2\beta_{k+1}^{-1}\bar{L}_A$, $v_{k+1} := v_\lambda = 2\lambda_{k+1} - \lambda_{k+1}^2\gamma_{k+1}^{-1}\bar{L}_g$, and the last term $\mathcal{R}_k$ is given as

$$\begin{aligned} \mathcal{R}_k := \left[\tau_k\beta_{k+1}\hat{q}_k^* + (1-\tau_k)\beta_{k+1}\tilde{q}_k^* - (1-\tau_k)(\beta_k - \beta_{k+1})\bar{q}_k^*\right]_{[1]} \\ + \left[\tau_k\gamma_{k+1}\hat{r}_k^* + (1-\tau_k)\gamma_{k+1}\tilde{r}_k^* - (1-\tau_k)(\gamma_k - \gamma_{k+1})\bar{r}_k^*\right]_{[2]}. \end{aligned} \tag{97}$$

The next step is to lower bound $\mathcal{R}_k$. Using the strong convexity of p_y and p_y^*, respectively, and $\bar{y}^c := \nabla p_y^*(\mathbf{0}^m)$, we have

$$\hat{q}_k^* = p_y(y^*_{\beta_{k+1}}(\hat{x}^k)) \geq \frac{1}{2}\|y^*_{\beta_{k+1}}(\hat{x}^k) - \bar{y}^c\|^2 = \frac{1}{2}\|\nabla p_y^*(\beta_{k+1}^{-1}(A\hat{x}^k - b)) - \nabla p_y^*(\mathbf{0}^m)\|^2$$
$$\geq \frac{1}{2\beta_{k+1}^2 L_{p_y}}\|A\hat{x}^k - b\|^2. \tag{98}$$

Alternatively, using the Lipschitz continuity of ∇p_y and ∇p_y^*, and $p_y(\bar{y}^c) = 0$, we can also derive

$$\bar{q}_k^* := p_y(\bar{y}^*_{\beta_{k+1}}(\bar{x}^k)) \leq \frac{L_{p_y}}{2}\|\bar{y}^*_{\beta_{k+1}}(\bar{x}^k) - \bar{y}^c\|^2 = \frac{L_{p_y}}{2}\|\nabla p_y^*(\beta_{k+1}^{-1}(A\bar{x}^k - b)) - \nabla p_y^*(\mathbf{0}^m)\|^2$$
$$\leq \frac{L_{p_y}}{2\beta_{k+1}^2}\|A\bar{x}^k - b\|^2. \tag{99}$$

Using the definition (85), (91) and (95) of $\tilde{q}_k^*$, $\bar{q}_k^*$ and $\hat{q}_k^*$, respectively, and the two estimates (98) and (99), the first term $[\cdot]_{[1]}$ of (97) can be lower bounded as

$$\begin{aligned}
[\cdot]_{[1]} &= \beta_{k+1}\left[\tau_k\hat{q}_k^* + (1-\tau_k)\tilde{q}_k^* - (1-\tau_k)(\beta_{k+1}^{-1}\beta_k - 1)\bar{q}_k^*\right] \\
&\geq \frac{1}{2L_{p_y}\beta_{k+1}}\left[\tau_k\|\hat{a}_k\|^2 + (1-\tau_k)\|\bar{a}_k - \hat{a}_k\|^2 - (1-\tau_k)\left(\beta_{k+1}^{-1}\beta_k - 1\right)L_{p_y}^2\|\bar{a}_k\|^2\right] \\
&= \frac{1}{2L_{p_y}\beta_{k+1}}\left[\|\hat{a}_k - (1-\tau_k)\bar{a}_k\|^2 + (1-\tau_k)\left[\tau_k - \left(\beta_{k+1}^{-1}\beta_k - 1\right)L_{p_y}^2\right]\|\bar{a}_k\|^2\right],
\end{aligned}$$

where $\hat{a}_k := A\hat{x}^k - b$ and $\bar{a}_k := A\bar{x}^k - b$. This expression shows that $[\cdot]_{[1]} \geq 0$ if

$$\left(1 + \frac{\tau_k}{L_{p_y}^2}\right)\beta_{k+1} \geq \beta_k, \tag{100}$$

which is the first condition in (40).

Next, using the property of p_i in Assumption A.2, it is easy to show that

$$\begin{cases}
\frac{1}{2\bar{L}_{A_i}}\|A_i(x_i - \bar{x}_i^c)\|^2 \leq p_i(x_i) \leq \frac{L_{p_i}}{2\sigma_i^2}\|A_i(x_i - \bar{x}_i^c)\|^2, & \forall i \in \mathcal{I}_1, \\
\frac{1}{2}\|A_i x_i - \bar{u}_i^c\| \leq p_i(A_i x_i) \leq \frac{L_{p_i}}{2}\|A_i x_i - \bar{u}_i^c\|^2, & \forall i \notin \mathcal{I}_1.
\end{cases}$$

Let us abbreviate $\hat{x}^k_{*,i} := x^*_{\gamma_{k+1},i}(\hat{y}^k)$, $\bar{x}^k_{*,i} := x^*_{\gamma_{k+1},i}(\bar{y}^k)$. Then, we define $\bar{e}_i^k := A_i(\bar{x}^k_{*,i} - \bar{x}_i^c)$ and $\hat{e}_i^k := A_i(\hat{x}^k_{*,i} - \bar{x}_i^c)$ for $i \in \mathcal{I}_1$, and $\bar{e}_i^k := A_i\bar{x}^k_{*,i} - \bar{u}_i^c$ and $\hat{e}_i^k := A_i\hat{x}^k_{*,i} - \bar{u}_i^c$ for $i \in \bar{\mathcal{I}}_1$. We can estimate the second term $[\cdot]_{[2]}$ of (97) using the definition (85), (91) and (95) of $\tilde{r}_k^*$, $\bar{r}_k^*$ and $\hat{r}_k^*$, respectively, and the two last inequalities as follows:

$$\begin{aligned}
[\cdot]_{[2]} &:= \gamma_{k+1}\left[\tau_k\hat{r}_k^* + (1-\tau_k)\tilde{r}_k^* - (1-\tau_k)\left(\gamma_{k+1}^{-1}\gamma_k - 1\right)\bar{r}_k^*\right] \\
&\geq \frac{\gamma_{k+1}}{2}\sum_{i\in\mathcal{I}_1}\left[\frac{\tau_k}{\bar{L}_{A_i}}\|\hat{e}_i\|^2 + \frac{(1-\tau_k)}{\bar{L}_{A_i}}\|\bar{e}_i^k - \hat{e}_i^k\|^2 - \frac{L_{p_i}(1-\tau_k)}{\sigma_i^2}\left(\frac{\gamma_k}{\gamma_{k+1}} - 1\right)\|\bar{e}_k^i\|^2\right] \\
&+ \frac{\gamma_{k+1}}{2}\sum_{i\notin\mathcal{I}_1}\left[\tau_k\|\hat{e}_i\|^2 + (1-\tau_k)\|\bar{e}_i^k - \hat{e}_i^k\|^2 - L_{p_i}(1-\tau_k)\left(\frac{\gamma_k}{\gamma_{k+1}} - 1\right)\|\bar{e}_k^i\|^2\right] \\
&= \frac{\gamma_{k+1}}{2}\sum_{i\in\mathcal{I}_1}\frac{1}{\bar{L}_{A_i}}\left[\|\hat{e}_k - (1-\tau_k)\bar{e}_k\|^2 + (1-\tau_k)\left[\tau_k - \left(\frac{\gamma_k}{\gamma_{k+1}} - 1\right)\frac{L_{p_i}\bar{L}_{A_i}}{\sigma_i^2}\right]\|\bar{e}_k^i\|^2\right] \\
&+ \frac{\gamma_{k+1}}{2}\sum_{i\notin\mathcal{I}_1}\left[\|\hat{e}_k - (1-\tau_k)\bar{e}_k\|^2 + (1-\tau_k)\left[\tau_k - \left(\frac{\gamma_k}{\gamma_{k+1}} - 1\right)L_{p_i}\right]\|\bar{e}_k^i\|^2\right].
\end{aligned}$$

This expression shows that $[\cdot]_{[2]} \geq 0$ if

$$\begin{cases} \left(1 + \frac{\sigma_i^2}{L_{p_i}\bar{L}_{A_i}}\tau_k\right)\gamma_{k+1} \geq \gamma_k, & \forall\ i \in \mathcal{I}_1, \\ \left(1 + \frac{\tau_k}{L_{p_i}}\right)\gamma_{k+1} \geq \gamma_k, & \forall\ i \notin \mathcal{I}_1. \end{cases} \tag{101}$$

Consequently, if both conditions (100) and (101) hold, then $\mathcal{R}_k \geq 0$. We summarize that the condition (101) holds if $\gamma_{k+1}\left(1 + \frac{\tau_k}{\hat{c}_1}\right) \geq \gamma_k$, which is the second condition of (40), where

$$\hat{c}_1 := \max\left\{\max_{i\in\mathcal{I}_1}\left\{\frac{L_{p_i}\bar{L}_{A_i}}{\sigma_i^2}\right\}, \max_{i\notin\mathcal{I}_1}\left\{L_{p_i}\right\}\right\}.$$

This quantity $\hat{c}_1$ is indeed (39).

Next, let us choose $u_{k+1} = 2\rho_{k+1} - \rho_{k+1}^2\beta_{k+1}^{-1}\bar{L}_A = 2\lambda_{k+1} - \lambda_{k+1}^2\gamma_{k+1}^{-1}\bar{L}_g = v_{k+1}$ as given in Lemma 4. We obtain from (96) that

$$G_{k+1}(\bar{w}^{k+1}) \leq (1-\tau_k)G_k(\bar{w}^k) + \tau_k G(w^\star) - \mathcal{R}_k + \frac{\tau_k^2}{u_{k+1}}D_k,$$

where $D_k := \frac{1}{2}\left[\|\tilde{y}^k - y^\star\|^2 - \|\tilde{y}^{k+1} - y^\star\|^2\right] + \frac{1}{2}\left[\|\tilde{x}^k - x^\star\|^2 - \|\tilde{x}^{k+1} - x^\star\|^2\right]$. This inequality is in fact (41) since $G(w^\star) = 0$ and $\mathcal{R}_k \geq 0$.

Finally, let $E_k^* := \frac{1}{2}\|\tilde{y}^k - y^\star\|^2 + \frac{1}{2}\|\tilde{x}^k - x^\star\|^2$. Then, on the one hand, by dividing both sides of (41) by $\tau_k^{-2}u_{k+1}$, we have

$$\frac{u_{k+1}}{\tau_k^2}G_{k+1}(\bar{w}^{k+1}) + E_{k+1}^* \leq \frac{(1-\tau_k)u_{k+1}}{\tau_k^2}G_k(\bar{w}^k) + E_k^*,$$

One the other hand, using the condition $(1-\tau_k)\tau_{k-1}^2u_{k+1} = \tau_k^2u_k$, we have $\frac{(1-\tau_{k+1})u_{k+2}}{\tau_{k+1}^2} = \frac{u_{k+1}}{\tau_k^2}$. Using this relation into the last inequality, we get

$$\frac{(1-\tau_{k+1})u_{k+2}}{\tau_{k+1}^2}G_{k+1}(\bar{w}^{k+1}) + E_{k+1}^* \leq \frac{u_{k+1}}{\tau_k^2}G_{k+1}(\bar{w}^{k+1}) + E_{k+1}^* \leq \frac{(1-\tau_k)u_{k+1}}{\tau_k^2}G_k(\bar{w}^k) + E_k^*.$$

By induction, we obtain (42) from this inequality. □

A.2.2 The proof of Lemma 5: The update rule for parameters

First, since $\hat{c}_1 = 1$ and $L_{p_y} = 1$, both conditions in (40) lead to $\beta_{k+1}(1+\tau_k) \geq \beta_k$ and $\gamma_{k+1}(1+\tau_k) \geq \gamma_k$. Now, assume that we choose $\tau_k = \frac{1}{k+1}$. Then $\tau_0 = 1$. This is the first update in (43).

Second, from the condition $(1-\tau_k)\tau_{k-1}^2u_{k+1} = \tau_k^2u_k$, we have

$$u_{k+1} = \frac{\tau_k^2}{\tau_{k-1}^2(1-\tau_k)}u_k = \frac{k}{k+1}u_k.$$

By induction, we obtain $u_{k+1} = \frac{u_1}{k+1}$, which is the last rule of (43).

Third, let us assume that β_k is updated by $\beta_{k+1} := \frac{\beta_k}{1+\tau_k} = \frac{(k+1)\beta_k}{(k+2)}$. By induction, we have $\beta_{k+1} = \frac{2\beta_1}{k+2}$. Similarly, if we update $\gamma_{k+1} := \frac{\gamma_k}{1+\tau_k}$, then we also have $\gamma_{k+1} = \frac{2\gamma_1}{k+2}$. There are two update rules in (43).

Fourth, we need to find ρ_k from the condition (44). We have $2\rho_k - \rho_k^2\beta_k^{-1}\bar{L}_A = 2\rho_k - (k+1)\rho_k^2\frac{\bar{L}_A}{2\beta_1} = u_k = \frac{u_1}{k}$. This leads to a quadratic equation in ρ_k as

$$(k+1)\frac{\bar{L}_A}{2\beta_1}\rho_k^2 - 2\rho_k + \frac{u_1}{k} = 0.$$

This equation has solution if $\left(\frac{k+1}{k}\right)\frac{\bar{L}_A u_1}{2\beta_1} \leq 1$. This condition holds for all $k \geq 1$ if $\bar{L}_A u_1 \leq \beta_1$. Under this condition, we obtain two positive solutions $\rho_k = \frac{2\beta_1}{\bar{L}_A(k+1)}\left[1 \pm \sqrt{1 - \frac{\bar{L}_A u_1(k+1)}{2\beta_1 k}}\right]$ of the above quadratic equation. This is indeed the update rule for ρ_k in (44).

Fifth, similar to ρ_k, we can find an update rule for λ_k from the condition (44). We have $2\lambda_k - \lambda_k^2\gamma_k^{-1}\bar{L}_g = v_k = u_k = \frac{u_1}{k}$. With the same argument as before, this quadratic equation in λ_k has solution if $\bar{L}_g u_1 \leq \gamma_1$. Under this condition, it has two positive solutions $\lambda_k = \frac{2\gamma_1}{\bar{L}_g(k+1)}\left[1 \pm \sqrt{1 - \frac{\bar{L}_g u_1(k+1)}{2\gamma_1 k}}\right]$, which is is the update rule for λ_k in (44).

Finally, we note that $\rho_k\beta_k^{-1}\bar{L}_A = 1 \pm \sqrt{1 - \frac{\bar{L}_A u_1(k+1)}{2\beta_1 k}} \in (0,2)$, which implies $\omega_k := 2 - \rho_k\beta_k^{-1}\bar{L}_A > 0$. Similarly, $\lambda_k\gamma_k^{-1}\bar{L}_g = 1 \pm \sqrt{1 - \frac{\bar{L}_g u_1(k+1)}{2\gamma_1 k}} \in (0,2)$, which leads to $\hat{\omega}_k := 2 - \lambda_k\gamma_k^{-1}\bar{L}_g > 0$. □

A.2.3 The proof of Theorem 1: The convergence guarantee of Algorithm 1.
Using $\tau_0 = 1$, $\tau_k = \frac{1}{k+1}$ and $u_{k+1} = \frac{u_1}{k+1} = \frac{\gamma_1}{\bar{L}_g(k+1)}$ into (42) of Lemma 4 we obtain

$$G_k(\bar{w}^k) \leq \frac{\tau_k^2}{(1-\tau_k)u_{k+1}}E_0^* = \frac{R_0(w^\star)^2}{2u_1 k} = \frac{\bar{L}_g R_0(w^\star)^2}{2\gamma_1 k}, \quad (\forall k \geq 1),$$

where $R_0(w^\star)^2 := 2E_0^* \equiv \|\tilde{y}^0 - y^\star\|^2 + \|\tilde{x}^0 - x^\star\|^2 = \|\bar{y}^0 - y^\star\|^2 + \|\bar{x}^0 - x^\star\|^2$ due to the choice $\tilde{x}^0 = \bar{x}^0$ and $\tilde{y}^0 = \bar{y}^0$. Finally, by using this bound, $\beta_k = \frac{2\beta_1}{k+1} = \frac{2\bar{L}_A\gamma_1}{\bar{L}_g(k+1)}$ and $\gamma_k = \frac{2\gamma_1}{k+1}$ from Lemma 5, and Lemma 3, we can easily derive (45) by noting that $\frac{k+1}{k} \leq 2$ for any $k \geq 1$. □

A.3 The convergence analysis of the ADSGARD method

In this appendix, we provide the full proof of Lemmas 6, 7 and 8 for (ADSGARD).

A.3.1 The proof of Lemma 6: Maintaining the gap reduction condition
We abbreviate $G_k := G_{\gamma_k\beta_k}$, $\bar{y}_k^* := y_{\beta_k}^*(\bar{x}^k)$, $\hat{x}_k^* := x_{\gamma_{k+1}}^*(\hat{y}^k)$ and $\bar{x}_k^* := x_{\gamma_{k+1}}^*(\bar{y}^k)$.

Using the definition of G_{k+1} and f_β, the third line $\bar{x}^{k+1} := (1-\tau_k)\bar{x}^k + \tau_k\hat{x}_k^*)$ in (ADSGARD), $p_y(y) \geq p_y(\bar{y}^c) = 0$, and $\beta_{k+1} \geq (1-\tau_k)\beta_k$, we can derive

$$\begin{aligned}
G_{k+1}(\bar{w}^{k+1}) &\overset{(24)}{:=} f_{\beta_{k+1}}(\bar{x}^{k+1}) - g_{\gamma_{k+1}}(\bar{y}^{k+1}) \\
&\overset{(23)}{=} \max\left\{f(\bar{x}^{k+1}) + \langle A\bar{x}^{k+1} - b, y\rangle - \beta_{k+1}p_y(y) : y \in \mathbb{R}^m\right\} - g_{\gamma_{k+1}}(\bar{y}^{k+1}) \\
&\overset{(48)}{\leq} \max\Big\{f((1-\tau_k)\bar{x}^k + \tau_k\hat{x}_k^*)) + \langle (1-\tau_k)(A\bar{x}^k - b) + \tau_k(A\hat{x}_k^* - b), y\rangle \\
&\qquad - (1-\tau_k)\beta_k p_y(y) : y \in \mathbb{R}^m\Big\} - g_{\gamma_{k+1}}(\bar{y}^{k+1}) \\
&\leq \max\Big\{(1-\tau_k)\left[f(\bar{x}^k) + \langle A\bar{x}^k - b, y\rangle - \beta_k p_y(y)\right]_{[1]} \\
&\qquad + \tau_k\left[f(\hat{x}_k^*) + \langle A\hat{x}_k^* - b, y\rangle\right]_{[2]} : y \in \mathbb{R}^m\Big\} - g_{\gamma_{k+1}}(\bar{y}^{k+1}). \quad (102)
\end{aligned}$$

Now, we estimate the terms $[\cdot]_{[1]}$ and $[\cdot]_{[2]}$ in (102) separately. Since $\bar{y}_k^*$ is the solution of the strongly concave maximization problem (25), using (23) we have

$$\begin{aligned}[\cdot]_{[1]} &:= f(\bar{x}^k) + \langle A\bar{x}^k - b, y\rangle - \beta_k p_y(y)\\ &\le \max_{y\in\mathbb{R}^m}\Big\{f(\bar{x}^k) + \langle A\bar{x}^k - b, y\rangle - \beta_k p_y(y)\Big\} - \frac{\beta_k}{2}\|y - \bar{y}_k^*\|^2\\ &= f_{\beta_k}(\bar{x}^k) - \frac{\beta_k}{2}\|y - \bar{y}_k^*\|^2. \end{aligned} \tag{103}$$

By (24), we have $G_k(\bar{w}^k) = f_{\beta_k}(\bar{x}^k) - g_{\gamma_k}(\bar{y}^k)$ which implies $f_{\beta_k}(\bar{x}^k) = G_k(\bar{w}^k) + g_{\gamma_k}(\bar{y}^k)$. Substituting this into (103) we get

$$[\cdot]_{[1]} \le G_k(\bar{w}^k) + g_{\gamma_k}(\bar{y}^k) - \frac{\beta_k}{2}\|y - \bar{y}_k^*\|^2, \quad \forall y \in \mathbb{R}^m. \tag{104}$$

Alternatively, we expand the term $[\cdot]_{[2]}$ of (102) as

$$\begin{aligned}[\cdot]_{[2]} &:= f(\hat{x}_k^*) + \langle A\hat{x}_k^* - b, y\rangle\\ &= f(\hat{x}_k^*) + \langle A\hat{x}_k^* - b, \hat{y}^k\rangle + \gamma_{k+1}p_x(A\hat{x}_k^*) + \langle A\hat{x}_k^* - b, y - \hat{y}^k\rangle - \gamma_{k+1}p_x(A\hat{x}_k^*)\\ &= g_{\gamma_{k+1}}(\hat{y}^k) + \langle \nabla g_{\gamma_{k+1}}(\hat{y}^k), y - \hat{y}^k\rangle - \gamma_{k+1}p_x(A\hat{x}_k^*), \end{aligned} \tag{105}$$

where, in the last line, we use $\nabla g_{\gamma_{k+1}}(\hat{y}^k) = A\hat{x}_k^* - b$. Let us denote $\hat{p}_k^* := p_x(A\hat{x}_k^*)$. Then, substituting (104) and (105) into (102) we get

$$\begin{aligned}G_{k+1}(\bar{w}^{k+1}) &\le (1-\tau_k)G_k(\bar{w}^k) - g_{\gamma_{k+1}}(\bar{y}^{k+1}) - \tau_k\gamma_{k+1}\hat{p}_k^*\\ &\quad + \max\Big\{(1-\tau_k)g_{\gamma_k}(\bar{y}^k) - \frac{(1-\tau_k)\beta_k}{2}\|y - \bar{y}_k^*\|^2\\ &\qquad + \tau_k\Big[g_{\gamma_{k+1}}(\hat{y}^k) + \langle\nabla g_{\gamma_{k+1}}(\hat{y}^k), y - \hat{y}^k\rangle\Big] : y \in \mathbb{R}^m\Big\}. \end{aligned} \tag{106}$$

Since $\nabla g_{\gamma_{k+1}}^i(\hat{y}^k) = A_i\hat{x}_{k,i}^* - b_i$ and $\nabla g_{\gamma_{k+1}}^i(\bar{y}^k) = A_i\bar{x}_{k,i}^* - b_i$ for $i \in \mathcal{I}_N$, we have

$$\tilde{p}_k^* := \sum_{i=1}^N \frac{1}{2\bar{L}_i}\|\nabla g_{\gamma_{k+1}}^i(\bar{y}^k) - \nabla g_{\gamma_{k+1}}^i(\hat{y}^k)\|_*^2 = \sum_{i=1}^N \frac{1}{2\bar{L}_i}\|A_i(\bar{x}_{k,i}^* - \hat{x}_{k,i}^*)\|_*^2. \tag{107}$$

For each $i \in \mathcal{I}_N$, using (30) of Lemma 2 with $y := \bar{y}^k$ and $\bar{y} := \hat{y}^k$, we obtain

$$g_\gamma^i(\bar{y}^k) \le g_\gamma^i(\hat{y}^k) + \langle\nabla g_\gamma^i(\hat{y}^k), \bar{y}^k - \hat{y}^k\rangle - \frac{\gamma}{2\bar{L}_i}\|\nabla g_\gamma^i(\bar{y}^k) - \nabla g_\gamma^i(\hat{y}^k)\|_*^2.$$

By summing up this inequality from $i = 1$ to $i = N$, and using $\gamma := \gamma_{k+1}$, the definition of $g_\gamma(\cdot) = \sum_{i=1}^N g_\gamma^i(\cdot)$ and (107), we get

$$g_{\gamma_{k+1}}(\bar{y}^k) \le g_{\gamma_{k+1}}(\hat{y}^k) + \langle\nabla g_{\gamma_{k+1}}(\hat{y}^k), \bar{y}^k - \hat{y}^k\rangle - \gamma_{k+1}\tilde{p}_k^*. \tag{108}$$

Now, using (32) of Lemma 2 with $\bar{y} := \bar{y}^k$, $\gamma := \gamma_k$ and $\bar{\gamma} := \gamma_{k+1}$, then summing up the results from $i = 1$ to $i = N$, and combining with (108), we have

$$\begin{aligned}g_{\gamma_k}(\bar{y}^k) &\overset{(32)}{\le} g_{\gamma_{k+1}}(\bar{y}^k) + (\gamma_k - \gamma_{k+1})p_x(A\bar{x}_k^*)\\ &\overset{(108)}{\le} g_{\gamma_{k+1}}(\hat{y}^k) + \langle\nabla g_{\gamma_{k+1}}(\hat{y}^k), \bar{y}^k - \hat{y}^k\rangle + (\gamma_k - \gamma_{k+1})\bar{p}_k^* - \gamma_{k+1}\tilde{p}_k^*, \end{aligned} \tag{109}$$

where $\bar{p}_k^* := p_x(A\bar{x}_k^*)$. Substituting (109) into (106) we further deduce

$$G_{k+1}(\bar{w}^{k+1}) \leq (1-\tau_k)G_k(\bar{w}^k) - \tau_k\gamma_{k+1}\hat{p}_k^* - (1-\tau_k)\big[\gamma_{k+1}\tilde{p}_k^* - (\gamma_k-\gamma_{k+1})\bar{p}_k^*\big] + \max\Big\{g_{\gamma_{k+1}}(\hat{y}^k) + \langle\nabla g_{\gamma_{k+1}}(\hat{y}^k), (1-\tau_k)\bar{y}^k + \tau_k y - \hat{y}^k\rangle - \frac{(1-\tau_k)\beta_k}{2}\|y-\bar{y}_k^*\|^2 : y\in\mathbb{R}^m\Big\} - g_{\gamma_{k+1}}(\bar{y}^{k+1}). \quad (110)$$

Let us define $u := (1-\tau_k)\bar{y}^k + \tau_k y$. Since $y\in\mathbb{R}^m$, we have $u\in\mathbb{R}^m$. Moreover, using the first line $\hat{y}^k := (1-\tau_k)\bar{y}^k + \tau_k\bar{y}_k^*$ of (ADSGARD), one has $u-\hat{y}^k = \tau_k(y-\bar{y}_k^*)$. Using this expression into (110) we obtain

$$\begin{aligned} G_{k+1}(\bar{w}^{k+1}) &\leq (1-\tau_k)G_k(\bar{w}^k) - g_{\gamma_{k+1}}(\bar{y}^{k+1}) - \mathcal{T}_k \\ &\quad + \max_{u\in\mathbb{R}^m}\Big\{g_{\gamma_{k+1}}(\hat{y}^k) + \langle\nabla g_{\gamma_{k+1}}(\hat{y}^k), u-\hat{y}^k\rangle - \frac{(1-\tau_k)\beta_k}{2\tau_k^2}\|u-\hat{y}^k\|^2\Big\} \\ &\leq (1-\tau_k)G_k(\bar{w}^k) - g_{\gamma_{k+1}}(\bar{y}^{k+1}) - \mathcal{T}_k \\ &\quad + \max_{u\in\mathbb{R}^m}\Big\{g_{\gamma_{k+1}}(\hat{y}^k) + \langle\nabla g_{\gamma_{k+1}}(\hat{y}^k), u-\hat{y}^k\rangle - \frac{L_{g_{\gamma_{k+1}}}}{2}\|u-\hat{y}^k\|^2\Big\}_{[3]}, \end{aligned} \quad (111)$$

where the last inequality follows from the last condition of (48), i.e., $\frac{(1-\tau_k)\beta_k}{\tau_k^2} \geq \frac{\bar{L}_g}{\gamma_{k+1}} = \gamma_{k+1}^{-1}\sum_{i=1}\bar{L}_i := L_{g_{\gamma_{k+1}}}$, and the quantity $\mathcal{T}_k$ is given by

$$\mathcal{T}_k := \tau_k\gamma_{k+1}\hat{p}_k^* + (1-\tau_k)\big[\gamma_{k+1}\tilde{p}_k^* - (\gamma_k-\gamma_{k+1})\bar{p}_k^*\big]. \quad (112)$$

Using line 2 of (ADSGARD), we can easily bound the term $[\cdot]_{[3]}$ of (111) as

$$\begin{aligned} [\cdot]_{[3]} &:= \max\Big\{g_{\gamma_{k+1}}(\hat{y}^k) + \langle\nabla g_{\gamma_{k+1}}(\hat{y}^k), u-\hat{y}^k\rangle - \frac{L_{g_{\gamma_{k+1}}}}{2}\|u-\hat{y}^k\|^2 : u\in\mathbb{R}^m\Big\} \\ &= g_{\gamma_{k+1}}(\hat{y}^k) + \langle\nabla g_{\gamma_{k+1}}(\hat{y}^k), \bar{y}^{k+1}-\hat{y}^k\rangle - \frac{L_{g_{\gamma_{k+1}}}}{2}\|\bar{y}^{k+1}-\hat{y}^k\|^2 \\ &\leq g_{\gamma_{k+1}}(\bar{y}^{k+1}), \end{aligned}$$

where the last inequality follows from Lipschitz continuity of ∇g_γ in Lemma 2. Using this inequality into (111) we eventually get

$$G_{k+1}(\bar{w}^{k+1}) \leq (1-\tau_k)G_k(\bar{w}^k) - \mathcal{T}_k. \quad (113)$$

Finally, we lower bound the quantity $\mathcal{T}_k$. From the definition (112) of $\mathcal{T}_k$, we write

$$\begin{aligned} \mathcal{T}_k = \gamma_{k+1}\Big\{ &\sum_{i\in\mathcal{I}_1}\Big[\tau_k p_i(\hat{x}_{k,i}^*) + \frac{(1-\tau_k)}{2\bar{L}_i}\|A_i(\hat{x}_{k,i}^* - \bar{x}_{k,i}^*)\|_*^2 - (1-\tau_k)\Big(\frac{\gamma_k}{\gamma_{k+1}}-1\Big)p_i(\bar{x}_{k,i}^*)\Big] \\ &+ \sum_{i\notin\mathcal{I}_1}\Big[\tau_k p_i(A_i\hat{x}_{k,i}^*) + \frac{(1-\tau_k)}{2\bar{L}_i}\|A_i(\hat{x}_{k,i}^* - \bar{x}_{k,i}^*)\|_*^2 - (1-\tau_k)\Big(\frac{\gamma_k}{\gamma_{k+1}}-1\Big)p_i(A_i\bar{x}_{k,i}^*)\Big]\Big\}. \end{aligned}$$

From Assumption A.2, since p_i is 1-strongly convex, and its gradient is L_{p_i}- Lipschitz continuous, we have

$$\begin{cases} \frac{1}{2\bar{L}_{A_i}}\|A_i(\hat{x}_{k,i}^* - \bar{x}_i^c)\|_*^2 \leq p_i(\bar{x}_{k,i}^*) \leq \frac{L_{p_i}}{2\sigma_i^2}\|A_i(\bar{x}_{k,i}^* - \bar{x}_i^c)\|_*^2, & i\in\mathcal{I}_1 \\ \frac{1}{2}\|A_i\hat{x}_{k,i}^* - \bar{u}_i^c\|_*^2 \leq p_i(A_i\bar{x}_{k,i}^*) \leq \frac{L_{p_i}}{2}\|A_i\bar{x}_{k,i}^* - \bar{u}_i^c\|_*^2, & i\notin\mathcal{I}_1. \end{cases}$$

Let us denote by $\hat{v}_i^k := A_i(x^*_{\gamma_{k+1},i}(\hat{y}^k) - \bar{x}_i^c)$ and $\bar{v}_i^k := A_i(x^*_{\gamma_{k+1},i}(\bar{y}^k) - \bar{x}_i^c)$ for $i \in \mathcal{I}_1$, and $\hat{v}_i^k := A_i x^*_{\gamma_{k+1},i}(\hat{y}^k) - \bar{u}_i^c$ and $\bar{v}_i^k := A_i x^*_{\gamma_{k+1},i}(\bar{y}^k) - \bar{u}_i^c$ for $i \in \bar{\mathcal{I}}_1$. Using the two last inequalities, we can lower bound $\mathcal{T}_k$ as

$$\begin{aligned}
\mathcal{T}_k &\geq \frac{\gamma_{k+1}}{2}\sum_{i\notin\mathcal{I}_1}\Big[\tau_k\|\hat{v}_i^k\|_*^2 + (1-\tau_k)\|\hat{v}_i^k - \bar{v}_i^k\|_*^2 - \frac{(1-\tau_k)(\gamma_k-\gamma_{k+1})L_{p_i}}{\gamma_{k+1}}\|\bar{v}_i^k\|_*^2\Big]\\
&+\frac{\gamma_{k+1}}{2}\sum_{i\in\mathcal{I}_1}\frac{1}{\bar{L}_{A_i}}\Big[\tau_k\|\hat{v}_i^k\|_*^2+(1-\tau_k)\|\hat{v}_i^k-\bar{v}_i^k\|_*^2-\frac{(1-\tau_k)(\gamma_k-\gamma_{k+1})L_{p_i}\bar{L}_{A_i}}{\gamma_{k+1}\sigma_i^2}\|\bar{v}_i^k\|_*^2\Big]\\
&= \frac{\gamma_{k+1}}{2}\sum_{i\notin\mathcal{I}_1}\Big[\|\hat{v}_i^k - (1-\tau_k)\bar{v}_i^k\|_*^2 + (1-\tau_k)\Big(\tau_k - \frac{(\gamma_k-\gamma_{k+1})L_{p_i}}{\gamma_{k+1}}\Big)\|\bar{v}_i^k\|_*^2\Big]\\
&+\frac{\gamma_{k+1}}{2}\sum_{i\in\mathcal{I}_1}\frac{1}{\bar{L}_{A_i}}\Big[\|\hat{v}_i^k - (1-\tau_k)\bar{v}_i^k\|_*^2 + (1-\tau_k)\Big(\tau_k - \frac{(\gamma_k-\gamma_{k+1})L_{p_i}\bar{L}_{A_i}}{\gamma_{k+1}\sigma_i^2}\Big)\|\bar{v}_i^k\|_*^2\Big].
\end{aligned}$$

From this estimate, we can see that if

$$\left(\frac{\tau_k}{\hat{c}_2} + 1\right)\gamma_{k+1} \geq \gamma_k, \quad \text{with} \quad \hat{c}_2 := \max\left\{\max_{i\in\mathcal{I}_1}\{L_{p_i}\}, \max_{i\notin\mathcal{I}_1}\left\{\frac{L_{p_i}\bar{L}_{A_i}}{\sigma_i^2}\right\}\right\}, \tag{114}$$

then $\mathcal{T}_k \geq \frac{\gamma_{k+1}}{2}\sum_{i=1}^N \frac{1}{\bar{L}_i}\|\hat{v}_i^k - (1-\tau_k)\bar{v}_i^k\|_*^2 \geq 0$. Hence, (113) leads to $G_{k+1}(\bar{w}^{k+1}) \leq (1-\tau_k)G_k(\bar{w}^k) + \tau_k\psi_k$, which is (49) with $\psi_k := -\mathcal{T}_k \leq 0$. Moreover, (114) is exactly the second condition of (48). □

A.3.2 The proof of Lemma 7: The update rule for parameters.
The tightest conditions obtained from (48) are $\gamma_{k+1} = (1+\hat{c}_2^{-1}\tau_k)^{-1}\gamma_k$, $\beta_{k+1} = (1-\tau_k)\beta_k$ and $(1-\tau_k)\gamma_{k+1}\beta_k = \bar{L}_g\tau_k^2$. By induction, we can derive from these equalities the tightest condition for τ_k as

$$\tau_k^2 = \frac{\tau_{k+1}^2(1+\hat{c}_2^{-1}\tau_{k+1})}{1-\tau_{k+1}}.$$

Similar to the proof of Lemma 5, we can show that $\tau_k = \mathcal{O}\left(\frac{1}{k}\right)$, which is optimal.

Let us choose $\tau_k := \frac{\hat{c}_2}{k+r}$ for some $r > \hat{c}_2$. With this choice, $\tau_0 = \frac{\hat{c}_2}{r} \in (0,1)$. We choose $\gamma_{k+1} := \frac{\gamma_k}{(1+\tau_k/\hat{c}_2)} = \frac{\gamma_k(k+r)}{k+r+1}$. By induction, we get $\gamma_k = \frac{\gamma_0 r}{k+r}$. We compute β_k from the last condition of (48) to get

$$\beta_k = \frac{\bar{L}_g\tau_k^2}{(1-\tau_k)\gamma_{k+1}} = \frac{\bar{L}_g\hat{c}_2^2(k+r+1)}{\gamma_0 r(k+r)(k+r-\hat{c}_2)}. \tag{115}$$

It remains to check the second condition of (48), i.e., $\beta_{k+1} \geq (1-\tau_k)\beta_k$. Using (115), this inequality is equivalent to

$$\frac{k+r+2}{(k+r+1)(k+r+1-\hat{c}_2)} \geq \left(1-\frac{\hat{c}_2}{k+r}\right)\left(\frac{k+r+1}{(k+r)(k+r-\hat{c}_2)}\right) = \frac{(k+r+1)}{(k+r)^2}. \tag{116}$$

Let $\hat{k} := k+r$, (116) is equivalent to $(\hat{c}_2-1)\hat{k}^2 + (2\hat{c}_2-3)\hat{k} + (\hat{c}_2-1) \geq 0$. This condition holds if $\hat{c}_2 \geq 3/2$ and $\hat{k} \geq 0$.

We now define $\bar{c}_2 := \max\left\{\hat{c}_2, \frac{3}{2}\right\}$ and take $\tau_k := \frac{\bar{c}_2}{k+\bar{c}_2+1}$. Then, $\tau_k \in (0,1)$ and all the conditions in (48) are satisfied. Using the update formula of τ_k, γ_k and β_k as above, we obtain the last conclusion of Lemma 7. □

A.3.3 The proof of Lemma 8: Finding a starting point
Given $\bar{y}^c = \nabla p_y^*(\mathbf{0}^m) \in \mathbb{R}^m$, we denote by $\bar{x}^0 \equiv \bar{x}_0^* := x_{\gamma_0}^*(\bar{y}^c)$. Using (30) with $L_{g_{\gamma_0}} = \gamma_0^{-1}\bar{L}_g$, we can derive

$$\begin{aligned} g_{\gamma_0}(\bar{y}^0) &\geq g_{\gamma_0}(\bar{y}^c) + \langle \nabla g_{\gamma_0}(\bar{y}^c), \bar{y}^0 - \bar{y}^c\rangle - \frac{L_{g_{\gamma_0}}}{2}\|\bar{y}^0 - \bar{y}^c\|^2 \\ &= f(\bar{x}_0^*) + \langle A\bar{x}_0^* - b, \bar{y}^c\rangle + \langle A\bar{x}_0^* - b, \bar{y}^0 - \bar{y}^c\rangle + \gamma_0 p_x(A\bar{x}_0^*) - \frac{L_{g_{\gamma_0}}}{2}\|\bar{y}^0 - \bar{y}^c\|^2 \\ &= f(\bar{x}^0) + \langle A\bar{x}^0 - b, \bar{y}^0\rangle - \frac{L_{g_{\gamma_0}}}{2}\|\bar{y}^0 - \bar{y}^c\|^2 + \gamma_0 p_x(A\bar{x}^0). \end{aligned}$$

Using this inequality, and $\bar{y}^0 := \nabla p_y^*(\beta_0^{-1}(A\bar{x}^0 - b))$, we have

$$\begin{aligned} G_{\gamma_0\beta_0}(\bar{w}^0) &:= f_{\beta_0}(\bar{x}^0) - g_{\gamma_0}(\bar{y}^0) = f(\bar{x}^0) + \beta_0 p_y^*(\beta_0^{-1}(A\bar{x}^0 - b)) - g_{\gamma_0}(\bar{y}^0) \\ &\leq \beta_0 p_y^*(\beta_0^{-1}(A\bar{x}^0 - b)) - \langle A\bar{x}^0 - b, \bar{y}^0\rangle - \gamma_0 p_x(A\bar{x}^0) + (L_{g_{\gamma_0}}/2)\|\bar{y}^0 - \bar{y}^c\|^2 \\ &= \beta_0\big[p_y^*(\beta_0^{-1}(A\bar{x}^0 - b)) - \langle \nabla p_y^*(\beta_0^{-1}(A\bar{x}^0 - b)), \beta_0^{-1}(A\bar{x}^0 - b)\rangle\big] \\ &\quad + (L_{g_{\gamma_0}}/2)\|\bar{y}^0 - \bar{y}^c\|^2 - \gamma_0 p_x(A\bar{x}^0). \end{aligned} \tag{117}$$

By the $L_{p_y}^{-1}$-strong convexity of p_y^*, we have

$$p_y^*(v) - \langle \nabla p_y^*(v), v\rangle \leq p_y^*(0) - \frac{1}{2L_{p_y}}\|v\|^2 = \max_{u\in\mathbb{R}^m}\{-p_y(u)\} - \frac{1}{2L_{p_y}}\|v\|^2 \leq -\frac{1}{2L_{p_y}}\|v\|^2.$$

Using this inequality with $v := \beta_0^{-1}(A\bar{x}^0 - b)$, and then substitute the result into (117) we obtain

$$G_{\gamma_0\beta_0}(\bar{w}^0) \leq -\gamma_0 p_x(A\bar{x}^0) - \frac{1}{2\beta_0 L_{p_y}}\|A\bar{x}^0 - b\|^2 + \frac{\bar{L}_g}{2\gamma_0}\|\bar{y}^0 - \bar{y}^c\|^2.$$

Finally, using $\bar{y}^c := \nabla p_y^*(\mathbf{0}^m)$ and the 1- Lipschitz continuity of ∇p_y^*, we have $\|\bar{y}^0 - \bar{y}^c\| = \|\nabla p_y^*(\beta_0^{-1}(A\bar{x}^0 - b)) - \nabla p_y^*(0)\| \leq \|\beta_0^{-1}(A\bar{x}^0 - b)\| = \beta_0^{-1}\|A\bar{x}^0 - b\|$. Substituting this bound into the last inequality we obtain (52). The remaining statement of this lemma consequently follows from (52). □

A.4 The proof of Theorem 3: The convergence analysis of the APSGARD method
Let us abbreviate $\bar{x}_k^* := \bar{x}_{\gamma_k}^*(\bar{y}^k)$, $\hat{y}_k^* := y_{\beta_{k+1}}^*(\hat{x}^k)$, $\bar{y}_k^* := y_{\beta_{k+1}}^*(\bar{x}^k)$, and $G_k := G_{\gamma_k\beta_k}$.

Using the definition of g_γ, the third line $\bar{y}^{k+1} = (1-\tau_k)\bar{y}^k + \tau_k\hat{y}_k^*$ of (APSGARD) and the first condition $\gamma_{k+1} \geq (1-\tau_k)\gamma_k$ of (58), we have

$$\begin{aligned} G_{k+1}(\bar{w}^{k+1}) &= f_{\beta_{k+1}}(\bar{x}^{k+1}) - g_{\gamma_{k+1}}(\bar{y}^{k+1}) \\ &\overset{\text{(APSGARD)+(58)}}{=} f_{\beta_{k+1}}(\bar{x}^{k+1}) - \min\Big\{\tau_k\big[f(x) + \langle Ax - b, \hat{y}_k^*\rangle\big]_{[1]} \\ &\quad + (1-\tau_k)\big[f(x) + \langle Ax - b, \bar{y}^k\rangle + \gamma_k p_x(x)\big]_{[2]} : x \in \mathcal{X}\Big\}. \end{aligned} \tag{118}$$

Since $f_\beta(\cdot) = f(\cdot) + \bar{p}_\beta(\cdot) \equiv f(\cdot) + \beta p_y^*(\beta^{-1}(A\cdot - b))$ due to (28), we can estimate the first term $[\cdot]_{[1]}$ of (118) as

$$\begin{aligned} [\cdot]_{[1]} &:= f(x) + \langle Ax - b, \hat{y}_k^*\rangle \\ &= f(x) + \langle A\hat{x}^k - b, \hat{y}_k^*\rangle - \beta_{k+1}p_y(\hat{y}_k^*) + \langle A^T\hat{y}_k^*, x - \hat{x}^k\rangle + \beta_{k+1}p_y(\hat{y}_k^*) \\ &= f(x) + \bar{p}_{\beta_{k+1}}(\hat{x}^k) + \langle \nabla\bar{p}_{\beta_{k+1}}(\hat{x}^k), x - \hat{x}^k\rangle + \beta_{k+1}p_y(\hat{y}_k^*). \end{aligned} \tag{119}$$

Next, we bound $[\cdot]_{[2]}$ of (118) as follows. Using the definition of g_γ with $p_x(A(\cdot)) \leftarrow p_x(\cdot)$ and $g_{\gamma_k}(\bar{y}^k) = f_{\beta_k}(\bar{x}^k) - G_k(\bar{w}^k)$, we can derive

$$\begin{aligned}[\cdot]_{[2]} &:= f(x) + \langle Ax - b, \bar{y}^k\rangle + \gamma_k p_x(x)\\ &\geq \min\Big\{f(x) + \langle Ax - b, \bar{y}^k\rangle + \gamma_k p_x(x) : x \in \mathcal{X}\Big\} + (\gamma_k/2)\|x - \bar{x}_k^*\|^2\\ &= g_{\gamma_k}(\bar{x}^k) + (\gamma_k/2)\|x - \bar{x}_k^*\|^2 = f_{\beta_k}(\bar{x}^k) - G_k(\bar{w}^k) + (\gamma_k/2)\|x - \bar{x}_k^*\|^2. \quad (120)\end{aligned}$$

Substituting $\beta := \beta_{k+1}$ and $x := \bar{x}^k$ into (81), we have

$$\bar{p}_{\beta_{k+1}}(\bar{x}^k) \geq \bar{p}_{\beta_{k+1}}(\hat{x}^k) + \langle \nabla\bar{p}_{\beta_{k+1}}(\hat{x}^k), \bar{x}^k - \hat{x}^k\rangle + \frac{1}{2L_{p_y}\beta_{k+1}}\|A(\bar{x}^k - \hat{x}^k)\|^2.$$

Using this inequality, the first inequality of (90) as in the proof of Lemma 4, and the convexity and Lipschitz gradient continuity of $\bar{p}_\beta$, we have

$$\begin{aligned}f_{\beta_k}(\bar{x}^k) &\geq f_{\beta_{k+1}}(\bar{x}^k) - (\beta_k - \beta_{k+1})p_y(y^*_{\beta_{k+1}}(\bar{x}^k))\\ &= f(\bar{x}^k) + \bar{p}_{\beta_{k+1}}(\bar{x}^k) - (\beta_k - \beta_{k+1})p_y(y^*_{\beta_{k+1}}(\bar{x}^k))\\ &\geq f(\bar{x}^k) + \bar{p}_{\beta_{k+1}}(\hat{x}^k) + \langle \nabla\bar{p}_{\beta_{k+1}}(\hat{x}^k), \bar{x}^k - \hat{x}^k\rangle + \hat{\mathcal{T}}_k, \quad (121)\end{aligned}$$

where $\hat{\mathcal{T}}_k := \frac{1}{2L_{p_y}\beta_{k+1}}\|A(\bar{x}^k - \hat{x}^k)\|^2 - (\beta_k - \beta_{k+1})p_y(y^*_{\beta_{k+1}}(\bar{x}^k))$. Substituting (121), (120) and (119) into (118), we obtain

$$\begin{aligned}G_{k+1}(\bar{w}^{k+1}) &\leq (1-\tau_k)G_k(\bar{w}^k) + f_{\beta_{k+1}}(\bar{x}^{k+1})\\ &\quad - \min_{x\in\mathcal{X}}\Big\{\tau_k\Big[f(x) + \bar{p}_{\beta_{k+1}}(\hat{x}^k) + \langle\nabla\bar{p}_{\beta_{k+1}}(\hat{x}^k), x - \hat{x}^k\rangle + \beta_{k+1}p_y(\hat{y}_k^*)\Big]\\ &\quad + (1-\tau_k)\Big[f(\bar{x}^k) + \bar{p}_{\beta_{k+1}}(\hat{x}^k) + \langle\nabla\bar{p}_{\beta_{k+1}}(\hat{x}^k), \bar{x}^k - \hat{x}^k\rangle + \frac{\gamma_k}{2}\|x - \bar{x}_k^*\|^2 + \hat{\mathcal{T}}_k\Big]\Big\}\\ &\leq (1-\tau_k)G_k(\bar{w}^k) + f_{\beta_{k+1}}(\bar{x}^{k+1}) - \tau_k\beta_{k+1}p_y(\hat{y}_k^*) - (1-\tau_k)\hat{\mathcal{T}}_k\\ &\quad - \min_{x\in\mathcal{X}}\Big\{(1-\tau_k)f(\bar{x}^k) + \tau_k f(x) + \bar{p}_{\beta_{k+1}}(\hat{x}^k) + \frac{(1-\tau_k)\gamma_k}{2}\|x - \bar{x}_k^*\|^2\\ &\quad + \langle\nabla\bar{p}_{\beta_{k+1}}(\hat{x}^k), (1-\tau_k)\bar{x}^k + \tau_k x - \hat{x}^k\rangle\Big\}. \quad (122)\end{aligned}$$

To further estimate (122), we now define the quantity $\widetilde{\mathcal{T}}_k$ as

$$\begin{aligned}\widetilde{\mathcal{T}}_k &:= \tau_k\beta_{k+1}p_y(\hat{y}_k^*) + (1-\tau_k)\hat{\mathcal{T}}_k\\ &= \tau_k\beta_{k+1}p_y(\hat{y}_k^*) + \frac{(1-\tau_k)}{2L_{p_y}\beta_{k+1}}\|A(\bar{x}^k - \hat{x}^k)\|^2 - (1-\tau_k)(\beta_k - \beta_{k+1})p_y(\bar{y}_k^*). \quad (123)\end{aligned}$$

For any $x \in \mathcal{X}$, the point $z := (1-\tau_k)\bar{x}^k + \tau_k x \in \mathcal{X}$. Moreover, by using the first line $\hat{x}^k = (1-\tau_k)\bar{x}^k + \tau_k\bar{x}_k^*$ of (APSGARD), we have $z - \hat{x}^k = \tau_k(x - \bar{x}_k^*)$. Using these relations and the convexity of f with $(1-\tau_k)f(\bar{x}^k) + \tau_k f(x) \geq f((1-\tau_k)\bar{x}^k + \tau_k x) = f(z)$ into (122), we can further derive

$$\begin{aligned}G_{k+1}(\bar{w}^{k+1}) &\leq (1-\tau_k)G_k(\bar{w}^k) + f_{\beta_{k+1}}(\bar{x}^{k+1}) - \widetilde{\mathcal{T}}_k\\ &\quad - \min_{z\in\mathcal{X}}\Big\{f(z) + \bar{p}_{\beta_{k+1}}(\hat{x}^k) + \langle\nabla\bar{p}_{\beta_{k+1}}(\hat{x}^k), z - \hat{x}^k\rangle + \frac{(1-\tau_k)\gamma_k}{2\tau_k^2}\|z - \hat{x}^k\|^2\Big\}\\ &\overset{(58)}{\leq} (1-\tau_k)G_k(\bar{w}^k) + f_{\beta_{k+1}}(\bar{x}^{k+1}) - \widetilde{\mathcal{T}}_k\\ &\quad - \min_{z\in\mathcal{X}}\Big\{f(z) + \bar{p}_{\beta_{k+1}}(\hat{x}^k) + \langle\nabla\bar{p}_{\beta_{k+1}}(\hat{x}^k), z - \hat{x}^k\rangle + \frac{\bar{L}_A}{2\beta_{k+1}}\|z - \hat{x}^k\|^2\Big\}, \quad (124)\end{aligned}$$

where, in the last inequality, we use $(1-\tau_k)\gamma_k\beta_{k+1} \geq \bar{L}_A\tau_k^2$ in (58). Since $\nabla\bar{p}_{\beta_{k+1}}(\hat{x}^k) = A^T\hat{y}_k^*$, the second line of (APSGARD) can be expressed as

$$\bar{x}^{k+1} = \mathop{\mathrm{argmin}}_{z\in\mathcal{X}}\Big\{f(z) + \langle\nabla\bar{p}_{\beta_{k+1}}(\hat{x}^k), z-\hat{x}^k\rangle + (\bar{L}_A/(2\beta_{k+1}))\|z-\hat{x}^k\|^2\Big\}. \tag{125}$$

Since $\nabla p_\beta(\cdot)$ is Lipschitz continuous with $L_{\bar{p}_\beta} = \beta^{-1}\|A\|^2 = \beta^{-1}\bar{L}_A$, we have

$$\begin{aligned}\mathcal{Q}_k(\bar{x}^{k+1}) &:= f(\bar{x}^{k+1}) + \bar{p}_{\beta_{k+1}}(\hat{x}^k) + \langle\nabla\bar{p}_{\beta_{k+1}}(\hat{x}^k), \bar{x}^{k+1}-\hat{x}^k\rangle + \frac{\bar{L}_A}{2\beta_{k+1}}\|\bar{x}^{k+1}-\hat{x}^k\|^2\\ &\geq f(\bar{x}^{k+1}) + \bar{p}_{\beta_{k+1}}(\bar{x}^{k+1}).\end{aligned}$$

Using this inequality and (125) into (124), we get

$$G_{k+1}(\bar{w}^{k+1}) \leq (1-\tau_k)G_k(\bar{w}^k) - \widetilde{\mathcal{T}}_k. \tag{126}$$

Finally, using (98) and (99) we can estimate the quantity $\widetilde{\mathcal{T}}_k$ as

$$\begin{aligned}\widetilde{\mathcal{T}}_k &:= \tau_k\beta_{k+1}p_y(\hat{y}_k^*) + \frac{(1-\tau_k)}{2L_{p_y}\beta_{k+1}}\|A(\bar{x}^k-\hat{x}^k)\|^2 - (1-\tau_k)(\beta_k-\beta_{k+1})p_y(\bar{y}_k^*)\\ &\geq \frac{1}{2L_{p_y}\beta_{k+1}}\left[\|\hat{r}_k-(1-\tau_k)\bar{r}_k\|^2 + (1-\tau_k)\left(\tau_k - \left(\frac{\beta_k}{\beta_{k+1}}-1\right)L_{p_y}^2\right)\|\bar{r}_k\|^2\right],\end{aligned} \tag{127}$$

where $\bar{r}_k := A\bar{x}^k - b$ and $\hat{r}_k := A\hat{x}^k - b$. Similar to the proof of Lemma 4, we can show that $\widetilde{\mathcal{T}}_k \geq 0$ if $\beta_{k+1}\left(1 + \frac{\tau_k}{L_{p_y}^2}\right) \geq \beta_k$, which is the second condition of (58).

Using (58) with the same argument as the proof of Lemma 7, we can derive the update rule for τ_k, γ_k and β_k as in (59). The remainder of this theorem is proved similarly as in Theorem 2. □

A.5 The proof of Theorem 4: The accelerated augmented Lagrangian method

Let $\mathcal{L}_\gamma(x,y) := f(x) + \langle y, Ax-b\rangle + \frac{\gamma}{2}\|Ax-b\|_2^2$ be the augmented Lagrangian of (1). Under Assumption A.1, by the well-known properties of $\mathcal{L}_\gamma$ [8], we have

$$\mathcal{L}_\gamma(x^\star,y) \leq \mathcal{L}_\gamma(x^\star,y^\star) \equiv \mathcal{L}(x^\star,y^\star) = f^\star = g^\star \leq \mathcal{L}_\gamma(x,y^\star),$$

for all $x\in\mathcal{X}, y\in\mathbf{R}^m$, $(x^\star,y^\star)\in\mathcal{W}^\star$ and $\gamma>0$. This expression leads to

$$g_\gamma(y) \leq f(x) + \langle y^\star, Ax-b\rangle + \frac{\gamma}{2}\|Ax-b\|_2^2 \leq f(x) + \|y^\star\|\|Ax-b\|_2 + \frac{\gamma}{2}\|Ax-b\|_2^2.$$

Hence, for any $y^\star\in\mathcal{Y}^\star$, we obtain

$$f(x) - g_\gamma(y) \geq f(x) - f^\star \geq -\|y^\star\|\|Ax-b\|_2 - \frac{\gamma}{2}\|Ax-b\|_2^2, \tag{128}$$

for all $(x,y)\in\mathcal{W}$. Let $t := \|A\bar{x}^k - b\|_2$. By combining (128) and $G_{\gamma\beta}(w) = f_\beta(x) - g_\gamma(y)$, we obtain $\frac{(1-\gamma_k\beta_k)}{\beta_k}t^2 - 2\|y^\star\|t - 2G_{\gamma_0\beta_k}(\bar{w}^k) \leq 0$. Since $\beta_{k+1} = (1-\tau_k)\beta_k$, we have $\gamma_k\beta_k = \gamma_0\beta_k = (1-\tau_0)(1-\tau_1)\cdots(1-\tau_{k-1})\beta_0\gamma_0 = \prod_{i=0}^{k-1}(1-\tau_i) < 1$ for $k\geq 1$. Hence, the last inequality leads to

$$\|A\bar{x}^k - b\|_2 \leq \left(\frac{\beta_k}{1-\beta_k\gamma_0}\right)\left[\|y^\star\| + \sqrt{\|y^\star\|^2 + \frac{2(1-\gamma_0\beta_k)}{\beta_k}G_{\gamma_0\beta_k}(\bar{w}^k)}\right]. \tag{129}$$

To prove (65), we note from the update rule (67) that $(1-\tau_k) = \frac{\tau_k^2}{\tau_{k-1}^2}$ for $k \geq 1$. Hence, $\beta_k = \beta_0 \prod_{i=0}^{k-1}(1-\tau_i) = \frac{(1-\tau_0)\tau_k^2}{\tau_0^2} = \beta_0\tau_k^2$. By elementary calculations, we can show that $\tau_k \leq \frac{2}{k+3}$ for $k \geq 0$. Hence, $\frac{\gamma_0\beta_k}{1-\gamma_0\beta_k} \leq \frac{\gamma_0\beta_0\tau_k^2}{1-\gamma_0\beta_0\tau_k^2} = \frac{\tau_k^2}{1-\tau_k^2} < \frac{4}{(k+1)(k+3)} < \frac{4}{(k+1)^2}$. In addition, $G_{\gamma_0\beta_k}(\bar{w}^k) \leq 0$ due to Lemma 6. Using these estimates into (129), we obtain $\|A\bar{x}^k - b\|_2 \leq \frac{8D_{\mathcal{Y}^\star}}{\gamma_0(k+1)^2}$, which is the first inequality of (65). From (128) and $G_{\gamma\beta}(w) = f_\beta(x) - g_\gamma(y)$ we have $f(\bar{x}^k) - f^\star \leq f(\bar{x}^k) - g_{\gamma_0}(\bar{y}^k) = G_{\gamma_0\beta_k}(\bar{w}^k) \leq 0$. This inequality and (128) imply the second inequality of (65). The remaining conclusion consequently follows from (65). □

A.6 The proof of Corollary 1: The fully strong convexity of the objective function Let us define $\bar{G}_\beta(\bar{w}) := f_\beta(\bar{x}) - g(\bar{y})$, where f_β is defined by (23) and g is defined by (10). By the update scheme (ADSGARD$_\mu$) and (67), the gap reduction condition $G_{\beta_{k+1}}(\bar{w}^{k+1}) \leq (1-\tau_k)G_{\beta_k}(\bar{w}^k) + \tau_k\psi_k \leq (1-\tau_k)G_{\beta_k}(\bar{w}^k)$ in Lemma 6 holds. Hence, $G_{\beta_k}(\bar{w}^k) \leq \omega_k G_{\beta_0}(\bar{w}^0)$. Since $\bar{w}^0$ is computed by (68), we have $G_{\beta_0}(\bar{w}^0) \leq 0$. Consequently, $G_{\beta_k}(\bar{w}^k) \leq 0$ for $k \geq 0$. Similar to the proof of Lemma 3, we can show that

$$\begin{aligned}-\|y^\star\|\|A\bar{x}^k - b\| \leq f(\bar{x}^k) - f^\star \leq f(\bar{x}^k) - g(\bar{y}^k) &= f_{\beta_k}(\bar{x}^k) - g(\bar{y}^k) - \frac{1}{2\beta_k}\|A\bar{x}^k - b\|^2\\ &= G_{\beta_k}(\bar{w}^k) - \frac{1}{2\beta_k}\|A\bar{x}^k - b\|^2 \leq -\frac{1}{2\beta_k}\|A\bar{x}^k - b\|^2.\end{aligned}$$

This inequality leads to

$$-D_{\mathcal{Y}^\star}\|A\bar{x}^k - b\| \leq f(\bar{x}^k) - f^\star \leq 0, \quad \text{and} \quad \|A\bar{x}^k - b\| \leq 2\beta_k D_{\mathcal{Y}^\star}. \tag{130}$$

Now, by using [54, Theorem 4], we can show that $\{\beta_k\}$ updated by (67) satisfies $\beta_k \leq 2\beta_0(k+2)^{-2}$. Since $\beta_0 := \hat{L}_g = \sum_{i=1}^N \mu_{f_i}^{-1}\|A_i\|^2$, substituting these expressions into (130), we obtain the first and the second estimates of (69).

Finally, we prove the last estimate of (69). Indeed, by the strong convexity of f_i, we have $f_i(\bar{x}_i^k) - f_i(x_i^\star) \geq \langle \xi_{f_i}(x_i^\star), \bar{x}_i^k - x_i^\star\rangle + (\mu_{f_i}/2)\|\bar{x}_i^k - x_i^\star\|_2^2$, where $\xi_{f_i}(x_i^\star) \in \partial f_i(x_i^\star)$ is one subgradient of f_i at $x_i^\star$ for $i \in \mathcal{I}_N$. On the other hand, since $x^\star$ is the optimal solution of (1), using the optimality condition of this problem, we have $\langle \xi_{f_i}(x_i^\star) + A_i^T y^\star, x_i - x_i^\star\rangle \geq 0$ for any $x_i \in \mathcal{X}_i$ and $y^\star \in \mathcal{Y}^\star$ and $Ax_i^\star = b_i$ for $i \in \mathcal{I}_N$. Using these expressions, we can show that

$$f(\bar{x}^k) - f(x^\star) \geq \sum_{i=1}^N \frac{\mu_{f_i}}{2}\|\bar{x}_i^k - x_i^\star\|^2 - \langle A\bar{x}^k - b, y^\star\rangle \geq \frac{\underline{\mu}_f}{2}\|\bar{x}^k - x^\star\|^2 - \|y^\star\|_*\|A\bar{x}^k - b\|,$$

where $\underline{\mu}_f := \min\{\mu_{f_i} : i \in \mathcal{I}_N\}$. This estimate leads to $\|\bar{x}^k - x^\star\|^2 \leq \frac{2}{\underline{\mu}_f}[f(\bar{x}^k) - f^\star] + \frac{2\|y^\star\|_*}{\underline{\mu}_f}\|A\bar{x}^k - b\| \leq \frac{4D_{\mathcal{Y}^\star}}{(k+2)}\sqrt{\hat{L}_g/\underline{\mu}_f}$, which is the third estimate of (69). □

Acknowledgments

This work is supported in part by 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. A. Auslender. *Optimisation: Méthodes Numériques*. Masson, Paris, 1976.
2. H.H. Bauschke and P. Combettes. *Convex analysis and monotone operators theory in Hilbert spaces*. Springer-Verlag, 2011.
3. A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.
4. A. Beck and M. Teboulle. Smoothing and first order methods: A unified framework. *SIAM J. Optim.*, 22(2):557–580, 2012.
5. A. Beck and M. Teboulle. A fast dual proximal gradient algorithm for convex minimization and applications. *Oper. Res. Letter*, 42(1):1–6, 2014.
6. S. Becker, E. J. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.
7. A. Ben-Tal and A.K. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM, 2001.
8. Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series)*. Athena Scientific, 1996.
9. D.P. Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation: Numerical methods*. Prentice Hall, 1989.
10. 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(1):1–122, 2011.
11. S. Boyd and L. Vandenberghe. *Convex Optimization*. University Press, Cambridge, 2004.
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(5):32–43, 2014.
13. A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.
14. V. Chandrasekaranm, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.
15. P. L. Combettes and J.-C. Pesquet. A proximal decomposition method for solving convex variational inverse problems. *Inverse Problems*, 24(6):065014, 2008.
16. P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4:1168–1200, 2005.
17. W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. Technical report, Rice University CAAM, 2012. TR12-14.
18. F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*, volume 1-2. Springer-Verlag, 2003.
19. M. Fukuda, M. Kojima, and M. Shida. Lagrangian dual interior-point methods for semidefinite programs. *SIAM J. Optimization*, 12:1007–1031, 2002.
20. D. Goldfarb and S. Ma. Fast alternating linearization methods of minimization of the sum of two convex functions. *Math. Program., Ser. A*, pages 1–34, 2012.
21. T. Goldstein, B. ODonoghue, and S. Setzer. Fast Alternating Direction Optimization Methods. *SIAM J. Imaging Sci.*, 7(3):1588–1623, 2012.
22. B. He and X. Yuan. On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers. (submitted for publication), 2012.
23. B.S. He and X.M. Yuan. On the $\mathcal{O}(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM J. Numer. Anal.*, 50:700–709, 2012.
24. M. Kojima, N. Megiddo, S. Mizuno, and et al. Horizontal and vertical decomposition in interior point methods for linear programs. Technical report., Information Sciences, Tokyo Institute of Technology, Tokyo, 1993.
25. G. Lan and R.D.C. Monteiro. Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Math. Program.*, 2013 (under revision).
26. M. B McCoy, V. Cevher, Q. Tran-Dinh, A. Asaei, and L. Baldassarre. Convexity in source separation: Models, geometry, and algorithms. *IEEE Signal Processing Magazine*, 31(3):87–95, 2014.
27. I. Necoara and J.A.K. Suykens. Applications of a smoothing technique to decomposition in convex optimization. *IEEE Trans. Automatic control*, 53(11):2674–2679, 2008.

28. I. Necoara and J.A.K. Suykens. Interior-point lagrangian decomposition method for separable convex optimization. *J. Optim. Theory and Appl.*, 143(3):567–588, 2009.
29. Ion Necoara and Andrei Patrascu. Iteration-complexity analysis of dual first order methods for convex programming. Tech. Report., pp. 1–37, 2014, (`http://arxiv.org/abs/1409.1462`).
30. 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(5):3109–3134, 2014.
31. A. Nemirovskii and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.
32. Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. *Doklady AN SSSR*, 269 (translated as Soviet Math. Dokl.):543–547, 1983.
33. Y. Nesterov. *Introductory lectures on convex optimization: a basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, 2004.
34. Y. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM J. Optimization*, 16(1):235–249, 2005.
35. Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.
36. Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Math. Program.*, 110(2):245–259, 2007.
37. Y. Nesterov. Barrier subgradient method. *Math. Program., Ser. B*, 127:31–56, 2011.
38. Y. Nesterov. Gradient methods for minimizing composite objective function. *Math. Program.*, 140(1):125–161, 2013.
39. J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2 edition, 2006.
40. H. Ouyang, N. He, Long Q. Tran, and A. Gray. Stochastic alternating direction method of multipliers. *JMLR W&CP*, 28:80–88, 2013.
41. Y. Ouyang, Y. Chen, G. Lan, and E. JR. Pasiliao. An accelerated linearized alternating direction method of multiplier. *Tech*, 2014.
42. N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.
43. R. T. Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematics Series*. Princeton University Press, 1970.
44. R. T. Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathemathics of Operations Research*, 1:97–116, 1976.
45. R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.
46. A. Ruszczyński. On convergence of an augmented lagrangian decomposition method for sparse convex optimization. *Mathematics of Operations Research*, 20:634–656, 1995.
47. 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(1):269–297, 2014.
48. Q. Tran-Dinh and V. Cevher. Constrained convex minimization via model-based excessive gap. In *Proc. the Neural Information Processing Systems Foundation conference* (NIPS2014), pages 1–9, Montreal, Canada, December 2014.
49. Q. Tran-Dinh and V. Cevher. A primal-dual algorithmic framework for constrained convex minimization. *Tech. Report.*, LIONS-EPFL, pp. 1–54, 2014 (`http://arxiv.org/pdf/1406.5403.pdf`).
50. Q. Tran-Dinh and V. Cevher. Splitting the Smoothed Primal-Dual Gap: Optimal Alternating Direction Methods. *Tech. Report.*, LIONS-EPFL, pp. 1–32, 2015 (`http://arxiv.org/pdf/1507.03734.pdf`).
51. Q. Tran-Dinh, A. Kyrillidis, and V. Cevher. An inexact proximal path-following algorithm for constrained convex minimization. *SIAM J. Optim.*, 24(4):1718–1745, 2014.
52. Q. Tran-Dinh, A. Kyrillidis, and V. Cevher. Composite self-concordant minimization. *J. Mach. Learn. Res.*, 15:374–416, 2015.
53. Q. Tran-Dinh, I. Necoara, and M. Diehl. Path-Following Gradient-Based Decomposition Algorithms For Separable Convex Optimization. *J. Global Optim.*, 59(1):59–80, 2014.
54. Q. Tran-Dinh, C. Savorgnan, and M. Diehl. Combining Lagrangian Decomposition and Excessive Gap Smoothing Technique for Solving Large-Scale Separable Convex Optimization Problems. *Compt. Optim. Appl.*, 55(1):75–111, 2013.

55. M. J. Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and its Applications*, 1:233–253, 2014.
56. H. Wang and A. Banerjee. Bregman Alternating Direction Method of Multipliers. Tech. Report., pp. 1–18, 2013, (`http://arxiv.org/pdf/1306.3203v1.pdf`).
57. A. Yurtsever, Q. Tran-Dinh, and V. Cevher. Universal primal-dual proximal-gradient methods. *Tech. Report.*, LIONS-EPFL, pp. 1–21, 2015. (`http://arxiv.org/pdf/1502.03123.pdf`).
58. G. Zhao. A Lagrangian dual method with self-concordant barriers for multistage stochastic convex programming. *Math. Progam.*, 102:1–24, 2005.