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

Proximal Alternating Penalty Algorithms for Nonsmooth Constrained Convex Optimization

Quoc Tran-Dinh

Received: date / Accepted: date

Abstract

We develop two new proximal alternating penalty algorithms to solve a wide range class of constrained convex optimization problems. Our approach mainly relies on a novel combination of the classical quadratic penalty, alternating, Nesterov's acceleration, and homotopy techniques. The first algorithm is designed to solve generic and possibly nonsmooth constrained convex problems without requiring any Lipschitz gradient continuity or strong convexity, while achieves the best-known $\mathcal{O}(1/k)$-convergence rate in the non-ergodic sense, where k is the iteration counter. The second algorithm is also designed to solve non-strongly convex problems, but with one strongly convex objective term. This algorithm achieves the $\mathcal{O}\left(1/k^2\right)$-convergence rate on the primal constrained problem. Such a rate is obtained in two cases: (i) averaging only on the iterate sequence of the strongly convex term, or (ii) using two proximal operators of this term without averaging. In both algorithms, we allow one to linearize the second subproblem to use the proximal operator of the corresponding objective term. Then, we customize our methods to solve different convex problems, and lead to new variants. As a byproduct, these algorithms preserve the same convergence guarantees as in our main algorithms. Finally, we verify our theoretical development via different numerical examples and compare our methods with some existing state-of-the-art algorithms.

Keywords Proximal alternating algorithm · quadratic penalty method · accelerated scheme · constrained convex optimization · first-order methods · convergence rate.

Mathematics Subject Classification (2000) 90C25 · 90-08

1 Introduction

Problem statement: We develop novel numerical methods to solve the following generic, and possibly nonsmooth constrained convex optimization problem:

$$F^{\star} := \begin{cases} \displaystyle\min_{z:=(x,y)\in\mathbb{R}^p} & \{F(z) := f(x) + g(y)\}, \\ \text{s.t.} & Ax + By - c \in \mathcal{K} \end{cases} \tag{1}$$

where $f : \mathbb{R}^{p_1} \to \mathbb{R}\cup\{+\infty\}$ and $g : \mathbb{R}^{p_2} \to \mathbb{R}\cup\{+\infty\}$ are two proper, closed, and convex functions; $p := p_1 + p_2$; $A \in \mathbb{R}^{n\times p_1}$, $B \in \mathbb{R}^{n\times p_2}$, and $c \in \mathbb{R}^n$ are given; and $\mathcal{K} \subseteq \mathbb{R}^n$ is a nonempty, closed, and convex subset.

Quoc Tran-Dinh
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill (UNC-Chapel Hill)
333 Hanes Hall, CB# 3260, UNC Chapel Hill, NC 27599-3260
Email: quoctd@email.unc.edu

Problem (1), on the one hand, covers a wide range class of classical constrained convex optimization problems in practice including conic programming (e.g., linear, convex quadratic, second-order cone, and semidefinite programming), convex optimization over graphs and networks, geometric programming, monotropic convex programming, and model predictive controls (MPC) [7,10,34]. On the other hand, it can be used as a unified template to describe many recent convex optimization models arising in signal/image processing, machine learning, and statistics ranging from unconstrained to constrained settings, see, e.g., [9,31,39]. In the latter case, the underlying convex problems obtained from these applications are often challenging to solve due to their high-dimensionality and nonsmoothness. Therefore, classical optimization methods such as sequential quadratic programming, and interior-point methods are no longer efficient to solve them [34]. This fundamental challenge has opened a door for the use of first-order methods [4,11, 31]. Various first-order methods have been proposed to solve large-scale instances of (1) including [proximal] gradient and fast gradient, primal-dual, splitting, conditional gradient, mirror descent, coordinate descent, and stochastic gradient-type methods, see, e.g, [3,11,24,25,31,32]. While discussing them all is out of scope of this paper, we focus on some strategies such as penalty, alternating direction, and primal-dual methods which most relate to our work proposed in this paper.

Our approach and related work: The approach in this paper relies on a novel combination of the quadratic penalty [19,34], alternating direction [3,46], homotopy [42], and Nesterov's accelerated methods [2,31,47]. The quadratic penalty method is a classical optimization scheme to handle constrained problems, and can be found in classical text books, e.g., [19,34]. It is often used in nonlinear optimization, and has recently been studied in first-order convex optimization methods, see [26,29]. This method is often inefficient if it stands alone. In this paper, we combine it with other ideas and show that it is indeed more efficient. Our second idea is to use the alternating strategy dated back from the work of J. von Neumann [9], but has recently become extremely popular, see, e.g., [9,18,21,23,38,36]. We exploit this old technique to split the coupling constraint $Ax + By - c \in \mathcal{K}$ and the proximal operator of $f + g$ into each individual one regarding x and y. However, the key idea is perhaps Nesterov's acceleration scheme [31] and the homotopy strategy in [42] that allow us to accelerate the convergence rate of our methods as well as to automatically update the penalty parameter without tuning.

In the context of primal-dual frameworks, our algorithms work on the primal problem (1) and also have convergence guarantees on this problem. Hence, they are different from primal-dual methods such as Chambolle-Pock's scheme [11], alternating minimization (AMA) [21,46], and alternating direction methods of multipliers (ADMM) [17,12,9,21,36]. Note that primal-dual algorithms, AMA, and ADMM are classical methods and their convergence guarantees were proved in many early works, e.g., [17,12,46]. Nevertheless, their convergence rate and iteration-complexity have only recently been studied under different assumptions including strong convexity, Lipschitz gradient continuity, and error bound-type conditions, see, e.g., [11,15,14,16,21,23,38] and the references quoted therein.

Existing state-of-the-art primal-dual methods often achieve the best known $\mathcal{O}\left(\frac{1}{k}\right)$-rate without strong convexity and Lipschitz gradient, where k is the iteration counter. However, such a rate is often obtained via an ergodic sense or a weighted averaging sequence [11,15,14,16,23,38,36]. Under a stronger condition such as either strong convexity or Lipschitz gradient, one can achieve the best known

$\mathcal{O}\left(\frac{1}{k^2}\right)$-convergence rate as shown in, e.g., [11,15,14,16,36].[1] A recent work by Xu [50] showed that ADMM methods can achieve the $\mathcal{O}\left(\frac{1}{k^2}\right)$ convergence rate requiring only the strong convexity on one objective term (either f or g). Such a rate is achieved via weighted averaging sequences. This is fundamentally different from the fast ADMM studied in [21]. Note that the $\mathcal{O}\left(\frac{1}{k^2}\right)$ rate is also attained in AMA methods [21] under the same assumption. Nevertheless, this rate is on the dual problem, and can be viewed as FISTA [4] applying to the dual problem of (1). To the best of our knowledge, the $\mathcal{O}\left(\frac{1}{k^2}\right)$ on the primal problem has not been shown yet. Recently, we proposed two algorithms in [42] to solve (1) that achieve $\mathcal{O}\left(\frac{1}{k}\right)$ convergence rate without any strong convexity or Lipschitz gradient continuity. Moreover, our guarantee for the first algorithm is given in a non-ergodic sequence.

While using the quadratic penalty method as in [26,29] to handle the constraints, our approach in this paper is fundamentally different from [26], where we apply the alternating scheme to decouple the joint variable $z = (x, y)$ and treat them alternatively between x and y. We also exploit the homotopy strategy in [42] to automatically update the penalty parameter instead of fixing or tuning as in [26, 29]. In terms of theoretical guarantee, [26] characterized the iteration-complexity by appropriately choosing a set of parameters depending on the desired accuracy and the feasible set diameters, while [29] assumed that the subproblem could be solved by Nesterov's schemes up to a certain accuracy. Our guarantee does not use any of these techniques, which avoids their drawbacks. Our methods are also different from AMA or ADMM where we do not require Lagrange multipliers, but rather stay in the primal space of (1). In fact, our idea is closely related to the alternating linearization methods in [20], but is still essentially different. We handle the constrained problem (1) directly and update the penalty parameter. We also do not require the smoothness of f and g. Our algorithm is also different from the dual smoothing methods in [27] and [5], where they simply added a proximity function to the primal objective to obtain a Lipschitz gradient dual function, and applied Nesterov's accelerated schemes. These methods work on the dual space.

In terms of structure assumption, our first algorithm achieves the same $\mathcal{O}\left(\frac{1}{k}\right)$-rate as in [11,15,14,16,23,38,36] without any assumption except for the existence of a saddle point. Moreover, the rate of convergence is on the last iterate, which is important for sparse and low-rank optimization. Under a partial strong convexity, i.e., either f or g is strongly convex, our second method can accelerate up to the $\mathcal{O}\left(\frac{1}{k^2}\right)$-convergence rate aka [50], but it has certain advantages compared to [50]. First, it is a primal method without Lagrange multipliers. Second, it linearizes the penalty term in the y-subproblem (see Algorithm 2 for details), which reduces the per-iteration complexity. Third, it either takes averaging only on the y-sequence or uses its last iterate with one additional proximal operator of g. Finally, the y-averaging sequence is weighted.

Our contribution: Our contribution can be summarized as follows:

(a) We propose a new proximal alternating penalty algorithm called PAPA to solve the generic constrained convex problem (1). We show that, under the existence of a saddle point, our method achieves the best known $\mathcal{O}\left(\frac{1}{k}\right)$ convergence rate on both the objective residual and the feasibility violation on (1) without strong convexity, Lipschitz gradient continuity, and the boundedness of the domain of f and g. Moreover, our guarantee is attained on the primal iterate

[1] A recent work in [1] showed an $o\left(\frac{1}{k}\right)$ or $o\left(\frac{1}{k^2}\right)$ rate depending on problem structures.

vectors without averaging. In addition, we allow one to linearize the penalty term in the second subproblem of y (see Step 5 of Algorithm 1 below) that significantly reduces the per-iteration complexity. We also flexibly update all the algorithmic parameters using analytical update rules.

(b) If one objective term of (1) is strongly convex (i.e., either f or g is strongly convex), then we propose a new variant that combines both Nesterov's optimal scheme (or FISTA) [4, 30] and Tseng's variants [2, 47] to solve (1). We prove that this variant can achieve up to $\mathcal{O}\left(\frac{1}{k^2}\right)$-convergence rate on both the objective residual and the feasibility violation. Such a rate is attained by either averaging only on the y-sequence or using one additional proximal operator of f/g.

(c) We customize our algorithms to obtain new variants for solving (1) and its extensions and special cases including the sum of three objective terms, and unconstrained composite convex problems. Some of these variants are new. We also interpret our algorithms as new variants of the primal-dual first-order method. As a byproduct, these variants preserve the same convergence rate as in the proposed algorithms. We also discuss restarting strategies for our methods to significantly improve their practical performance. The convergence guarantee of this strategy will be presented in our forthcoming work [40].

Let us clarify the following points of our contribution. First, although our convergence guarantee is $\mathcal{O}\left(\frac{1}{k}\right)$, it is the best known so far for (1) under only the convexity and the existence of a saddle point. Moreover, the non-ergodic rate is very important for sparse and low-rank optimization since averaging often destroys the sparsity or low-rankness. Second, the linearization of the y-subproblem in Algorithm 1 and Algorithm 2 is useful when A is an orthogonal operator. This allows us to only use the proximal operator of both f and g and significantly reduces the per-iteration complexity compared to classical AMA and ADMM. Third, when applying our method to a composite convex problem, we obtain new variants which are different from existing work. Finally, we allow one to handle general constraints in $\mathcal{K}$ without shifting the problem into linear equality constraints. This is very convenient to handle inequality constraints, convex cones, or boxed constraints as long as the projection onto $\mathcal{K}$ is efficient to compute.

Paper organization: The rest of this paper is organized as follows. Section 2 recalls the dual problem of (1), a fundamental assumption, and its optimality condition. It also defines the quadratic penalty function for (1) and proves a key lemma. Section 3 presents the main contribution with two algorithms and their convergence analysis. Section 4 deals with some extensions and variants of these two methods. Section 5 provides several numerical examples to illustrate our theoretical development and compares with existing methods. For clarity of exposition, all technical proofs are deferred to Appendix A.

2 Preliminaries: Duality, optimality condition, and quadratic penalty

We first define the dual problem of (1) and recall its optimality condition. Then, we define the quadratic penalty function for (1) and prove a key lemma on the objective residual and the feasibility violation.

2.1 Basic notation

We work on finite dimensional Euclidean spaces, $\mathbb{R}^p$, equipped with a standard inner product $\langle\cdot,\cdot\rangle$ and a norm $\|\cdot\| := \langle\cdot,\cdot\rangle^{1/2}$. Given a nonempty, closed and convex set $\mathcal{K} \subseteq \mathbb{R}^n$, we use $\mathcal{N}_{\mathcal{K}}(x)$ to denote its normal cone at x, use $\mathrm{ri}\,(\mathcal{K})$ for its relative interior, and define $\mathcal{K}^{\circ} := \{x \in \mathbb{R}^n \mid \langle x, u\rangle \leq 1,\ u \in \mathcal{K}\}$ for its polar set. We also use $\delta_{\mathcal{K}}(\cdot)$ and $s_{\mathcal{K}}(\cdot)$ to denote its indicator and support function, respectively. If $\mathcal{K}$ is a cone, then $\mathcal{K}^* := \{x \in \mathbb{R}^n \mid \langle x, u\rangle \geq 0,\ u \in \mathcal{K}\}$ stands for its dual cone. Given a proper, closed and convex function f, $\mathrm{dom}(f)$ denotes its domain, $\partial f(\cdot)$ is its subdifferential, $f^*(y) := \sup_x \{\langle y, x\rangle - f(x)\}$ is its Fenchel conjugate, and

$$\mathrm{prox}_{\gamma f}(x) := \arg\min_{u \in \mathbb{R}^p} \left\{ f(u) + \tfrac{1}{2\gamma}\|u - x\|^2 \right\}, \tag{2}$$

is called its the proximal operator, where $\gamma > 0$. In this case, we have

$$\mathrm{prox}_{\gamma f}(x) + \gamma \mathrm{prox}_{f^*/\gamma}(x/\gamma) = x, \tag{3}$$

which is called the Moreau's identity. We say that f is L_f-Lipschitz gradient if it is differentiable, and its gradient ∇f is Lipschitz continuous on its domain with the Lipschitz constant $L_f \in [0, +\infty)$. We say that f is μ_f-strongly convex if $f(\cdot) - \frac{\mu_f}{2}\|\cdot\|^2$ is convex, where $\mu_f > 0$ is its strong convexity parameter. Without loss of generality, we assume that f is μ_f-strongly convex with $\mu_f \geq 0$ to cover also convex functions. For the detail of these notions, we refer the reader to [3, 37].

2.2 Dual problem, fundamental assumption, and KKT condition

Let us define the Lagrange function associated with (1) as

$$\mathcal{L}(x, y, r, \lambda) := f(x) + g(y) - \langle Ax + By - r - c, \lambda\rangle,$$

where λ is the vector of Lagrange multipliers. The dual function is defined as

$$d(\lambda) := \max_{(x,y) \in \mathrm{dom}(F)} \{\langle Ax + By - c, \lambda\rangle - f(x) - g(y)\} = f^*(A^\top\lambda) + g^*(B^\top\lambda) - \langle c, \lambda\rangle,$$

where $\mathrm{dom}(F) := \mathrm{dom}(f) \times \mathrm{dom}(g)$, and f^* and g^* are the Fenchel conjugates of f and g, respectively. The dual problem of (1) is

$$D^\star := \min_{\lambda \in \mathbb{R}^n} \left\{ D(\lambda) := d(\lambda) + s_{\mathcal{K}}(-\lambda) \equiv f^*(A^\top\lambda) + g^*(B^\top\lambda) - \langle c, \lambda\rangle + s_{\mathcal{K}}(-\lambda) \right\}, \tag{4}$$

where $s_{\mathcal{K}}(v) := \sup\{\langle v, r\rangle \mid r \in \mathcal{K}\}$ is the support function of $\mathcal{K}$. If $\mathcal{K}$ is a nonempty, closed, and convex cone, then (4) reduces to

$$D^\star := \min_{\lambda \in -\mathcal{K}^*} \left\{ D(\lambda) := f^*(A^\top\lambda) + g^*(B^\top\lambda) - \langle c, \lambda\rangle \right\},$$

where $\mathcal{K}^*$ is the dual cone of $\mathcal{K}$.

We say that a point $(x^\star, y^\star, r^\star, \lambda^\star) \in \mathrm{dom}(f) \times \mathrm{dom}(g) \times \mathcal{K} \times \mathbb{R}^n$ is a saddle point of the Lagrange function $\mathcal{L}$ if for $(x, y) \in \mathrm{dom}(F)$, $r \in \mathcal{K}$ and $\lambda \in \mathbb{R}^n$, one has

$$\mathcal{L}(x^\star, y^\star, r^\star, \lambda) \leq \mathcal{L}(x^\star, y^\star, r^\star, \lambda^\star) \leq \mathcal{L}(x, y, r, \lambda^\star). \tag{5}$$

We denote by $\mathcal{S}^\star := \{(x^\star, y^\star, r^\star, \lambda^\star)\}$ the set of saddle points of $\mathcal{L}$ satisfying (5), $\mathcal{Z}^\star := \{(x^\star, y^\star)\}$, and by $\Lambda^\star := \{\lambda^\star\}$ the set of the multipliers $\lambda^\star$.

In this paper, we rely on the following mild assumption.

Assumption 1 *Both functions f and g are proper, closed and convex, and $\mathcal{K}$ is a nonempty, closed and convex set in $\mathbb{R}^n$. The set of saddle points $\mathcal{S}^\star$ of $\mathcal{L}$ is nonempty, and the optimal value $F^\star$ is finite and is attainable at some $(x^\star, y^\star) \in \mathcal{Z}^\star$.*

We assume that Assumption 1 holds throughout this paper without recalling it in the sequel.

The optimality condition (or the KKT condition) of (1) can be written as

$$0 \in \partial f(x^\star) - A^\top\lambda^\star,\quad 0 \in \partial g(y^\star) - B^\top\lambda^\star,\quad \lambda^\star \in \mathcal{N}_{\mathcal{K}}(Ax^\star + By^\star - c), \tag{6}$$

where $\mathcal{N}_{\mathcal{K}}(\cdot)$ is the normal cone of $\mathcal{K}$. Let us assume that the following Slater condition holds:

$$\mathrm{ri}\left(\mathrm{dom}(F)\right) \cap \{(x,y) \mid Ax + By - c \in \mathrm{ri}\left(\mathcal{K}\right)\} \neq \emptyset.$$

Then the optimality condition (6) is necessary and sufficient for the strong duality of (1) and (4) to hold, i.e., $F^\star + D^\star = 0$, and the dual solution is attainable and $\Lambda^\star$ is bounded, see, e.g., [8].

2.3 Quadratic penalty function and its properties

Let us define the quadratic penalty function Φ_ρ for the constrained problem (1) as

$$\Phi_\rho(z) := f(x) + g(y) + \rho\psi(x,y), \quad \text{where} \;\; \psi(x,y) := \frac{1}{2}\mathrm{dist}_{\mathcal{K}}\big(Ax + By - c\big)^2, \tag{7}$$

and $z := (x,y)$, and $\rho > 0$ is a penalty parameter. Let us denote by $\mathrm{proj}_{\mathcal{K}}(\cdot)$ the projection operator onto $\mathcal{K}$. Then, we can write $\psi(\cdot)$ in (7) as

$$\psi(x,y) := \frac{1}{2}\min_{r\in\mathcal{K}}\|r - (Ax + By - c)\|^2 = \frac{1}{2}\|Ax + By - c - \mathrm{proj}_{\mathcal{K}}\left(Ax + By - c\right)\|^2.$$

From the definition of Φ_ρ, we have the following result, whose proof is similar to [42, Lemma 1]; however, we provide here a short proof for completeness.

Lemma 1 *Let $\Phi_\rho(\cdot)$ be the quadratic penalty function defined by (7), and $S_\rho(z) := \Phi_\rho(z) - F^\star$. Then, for any $z = (x,y) \in \mathbb{R}^p$, and $\lambda^\star \in \Lambda^\star$, we have*

$$\begin{cases} -\|\lambda^\star\|\mathrm{dist}_{\mathcal{K}}\big(Ax + By - c\big) & \le F(z) - F^\star \le S_\rho(z) - \frac{\rho}{2}\mathrm{dist}_{\mathcal{K}}\big(Ax + By - c\big)^2, \\ \mathrm{dist}_{\mathcal{K}}\big(Ax + By - c\big) & \le \frac{1}{\rho}\left[\|\lambda^\star\| + \sqrt{\|\lambda^\star\|^2 + 2\rho S_\rho(z)}\right], \end{cases} \tag{8}$$

where $\|\lambda^\star\|^2 + 2\rho S_\rho(z) \ge \frac{\rho^2}{2}\|Ax + By - c - \mathrm{proj}_{\mathcal{K}}\left(Ax + By - c\right) + \frac{1}{\rho}\lambda^\star\|^2 \ge 0.$

Proof Since $F(z^\star) = \mathcal{L}(z^\star, r^\star, \lambda^\star) \le \mathcal{L}(z, r, \lambda^\star) = F(z) - \langle\lambda^\star, Ax + By - r - c\rangle$ holds for any $r \in \mathcal{K}$ due to (5), using $r = r^* = \mathrm{proj}_{\mathcal{K}}\left(Ax + By - c\right)$, and $S_\rho(\cdot)$, we obtain

$$\begin{aligned} S_\rho(z) - \tfrac{\rho}{2}\mathrm{dist}_{\mathcal{K}}\big(Ax + By - c\big)^2 &= F(z) - F(z^\star) \ge \langle\lambda^\star, Ax + By - r^* - c\rangle \\ &\ge -\|\lambda^\star\|\|Ax + By - c - r^*\| \\ &= -\|\lambda^\star\|\mathrm{dist}_{\mathcal{K}}\big(Ax + By - c\big), \end{aligned} \tag{9}$$

which is the first inequality of (8). Next, since $\frac{\rho}{2}\|u - r^*\|^2 + \frac{1}{2\rho}\left\|\lambda^\star\right\|^2 + \langle\lambda^\star, u\rangle \ge \frac{\rho}{2}\|u - r^* + \frac{1}{\rho}\lambda^\star\| \ge 0$ for $u = Ax + By - c$, we obtain

$$\tfrac{\rho}{2}\mathrm{dist}_{\mathcal{K}}\big(Ax{+}By{-}c\big)^2 + \tfrac{1}{2\rho}\|\lambda^\star\|^2 + \langle\lambda^\star, Ax{+}By{-}c{-}r^*\rangle = \tfrac{\rho}{2}\|Ax{+}By{-}c{-}r^* + \tfrac{1}{\rho}\lambda^\star\|^2 \ge 0.$$

Summing up this estimate and the first inequality of (9), we obtain

$$S_\rho(z) + \tfrac{1}{2\rho}\|\lambda^\star\|^2 \ge \tfrac{\rho}{2}\|Ax + By - c - \mathrm{proj}_{\mathcal{K}}\left(Ax + By - c\right) + \tfrac{1}{\rho}\lambda^\star\|^2 \ge 0.$$

The second inequality of (8) is a consequence of the first one by solving the following quadratic inequation $\rho t^2 - 2\|\lambda^\star\|t - 2S_\rho(z) \le 0$ in $t \ge 0$. □

3 Proximal Alternating Penalty Algorithms

Our algorithms rely on an alternating strategy applying to the quadratic penalty function $\Phi_\rho(\cdot)$ defined by (7), Nesterov's accelerated scheme [31], and the homotopy strategy in [42]. We first present the main alternating linearization step of the algorithms, and then describe two algorithms in the two subsections, respectively.

3.1 Alternating linearization scheme for the quadratic penalty function

Given $\hat{z} := (\hat{x}, \hat{y}) \in \mathbb{R}^p$ and $\gamma \geq 0$, we consider the following x-subproblem:

$$x_+ \in \mathcal{S}_\gamma(\hat{x}, \hat{y}; \rho) := \arg\min_{x \in \mathbb{R}^{p_1}} \left\{ f(x) + \rho\psi(x, \hat{y}) + \frac{\gamma}{2}\|x - \hat{x}\|^2 \right\}. \tag{10}$$

When $\gamma = 0$, $\mathcal{S}_\gamma(\cdot)$ can be a multivalued mapping. Since $\mathcal{S}_\gamma(\cdot) \neq \emptyset$ for $\gamma > 0$, without loss of generality, we assume that $\mathcal{S}_\gamma(\cdot)$ is nonempty for any $\gamma \geq 0$.

Alternatively, given $x_+ \in \mathbb{R}^{p_1}$, $\hat{y}, \tilde{y} \in \mathbb{R}^{p_2}$, and $\beta > 0$, we consider the y-subproblem:

$$\begin{aligned} y_+ &= \mathrm{prox}_{g/\beta}\big(\tilde{y} - \tfrac{\rho}{\beta}\nabla_y\psi(x_+, \hat{y})\big) \\ &:= \arg\min_{y \in \mathbb{R}^{p_2}} \left\{ g(y) + \rho\langle \nabla_y\psi(x_+, \hat{y}), y - \hat{y}\rangle + \frac{\beta}{2}\|y - \tilde{y}\|^2 \right\}. \end{aligned} \tag{11}$$

Since $\psi(x_+, \cdot)$ is linearized in y, this y-subproblem reduces to evaluating the proximal operator of g at $\tilde{y} - \frac{\rho}{\beta}\nabla_y\psi(x_+, \hat{y})$ as defined by (2). Because problem (11) uses x_+ computed from (10), we obtain an alternating scheme between x and y.

3.2 PAPA for non-strongly convex problems

We first consider the case where both f and g in (1) are neither necessarily strongly convex (i.e., both μ_f and μ_g can be zero) nor Lipschitz gradient continuous.

3.2.1 The algorithm

We call our first method the Proximal Alternating Penalty Algorithm (PAPA), which is presented in Algorithm 1 below.

Algorithm 1 (***P**roximal **A**lternating **P**enalty **A**lgorithm* - Nonstrong convexity)

1: **Initialization:**
2: Choose $x^0 \in \mathbb{R}^{p_1}$, $y^0 \in \mathbb{R}^{p_2}$, $\rho_0 > 0$, and $\gamma_0 \geq 0$.
3: Initialize $\beta_0 := \|B\|^2\rho_0$, $\hat{x}^0 := x^0$, and $\hat{y}^0 := y^0$.
4: **For** $k := 0$ **to** $k_{\max}$ **perform**
5: Update $\begin{cases} x^{k+1} & \in \mathcal{S}_{\gamma_k}(\hat{x}^k, \hat{y}^k; \rho_k), \\ y^{k+1} & := \mathrm{prox}_{g/\beta_k}\big(\hat{y}^k - \frac{\rho_k}{\beta_k}\nabla_y\psi(x^{k+1}, \hat{y}^k)\big), \\ (\hat{x}^{k+1}, \hat{y}^{k+1}) & := (x^{k+1}, y^{k+1}) + \frac{k}{k+2}(x^{k+1} - x^k, y^{k+1} - y^k). \end{cases}$
6: Update $\rho_{k+1} := (k+2)\rho_0$, $\gamma_{k+1} := \frac{(\gamma_k + \mu_f)(k+2)}{\rho_0\|B\|^2(k+1) + \mu_g}$, and $\beta_{k+1} := \|B\|^2\rho_{k+1}$.
7: **End for**

Before analyzing the convergence of Algorithm 1, we make the following comments:

(a) First, Algorithm 1 adopts the idea of Nesterov's first accelerated method in [4,30] to accelerate the penalized problem $\min_{x,y}\Phi_\rho(x, y)$ studied, e.g., in [26, 34]. However, it first alternates between x and y to decouple the quadratic penalty term $\psi(x,y)$ compared to [26]. Next, it linearizes the second subproblem (11) in y to use prox_g. Finally, it is combined with the homotopy strategy in [42] to update the penalty parameter ρ so that its last iterate sequence $\{(x^k,y^k)\}$ converges to a solution $(x^\star,y^\star)$ of the original problem (1).

(b) Second, if the x-subproblem (10) with $\gamma_k=0$, i.e.:

$$x^{k+1}\in\mathcal{S}(\hat{y}^k;\rho_k):=\operatorname*{argmin}_x\left\{f(x)+\rho_k\psi(x,\hat{y}^k)\right\} \tag{12}$$

is solvable (not necessarily unique, e.g., when $\mathrm{dom}(f)$ is compact or A is orthogonal), then the main step, Step 5, in Algorithm 1 reduces to

$$\begin{cases} x^{k+1}\in\mathcal{S}(\hat{y}^k;\rho_k),\\ y^{k+1}:=\mathrm{prox}_{g/\beta_k}\left(\hat{y}^k-\frac{\rho_k}{\beta_k}\nabla_y\psi(x^{k+1},\hat{y}^k)\right),\\ \hat{y}^{k+1}:=y^{k+1}+\frac{k}{k+2}(y^{k+1}-y^k).\end{cases} \tag{13}$$

In this case, the term $\|x^0-x^\star\|$ disappears in the bounds of Theorem 1 below. If $A=\mathbb{I}$, the identity operator, then the two first steps of (13) becomes

$$x^{k+1}:=\mathrm{prox}_{f/\rho_k}\left(c-B\hat{y}^k\right)\quad\text{and}\quad y^{k+1}:=\mathrm{prox}_{g/\beta_k}\left(\hat{y}^k-\tfrac{\rho_k}{\beta_k}\nabla_y\psi(x^{k+1},\hat{y}^k)\right),$$

which only require the proximal operator of f and g.

(c) Third, the gradient $\nabla_y\psi(x^{k+1},\hat{y}^k)$ is computed explicitly as

$$\nabla_y\psi(x^{k+1},\hat{y}^k)=B^\top(u^k-\mathrm{proj}_{\mathcal{K}}(u^k)),\quad\text{where}\quad u^k:=Ax^{k+1}+B\hat{y}^k-c,$$

which requires matrix-vector products Ax, By, and $B^\top u$ each, and one projection onto $\mathcal{K}$. When $\mathcal{K}$ is a simple set (e.g., box, cone, or simplex), the cost of computing $\mathrm{proj}_{\mathcal{K}}$ is minor. As a special case, if $A=\mathbb{I}$, the identity operator, then the per-iteration complexity of Algorithm 1 consists of one prox_f, one prox_g, one By, one $B^T\lambda$ and one $\mathrm{proj}_{\mathcal{K}}(\cdot)$.

(d) Fourth, if $\mu_f=\mu_g=0$, then $\omega=1$, $\gamma_{k+1}=(k+2)\gamma_0$, and the $\mathcal{O}\left(\frac{1}{k}\right)$-convergence rate of Algorithm 1 stated in Theorem 1 below remains unchanged.

(e) Fifth, the convergence guarantee in Theorem 1 is on the last iterate (x^k,y^k) (i.e., without averaging) compared to, e.g., [11,15,23,38].

(f) Sixth, the update rule of ρ_k, γ_k, and β_k at Step 6 is not heuristically tuned. The choice of ρ_0 trades off the feasibility and the objective residual in the bound (14) below. In our implementation, we choose $\rho_0:=\frac{1}{\|B\|}$ by default.

(g) Finally, for Algorithm 1, we can also linearize the subproblem (10) at Step 5 to obtain the following closed form solution using the proximal operator of f:

$$x^{k+1}:=\mathrm{prox}_{f/\hat{\gamma}_k}\left(\hat{x}^k-\tfrac{\rho_k}{\hat{\gamma}_k}\nabla_x\psi(\hat{x}^k,\hat{y}^k)\right).$$

In this case, the analysis of this variant is similar to [42, Theorem 3], but in the alternating manner between x and y. We omit the detail analysis in this paper.

We highlight that Algorithm 1 is different from alternating linearization method in [20], alternating minimization (AMA) [46], and alternating direction methods of multipliers (ADMM) in the literature [23,36,50] as discussed in the introduction.

3.2.2 Convergence analysis

The convergence of Algorithm 1 is presented as follows.

Theorem 1 *Let $\{(x^k, y^k)\}$ be the sequence generated by Algorithm 1 for solving* (1)*. Then, for $k \geq 1$, we have*

$$\begin{cases} \left|F(z^k) - F^\star\right| & \leq \dfrac{\max\{R_p^2,\ \|\lambda^\star\| R_d\}}{2k}, \\ \mathrm{dist}_{\mathcal{K}}\left(Ax^k + By^k - c\right) & \leq \dfrac{R_d}{\rho_0 k}, \end{cases} \tag{14}$$

where $R_p^2 := \gamma_0\|x^0 - x^\star\|^2 + \rho_0\|B\|^2\|y^0 - y^\star\|^2$ and $R_d := \|\lambda^\star\| + \sqrt{\|\lambda^\star\|^2 + \rho_0 R_p^2}$. Consequently, the convergence rate of Algorithm 1 is $\mathcal{O}\left(\frac{1}{k}\right)$, i.e., $|F(z^k) - F^\star| \leq \mathcal{O}\left(\frac{1}{k}\right)$ and $\mathrm{dist}_{\mathcal{K}}\left(Ax^k + By^k - c\right) \leq \mathcal{O}\left(\frac{1}{k}\right)$ even when $\mu_f = \mu_g = 0$.

The proof of Theorem 1 requires the following key lemma, whose proof can be found in Appendix A.3.

Lemma 2 *Let $\{(x^k, y^k, \hat{x}^k, \hat{y}^k)\}$ be the sequence generated by Algorithm 1. Then, $(\hat{x}^k, \hat{y}^k)$ can be interpreted as*

$$\begin{aligned} &(\hat{x}^k, \hat{y}^k) &&= (1-\tau_k)(x^k, y^k) + \tau_k(\tilde{x}^k, \tilde{y}^k), \\ \textit{with}\ \ &(\tilde{x}^{k+1}, \tilde{y}^{k+1}) &&:= (\tilde{x}^k, \tilde{y}^k) + \tfrac{1}{\tau_k}(x^{k+1} - \hat{x}^k, y^{k+1} - \hat{y}^k), \end{aligned} \tag{15}$$

and $(\tilde{x}^0, \tilde{y}^0) := (x^0, y^0)$, where $\tau_k := \frac{1}{k+1} \in (0, 1]$. Moreover, Φ_ρ defined by (7) *satisfies*

$$\begin{aligned} \Phi_{\rho_k}(z^{k+1}) \leq\ & (1-\tau_k)\Phi_{\rho_{k-1}}(z^k) + \tau_k F(z^\star) + \tfrac{\gamma_k\tau_k^2}{2}\|\tilde{x}^k - x^\star\|^2 \\ & - \tfrac{(\gamma_k+\mu_f)\tau_k^2}{2}\|\tilde{x}^{k+1} - x^\star\|^2 + \tfrac{\beta_k\tau_k^2}{2}\|\tilde{y}^k - y^\star\|^2 - \tfrac{(\beta_k+\mu_g)\tau_k^2}{2}\|\tilde{y}^{k+1} - y^\star\|^2 \\ & - \tfrac{(\beta_k - \|B\|^2\rho_k)}{2}\|y^{k+1} - \hat{y}^k\|^2 - \tfrac{(1-\tau_k)}{2}\left[\rho_{k-1} - \rho_k(1-\tau_k)\right]\|s^k\|^2, \end{aligned} \tag{16}$$

where $s^k := Ax^k + By^k - c - \mathrm{proj}_{\mathcal{K}}\left(Ax^k + By^k - c\right)$.

Proof (The proof of Theorem 1) The update rules $\tau_k := \frac{1}{k+1}$ and $\rho_k := (k+1)\rho_0$ from Algorithm 1 show that

$$\tau_0 := 1, \quad \tau_k = \frac{\tau_{k-1}}{1 + \tau_{k-1}}, \quad \text{and} \quad \rho_k = \frac{\rho_{k-1}}{1 - \tau_k}.$$

These relations also lead to

$$\frac{(1-\tau_k)}{\rho_k\tau_k^2} = \frac{1}{\rho_{k-1}\tau_{k-1}^2}, \quad \rho_{k-1} = (1-\tau_k)\rho_k, \quad \text{and} \quad \beta_k := \|B\|^2\rho_k.$$

Let $\nu_k := 1 + \frac{\mu_g}{\rho_0\|B\|^2(k+1)} \geq 1$. If we update γ_k as $\gamma_{k+1} = \frac{(\gamma_k+\mu_f)(k+2)}{\rho_0\|B\|^2(k+1)+\mu_g}$, then

$$\frac{\nu_k\gamma_{k+1}}{k+2} = \left(1 + \frac{\mu_g}{\rho_0\|B\|^2(k+1)}\right)\frac{\gamma_{k+1}}{k+2} = \frac{\gamma_k + \mu_f}{k+1}.$$

Using these four equalities, we obtain from (16) that

$$\begin{aligned} (k+1)S_{\rho_k}(z^{k+1}) + \nu_k\left[\tfrac{\gamma_{k+1}}{2(k+2)}\|\tilde{x}^{k+1} - x^\star\|^2 + \tfrac{\rho_0\|B\|^2}{2}\|\tilde{y}^{k+1} - y^\star\|^2\right] &\leq kS_{\rho_{k-1}}(z^k) \\ + \left[\tfrac{\gamma_k}{2(k+1)}\|\tilde{x}^k - x^\star\|^2 + \tfrac{\rho_0\|B\|^2}{2}\|\tilde{y}^k - y^\star\|^2\right]. \end{aligned}$$

Let us denote by $S_k := S_{\rho_{k-1}}(z^k)$, $a_k := \frac{\gamma_k}{2(k+1)}\|\tilde{x}^k - x^\star\|^2 + \frac{\rho_0\|B\|^2}{2}\|\tilde{y}^k - y^\star\|^2$. Then, the last estimate can be simplified as

$$(k+1)S_{k+1} + a_{k+1} \le (k+1)S_{k+1} + \nu_k a_{k+1} \le kS_k + a_k.$$

By induction, we can easily show that $S_{k+1} \le \frac{a_0}{k+1}$, which leads to

$$S_{\rho_k}(z^{k+1}) \le \frac{1}{2(k+1)}\left[\gamma_0\|\tilde{x}^0 - x^\star\|^2 + \rho_0\|B\|^2\|\tilde{y}^0 - y^\star\|^2\right]. \qquad (17)$$

Using this estimate into Lemma 1 and note that $\rho_k = \rho_0(k+1)$, $\tilde{x}^0 = \hat{x}^0 = x^0$, and $\tilde{y}^0 = \hat{y}^0 = y^0$, we obtain (14). □

Remark 1 We can replace the update of $\tau_k := \frac{1}{k+1}$ in (15) and $\rho_k := (k+1)\rho_0$ in Algorithm 1 by imposing a tighter condition:

$$\|B\|^2\rho_k\tau_k^2 = (\|B\|^2\rho_{k-1} + \mu_g)\tau_{k-1}^2(1-\tau_k) \quad \text{and} \quad \rho_k(1-\tau_k) = \rho_{k-1}.$$

In this case, we obtain the following update rule for τ_k and ρ_k:

$$\tau_k := \frac{\tau_{k-1}\sqrt{1+\nu_k}}{1+\tau_{k-1}\sqrt{1+\nu_k}} \quad \text{and} \quad \rho_k := \frac{\rho_{k-1}}{1-\tau_k}, \quad \text{where} \quad \nu_k := \frac{\mu_g}{\|B\|^2\rho_{k-1}}.$$

However, as we will prove in Appendix A.5, it does not theoretically improve the $\mathcal{O}\left(\frac{1}{k}\right)$ rate of Algorithm 1 in Theorem 1.

3.3 PAPA for the strong convexity of the g objective term

In Algorithm 1, we have not been able to prove a better convergence rate than $\mathcal{O}\left(\frac{1}{k}\right)$ when g is strongly convex. In this subsection, we propose a new algorithm that allows us to exploit the strong convexity of g in order to improve the convergence rate from $\mathcal{O}\left(\frac{1}{k}\right)$ to $\mathcal{O}\left(\frac{1}{k^2}\right)$. This algorithm can be viewed as a hybrid between Tseng's accelerated proximal gradient [47], and Nesterov's scheme in [30].

3.3.1 The algorithm

The details of the algorithm are presented in Algorithm 2 below.

Before analyzing the convergence of Algorithm 2, we make the following comments.

(a) Similar to Algorithm 1, when the x-subproblem (12) is solvable, we do not need to add the regularization term $\frac{\gamma_k}{2}\|x - \hat{x}^k\|^2$. In this case, the term $\|x^0 - x^\star\|$ also disappears in the convergence bound (18) of Theorem 2 below.

(b) The update of τ_k at Step 5 is standard in accelerated methods. Indeed, if we define $t_k := \frac{1}{\tau_k}$, then we obtain the well-known Nesterov update rule [30] for t_k as $t_{k+1} = 0.5(1 + (1+4t_k^2)^{1/2})$ with $t_0 := 1$. However, as shown in our proof below, we can update τ_k and ρ_k based on the following tighter conditions:

$$\frac{\|B\|^2\rho_k\tau_k^2}{1-\tau_k} = \|B\|^2\rho_{k-1}\tau_{k-1}^2 + \mu_g\tau_{k-1} \quad \text{and} \quad \rho_k = \frac{\rho_{k-1}}{1-\tau_k}.$$

These conditions lead to a new update rule for ρ_k and τ_k as

$$\tau_k := \frac{\sqrt{\tau_{k-1}^2 + \kappa\tau_{k-1}/\rho_{k-1}}}{1+\sqrt{\tau_{k-1}^2 + \kappa\tau_{k-1}/\rho_{k-1}}}, \quad \text{and} \quad \rho_k := \frac{\rho_{k-1}}{1-\tau_k},$$

Algorithm 2 (***P****roximal* ***A****lternating* ***P****enalty* ***A****lgorithm* - Strong convexity)

1: **Initialization:**

2: Choose $x^0 \in \mathbb{R}^{p_1}$, $y^0 \in \mathbb{R}^{p_2}$, $\rho_0 \in \left(0, \frac{\mu_g}{2\|B\|^2}\right]$, and $\gamma_0 \geq 0$.

3: Initialize $\tau_0 := 1$, $\beta_0 := \|B\|^2\rho_0$, $\hat{x}^0 := x^0$, and $\tilde{y}^0 := y^0$.

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

5: Update $\tau_{k+1} := \frac{\tau_k}{2}\left(\sqrt{\tau_k^2 + 4} - \tau_k\right)$.

6: Update
$$\begin{cases} \hat{y}^k & := (1-\tau_k)y^k + \tau_k\tilde{y}^k, \\ x^{k+1} & \in \mathcal{S}_{\gamma_k}(\hat{x}^k, \hat{y}^k; \rho_k), \\ \hat{x}^{k+1} & := x^{k+1} + \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}(x^{k+1} - x^k), \\ \tilde{y}^{k+1} & := \mathrm{prox}_{g/(\tau_k\beta_k)}\big(\tilde{y}^k - \frac{\rho_k}{\tau_k\beta_k}\nabla_y\psi(x^{k+1}, \hat{y}^k)\big). \end{cases}$$

7: Perform **one** of the following two steps:

Choice 1: $y^{k+1} := (1-\tau_k)y^k + \tau_k\tilde{y}^{k+1}$ (Averaging step).

Choice 2: $y^{k+1} := \mathrm{prox}_{g/\beta_k}\big(\hat{y}^k - \frac{\rho_k}{\beta_k}\nabla_y\psi(x^{k+1}, \hat{y}^k)\big)$ (Proximal step).

8: Update $\rho_{k+1} := \frac{\rho_k}{1-\tau_{k+1}}$, $\beta_{k+1} := \|B\|^2\rho_{k+1}$, and $\gamma_{k+1} := \gamma_k + \mu_f$.

9: **End for**

where $\kappa := \frac{\mu_g}{\|B\|^2}$. In this case, we still has the same guarantee as in Theorem 2.

(c) The update of ρ_k at Step 8 is as the same as in Algorithm 1. But the update of γ_k is different. When $\mu_f = 0$, we can fix $\gamma_k := \gamma_0 > 0$ for all $k \geq 0$.

(d) To achieve the $\mathcal{O}\left(\frac{1}{k^2}\right)$-convergence rate, we only require the strong convexity on one objective term, i.e., $\mu_g > 0$. In addition, we can compute $\{y^k\}$ with averaging as in **Choice 1** or with one additional proximal operator prox_g of g as in **Choice 2**. For **Choice 1**, the weighted averaging sequence is only taken on $\{y^k\}$ but not on $\{x^k\}$. This is different from a recent work in [50], where the same convergence rate of ADMM is obtained for $\mu_g > 0$. We emphasize that Algorithm 2 is fundamentally different from [50] as shown in the introduction. The $\mathcal{O}\left(\frac{1}{k^2}\right)$ rate was also known for AMA [21], but the guarantee is on the dual problem (4). To achieve the same rate on (1), an extra step is required, see [44].

(e) The strong convexity of g can be relaxed to a quasi-strong convexity as studied in [28], where we assume that there exists $\mu_g > 0$ such that

$$g(y) + \langle\nabla g(y), y^\star - y\rangle + \frac{\mu_g}{2}\|y - y^\star\|^2 \leq g(y^\star), \quad \forall y \in \mathrm{dom}(g),\ y^\star \in \mathcal{Y}^\star,$$

where $\mathcal{Y}^\star$ is the projection of the primal solution set $\mathcal{Z}^\star$ onto y, and $\nabla g(y) \in \partial g(y)$. As shown in [28], this condition is weaker than the strong convexity of g.

3.3.2 Convergence analysis

We prove the following convergence result for Algorithm 2.

Theorem 2 *Let $\{(x^k, y^k)\}$ be the sequence generated by Algorithm 2 for solving* (1). *Then*

$$\begin{cases} |F(z^k) - F^{\star}| & \leq \dfrac{2\max\left\{R_p^2,\ 2\|\lambda^{\star}\|R_d\right\}}{(k+1)^2}, \\ \mathrm{dist}_{\mathcal{K}}\big(Ax^k + By^k - c\big) & \leq \dfrac{4R_d}{\rho_0(k+1)^2}, \end{cases} \tag{18}$$

where $R_p^2 := \gamma_0\|x^0 - x^{\star}\|^2 + \rho_0\|B\|^2\|y^0 - y^{\star}\|^2$ and $R_d := \|\lambda^{\star}\| + \sqrt{\|\lambda^{\star}\|^2 + \rho_0 R_p^2}$. Consequently, the convergence rate of Algorithm 2 is $\mathcal{O}\left(\frac{1}{k^2}\right)$, i.e., $|F(z^k) - F^{\star}| \leq \mathcal{O}\left(\frac{1}{k^2}\right)$ and $\mathrm{dist}_{\mathcal{K}}\big(Ax^k + By^k - c\big) \leq \mathcal{O}\left(\frac{1}{k^2}\right)$.

To prove Theorem 2 we need the following lemma (*cf.* Appendix A.4).

Lemma 3 *Let $\{(x^k, y^k, \hat{x}^k, \hat{y}^k, \tilde{y}^k)\}$ be the sequence generated by Algorithm 2. Then, $\hat{x}^k$ can be interpreted as*

$$\hat{x}^k = (1-\tau_k)x^k + \tau_k\tilde{x}^k, \quad \text{with } \tilde{x}^0 := x^0, \quad \text{and } \tilde{x}^{k+1} := \tilde{x}^k + \tfrac{1}{\tau_k}(x^{k+1} - \hat{x}^k). \tag{19}$$

Moreover, the following estimate holds:

$$\begin{aligned} \Phi_{\rho_k}(z^{k+1}) \leq{} & (1-\tau_k)\Phi_{\rho_{k-1}}(z^k) + \tau_k F(z^{\star}) + \tfrac{\gamma_k\tau_k^2}{2}\|\tilde{x}^k - x^{\star}\|^2 \\ & - \tfrac{(\gamma_k + \mu_f)\tau_k^2}{2}\|\tilde{x}^{k+1} - x^{\star}\|^2 - \tfrac{\gamma_k\tau_k^2}{2}\|\tilde{x}^{k+1} - \tilde{x}^k\|^2 + \tfrac{\beta_k\tau_k^2}{2}\|\tilde{y}^k - y^{\star}\|^2 \\ & - \tfrac{\beta_k\tau_k^2 + \mu_g\tau_k}{2}\|\tilde{y}^{k+1} - y^{\star}\|^2 - \tfrac{(\beta_k - \rho_k\|B\|^2)\tau_k^2}{2}\|\tilde{y}^{k+1} - \tilde{y}^k\|^2 \\ & - \tfrac{(1-\tau_k)}{2}\left[\rho_{k-1} - \rho_k(1-\tau_k)\right]\|s^k\|^2, \end{aligned} \tag{20}$$

where $s^k := Ax^k + By^k - c - \mathrm{proj}_{\mathcal{K}}\big(Ax^k + By^k - c\big)$.

Proof (The proof of Theorem 2) For simplicity of notation, we denote by $S_k := S_{\rho_{k-1}}(z^k) = \Phi_{\rho_{k-1}}(z^k) - F^{\star}$. Assume that ρ_k and β_k are updated by $\rho_{k-1} = \rho_k(1-\tau_k)$, and $\beta_k = \|B\|^2\rho_k$. Then, we can simplify (20) as follows:

$$\begin{aligned} S_{k+1} &+ \tfrac{(\gamma_k+\mu_f)\tau_k^2}{2}\|\tilde{x}^{k+1} - x^{\star}\|^2 + \tfrac{(\|B\|^2\rho_k\tau_k^2 + \mu_g\tau_k)}{2}\|\tilde{y}^{k+1} - y^{\star}\|^2 \leq (1-\tau_k)S_k \\ &+ \tfrac{\gamma_k\tau_k^2}{2}\|\tilde{x}^k - x^{\star}\|^2 + \tfrac{\|B\|^2\rho_k\tau_k^2}{2}\|\tilde{y}^k - y^{\star}\|^2. \end{aligned}$$

Let us assume that the parameters τ_k, ρ_k and γ_k are updated such that

$$\frac{\|B\|^2\rho_k\tau_k^2}{1-\tau_k} \leq \|B\|^2\rho_{k-1}\tau_{k-1}^2 + \mu_g\tau_{k-1} \quad \text{and} \quad \frac{\gamma_k\tau_k^2}{1-\tau_k} \leq (\gamma_{k-1} + \mu_f)\tau_{k-1}^2. \tag{21}$$

Then, we can write the above inequality as

$$\begin{aligned} S_{k+1} &+ \tfrac{(\gamma_k+\mu_f)\tau_k^2}{2}\|\tilde{x}^{k+1} - x^{\star}\|^2 + \tfrac{(\|B\|^2\rho_k\tau_k^2 + \mu_g\tau_k)}{2}\|\tilde{y}^{k+1} - y^{\star}\|^2 \leq (1-\tau_k)\Big[S_k \\ &+ \tfrac{(\gamma_{k-1}+\mu_f)\tau_{k-1}^2}{2}\|\tilde{x}^k - x^{\star}\|^2 + \tfrac{\left(\|B\|^2\rho_{k-1}\tau_{k-1}^2 + \mu_g\tau_{k-1}\right)}{2}\|\tilde{y}^k - y^{\star}\|^2\Big]. \end{aligned}$$

Hence, if we define $A_k := S_k + \frac{(\gamma_{k-1}+\mu_f)\tau_{k-1}^2}{2}\|\tilde{x}^k - x^{\star}\|^2 + \frac{\left(\|B\|^2\rho_{k-1}\tau_{k-1}^2 + \mu_g\tau_{k-1}\right)}{2}\|\tilde{y}^k - y^{\star}\|^2$, then, we have $A_{k+1} \leq (1-\tau_k)A_k$. By induction, we obtain

$$A_{k+1} \leq \omega_k\left[(1-\tau_0)S_0 + \tfrac{\gamma_0\tau_0^2}{2}\|\tilde{x}^0 - x^{\star}\|^2 + \tfrac{\|B\|^2\rho_0\tau_0^2}{2}\|\tilde{y}^0 - y^{\star}\|^2\right], \tag{22}$$

where $\omega_k := \prod_{i=1}^k (1-\tau_i)$.

Now, we assume that we update τ_k and ρ_k as follows:

$$\tau_0 = 1, \quad \tau_k := \frac{\tau_{k-1}}{2}\left(\sqrt{\tau_{k-1}^2 + 4} - \tau_{k-1}\right), \text{ and } \rho_k := \frac{\rho_{k-1}}{1-\tau_k} = \frac{\rho_{k-1}\tau_{k-1}^2}{\tau_k^2}.$$

This update leads to $1 - \tau_k = \frac{\tau_k^2}{\tau_{k-1}^2}$. By induction and $\tau_0 = 1$, we can show that $\frac{1}{k+1} \leq \tau_k \leq \frac{2}{k+2}$. Moreover, we also have $\omega_k = \prod_{i=1}^k (1-\tau_i) = \prod_{i=1}^k \frac{\tau_i^2}{\tau_{i-1}^2} = \frac{\tau_k^2}{\tau_0^2} = \tau_k^2$. Since $\rho_k = \frac{\rho_{k-1}}{1-\tau_k}$, by induction, we obtain $\rho_k = \frac{\rho_0}{\tau_k^2}$.

Next, we find the condition on ρ_0 such that the first condition of (21) holds. Indeed, using $1-\tau_k = \frac{\tau_k^2}{\tau_{k-1}^2}$ and $\rho_k = \frac{\rho_0}{\tau_k^2}$, this condition is equivalent to

$$\|B\|^2 \rho_0 \frac{\tau_{k-1}}{\tau_k} \leq \mu_g.$$

Clearly, since $1 \leq \frac{\tau_{k-1}}{\tau_k} \leq 2$, if $2\|B\|^2\rho_0 \leq \mu_g$, then $\|B\|^2\rho_0\frac{\tau_{k-1}}{\tau_k} \leq \mu_g$ holds. This condition is equivalent to $\rho_0 \leq \frac{\mu_g}{2\|B\|^2}$.

The second condition of (21) holds if we choose $\gamma_k \leq \frac{(\gamma_{k-1}+\mu_f)\tau_{k-1}^2(1-\tau_k)}{\tau_k^2} = \gamma_{k-1} + \mu_f$. In this case, since $\tau_0 = 1$, $\tilde{x}^0 = x^0$ and $\tilde{y}^0 = y^0$, (22) leads to

$$S_{\rho_k}(z^{k+1}) \leq \frac{\tau_k^2}{2}\left[\gamma_0\|x^0 - x^\star\|^2 + \rho_0\|B\|^2\|y^0 - y^\star\|^2\right].$$

Using this, $\rho_k = \frac{\rho_0}{\tau_k^2}$, and $\frac{1}{k+1} \leq \tau_k \leq \frac{2}{k+2}$ into Lemma 1, we obtain (18). □

4 Variants and extensions

Algorithms 1 and 2 can be customized into different variants. Let us provide some examples on how to customize these algorithms to handle instances of (1).

4.1 Application to composite convex minimization

Let us consider the following composite convex problem

$$P^\star := \min_{y\in\mathbb{R}^p}\left\{P(y) := f(y) + g(y)\right\}, \tag{23}$$

where $f : \mathbb{R}^p \to \mathbb{R}\cup\{+\infty\}$ and $g : \mathbb{R}^p \to \mathbb{R}\cup\{+\infty\}$ are proper, closed and convex. Let us introduce $x = y$ and write (23) as (1) with $F(z) := f(x)+g(y)$ and $x-y = 0$.

Now we apply Algorithm 1 to solve the resulting problem, and obtain

$$x^{k+1} := \mathrm{prox}_{f/\rho_k}(\hat{y}^k) \;\text{ and }\; y^{k+1} := \mathrm{prox}_{g/\beta_k}\big(\hat{y}^k - \tfrac{\rho_k}{\beta_k}(\hat{y}^k - x^{k+1})\big).$$

Plugging the first expression into the second one and using the fact that $\beta_k = \|B\|^2\rho_k = \rho_k$ in the update rule of Step 6, we get

$$y^{k+1} := \mathrm{prox}_{g/\beta_k}\big(\mathrm{prox}_{f/\beta_k}(\hat{y}^k)\big). \tag{24}$$

Hence, we obtain the following scheme to solve (23):

$$\begin{cases} y^{k+1} := \mathrm{prox}_{g/\beta_k}\big(\mathrm{prox}_{f/\beta_k}(\hat{y}^k)\big) \;\text{ with } \beta_k := \beta_0(k+1),\\ \hat{y}^{k+1} := y^{k+1} + \frac{k}{k+2}\big(y^{k+1} - y^k\big). \end{cases} \tag{25}$$

Here, $\beta_0 > 0$ is an initial value. This scheme was studied in [45].

Similarly, when g is μ_g-strongly convex with $\mu_g > 0$, we can apply Algorithm 2 to solve (23). Let us consider **Choice 1**. Then, after eliminating $\{x^k\}$, we obtain

$$\begin{cases} \hat{y}^k &:= (1-\tau_k)y^k + \tau_k\tilde{y}^k, \\ \tilde{y}^{k+1} &:= \mathrm{prox}_{g/(\tau_k\beta_k)}\left(\frac{1}{\tau_k}\mathrm{prox}_{f/\beta_k}(\hat{y}^k) - \frac{(1-\tau_k)}{\tau_k}y^k)\right), \\ y^{k+1} &:= (1-\tau_k)y^k + \tau_k\tilde{y}^{k+1}. \end{cases} \tag{26}$$

Here, $\beta_k := \frac{\beta_0}{\tau_k^2}$ for $\beta_0 \in (0, \frac{\mu_g}{2}]$, and $\tau_0 := 1$ and $\tau_{k+1} := 0.5\tau_k(\sqrt{\tau_k^2+4} - \tau_k)$. The following corollary provides the convergence rate of these two variants, whose proof can be found in Appendix A.6.

Corollary 1 *Assume that f is Lipschitz continuous with the Lipschitz constant $L_f \in [0, +\infty)$, i.e., $|f(y) - f(\hat{y})| \le L_f\|y - \hat{y}\|$ for all $y, \hat{y} \in \mathrm{dom}(f)$. Let $\{y^k\}$ be generated by (25) to solve (23). Then, we have*

$$P(y^k) - P^\star \le \frac{\rho_0\|y^0 - y^\star\|^2}{2k} + \frac{2L_f^2 + \sqrt{2}L_f\rho_0\|y^0 - y^\star\|}{\rho_0(k+1)}. \tag{27}$$

If f is L_f-Lipschitz continuous on $\mathrm{dom}(f)$ and g is μ_g-strongly convex, then $\{y^k\}$ generated by (26) to solve the composite convex minimization problem (23) satisfies:

$$P(y^k) - P^\star \le \frac{2\rho_0\|y^0 - y^\star\|^2}{(k+1)^2} + \frac{8L_f^2 + 4\sqrt{2}L_f\rho_0\|y^0 - y^\star\|}{\rho_0(k+2)^2}. \tag{28}$$

Note that we can use **Choice 2** to replace the averaging on y^k by prox_g. In this case, we still have the same guarantee as in (28), but the scheme (26) is slightly changed. We can also eliminate y^k in Algorithm 2 instead of x^k. In this case, the convergence guarantee is on $\{x^k\}$ and it requires g to be L_g-Lipschitz continuous instead of f. The proof is rather similar and we skip the details in this paper.

4.2 Application to composite convex minimization with linear operator

We tackle a more general form of (23) by considering the following problem:

$$P^\star := \min_{y\in\mathbb{R}^p}\left\{P(y) := f(By) + g(y)\right\}, \tag{29}$$

where $f : \mathbb{R}^p \to \mathbb{R}\cup\{+\infty\}$ and $g : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ are proper, closed and convex, and $B \in \mathbb{R}^{n\times p}$ is a linear operator.

Using the same trick by introducing $x = By$, we obtain $F(x, y) = f(x) + g(y)$ and a linear constraint $x - By = 0$. Now we apply Algorithm 1 to solve the resulting problem, and obtain

$$x^{k+1} := \mathrm{prox}_{f/\rho_k}\left(B\hat{y}^k\right) \ \text{ and } \ y^{k+1} := \mathrm{prox}_{g/\beta_k}\left(\hat{y}^k - \tfrac{\rho_k}{\beta_k}B^\top(B\hat{y}^k - x^{k+1})\right).$$

Plugging the first expression into the second one and using the fact that $\beta_k = \|B\|^2\rho_k$ in the update rule of Step 6, we get

$$\begin{cases} y^{k+1} := \mathrm{prox}_{g/(\|B\|^2\rho_k)}\left((\mathbb{I} - \frac{1}{\|B\|^2}B^\top B)\hat{y}^k + \frac{1}{\|B\|^2}B^\top\mathrm{prox}_{f/\rho_k}\left(B\hat{y}^k\right)\right), \\ \hat{y}^{k+1} := y^{k+1} + \frac{k}{k+2}(y^{k+1} - y^k). \end{cases} \tag{30}$$

Similarly, we can also customize Algorithm 2 to solve (29) when g is μ_g-strongly convex as:

$$\begin{cases} \tilde{y}^{k+1} := \mathrm{prox}_{g/(\tau_k\beta_k)}\big(\tilde{y}^k - \frac{\rho_k}{\tau_k\beta_k}B^\top(B\hat{y}^k - \mathrm{prox}_{f/\rho_k}\big(B\hat{y}^k\big))\big), \\ y^{k+1} := (1-\tau_k)y^k + \tau_k\tilde{y}^{k+1}, \\ \hat{y}^{k+1} := y^{k+1} + \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}(y^{k+1} - y^k). \end{cases} \tag{31}$$

The convergence of (30) and (31) can be proved as in Corollary 1 under the Lipschitz continuity of f. We omit the details here.

4.3 A primal-dual interpretation of Algorithm 1 and Algorithm 2

We show that Algorithms 1 and 2 can be interpreted as a primal-dual method for solving (29). We consider the x-subproblem (10) with $\gamma = 0$ as:

$$x^{k+1} := \mathrm{prox}_{f/\rho_k}\big(B\hat{y}^k\big) \overset{(3)}{=} B\hat{y}^k - \frac{1}{\rho_k}\mathrm{prox}_{\rho_k f^*}\big(\rho_k B\hat{y}^k\big). \tag{32}$$

Let $\bar{x}^{k+1} := \mathrm{prox}_{\rho_k f^*}\big(\rho_k B\hat{y}^k\big)$. Then, by using (32) and a notation $\dot{x} := 0^{p_1}$, we can rewrite Algorithm 1 for solving (29) as

$$\begin{cases} \bar{x}^{k+1} := \mathrm{prox}_{\rho_k f^*}\big(\dot{x} + \rho_k B\hat{y}^k\big), \\ y^{k+1} := \mathrm{prox}_{g/\beta_k}\big(\hat{y}^k - \frac{1}{\beta_k}B^\top\bar{x}^{k+1}\big), \\ \hat{y}^{k+1} := y^{k+1} + \frac{k}{k+2}(y^{k+1} - y^k). \end{cases} \tag{33}$$

This scheme can be considered as a new primal-dual method for solving (29), and it is different from existing methods in the literature.

Similarly, we can also interpret Algorithm 2 with **Choice 1** as a primal-dual variant. Using the same idea as above, we can arrive at

$$\begin{cases} \bar{x}^{k+1} := \mathrm{prox}_{\rho_k f^*}\big(\dot{x} + \rho_k B\hat{y}^k\big), \\ \tilde{y}^{k+1} := \mathrm{prox}_{g/(\tau_k\beta_k)}\big(\tilde{y}^k - \frac{1}{\tau_k\beta_k}B^\top\bar{x}^{k+1}\big), \\ y^{k+1} := (1-\tau_k)y^k + \tau_k\tilde{y}^{k+1}, \\ \hat{y}^{k+1} := y^{k+1} + \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}(y^{k+1} - y^k). \end{cases} \tag{34}$$

The convergence guarantee of both schemes (33) and (34) can be proved as in Corollary 1 under the L_f-Lipschitz continuity assumption of f. We again omit the detail analysis here.

4.4 Extension to the sum of three objectives

Let us consider the following constrained convex optimization problem:

$$F^\star := \min_{z:=[x,y]}\left\{F(z) := f(x) + g(y) + h(y) \mid Ax + By - c \in \mathcal{K}\right\}, \tag{35}$$

where f, g, A, B, c and $\mathcal{K}$ are defined as in (1), and $h : \mathbb{R}^{p_2} \to \mathbb{R}$ is convex and Lipschitz gradient with the Lipschitz constant $L_h > 0$. In this case, we modify the y-subproblem in Algorithm 1 as

$$y^{k+1} := \mathrm{prox}_{g/\hat{\beta}_k}\left(\hat{y}^k - \frac{1}{\hat{\beta}_k}\left(\nabla h(\hat{y}^k) + \rho_k\nabla_y\psi(x^{k+1},\hat{y}^k)\right)\right), \tag{36}$$

where $\hat{\beta}_k := \beta_k + L_h$. Other steps remain as in Algorithm 1.

When either g is μ_g-strongly convex or h is μ_h-strongly convex such that $\mu_g + \mu_h > 0$, we can applied Algorithm 2 to solve (35). In this case, the y-subproblem in Algorithm 2 becomes

$$\tilde{y}^{k+1} := \mathrm{prox}_{g/(\tau_k\hat{\beta}_k)}\left(\tilde{y}^k - \tfrac{1}{\tau_k\hat{\beta}_k}\left(\nabla h(\bar{y}^k) + \rho_k\nabla_y\psi(x^{k+1},\hat{y}^k)\right)\right). \quad (37)$$

Here, $\bar{y}^k$ is chosen such that $\bar{y}^k := \tilde{y}^k$ if $\mu_g + 2\mu_h - L_h > 0$, or $\bar{y}^k := \hat{y}^k$ if $\mu_g > 0$. In addition, if we use **Choice 2**, then we compute y^{k+1} as

$$y^{k+1} := \mathrm{prox}_{g/\hat{\beta}_k}\left(\hat{y}^k - \tfrac{1}{\hat{\beta}_k}\left(\nabla h(\bar{y}^k) + \rho_k\nabla_y\psi(x^{k+1},\hat{y}^k)\right)\right). \quad (38)$$

Other steps remain the same as in Algorithm 2.

The following corollary shows the convergence of these variants, whose proof can be found in Appendix A.7.

Corollary 2 *Let $\{(x^k, y^k)\}$ be the sequence generated by Algorithm 1 to solve* (35) *using* (36) *for y^k and $\hat{\beta}_k := \rho_k\|B\|^2 + L_h$. Then the bound* (14) *in Theorem 1 still holds with $R_p^2 := \gamma_0\|x^0 - x^\star\|^2 + (L_h + \rho_0\|B\|^2)\|y^0 - y^\star\|^2$.*

Assume that either g is μ_g-strongly convex or h is μ_h-strongly convex such that $\mu_g + \mu_h > 0$. Let $\{(x^k, y^k)\}$ be the sequence generated by Algorithm 2 to solve (35) *using* (37) *for y^k so that:*

(i) If $\mu_g > 0$, then we choose $\bar{y}^k := \hat{y}^k$ and $0 < \rho_0 \le \frac{\mu_g}{2\|B\|^2}$, and update $\hat{\beta}_k := \rho_k\|B\|^2 + L_h$.

(ii) If $L_h < 2\mu_h$, then we choose $\bar{y}^k := \tilde{y}^k$ and $0 < \rho_0 \le \frac{\mu_g + 2\mu_h - L_h}{2\|B\|^2}$, and update $\hat{\beta}_k := \rho_k\|B\|^2 + \frac{L_h}{\tau_k}$.

Then the bound (18) *in Theorem 2 still holds with $R_p^2 := \gamma_0\|x^0 - x^\star\|^2 + (L_h + \rho_0\|B\|^2)\|y^0 - y^\star\|^2$.*

Note that we can extend Algorithms 1 and 2 to handle the case where $f(x)$ is replaced by $f(x) + h(x)$, where h is convex and L_h-Lipschitz gradient.

4.5 Shifting the initial dual variable and restarting

As we can see from (8) of Lemma 1 that the bound on $\mathrm{dist}_{\mathcal{K}}(Ax + By - c)$ depends on $\|\lambda^\star\|$ instead of $\|\lambda^\star - \lambda^0\|$ from an initial dual variable λ^0. We use the idea of "restarting the prox-center point" from [41] to adaptively update λ^0. This idea has recently been used in [42, 48] as a restarting strategy and it has significantly improve the performance of the algorithms.

The main idea is to replace φ defined by (46) by

$$\varphi_\rho(u;\lambda^0) := \max_{\lambda\in\mathbb{R}^n}\min_{r\in\mathcal{K}}\left\{\langle u - r, \lambda\rangle - \tfrac{\rho}{2}\|\lambda - \lambda^0\|^2\right\} = \tfrac{\rho}{2}\mathrm{dist}_{\mathcal{K}}\big(u + \tfrac{1}{\rho}\lambda\big).$$

and redefine $\psi(\cdot,\cdot)$ in (7) by $\psi_\rho(x, y;\lambda^0) := \varphi_\rho(Ax + By - c;\lambda^0) = \frac{\rho}{2}\mathrm{dist}_{\mathcal{K}}\big(Ax + By - c + \frac{1}{\rho}\lambda^0\big)^2$. Then the main steps (10)-(11) of Algorithm 1 or Algorithm 2 become

$$\begin{aligned}
&x^{k+1} \in \operatorname*{argmin}_x\left\{f(x) + \psi_{\rho_k}(x, \hat{y}^k;\lambda^0)\right\},\\
&y^{k+1} := \operatorname*{argmin}_y\left\{g(y) + \langle\nabla_y\psi_{\rho_k}(x^{k+1}, \hat{y}^k;\lambda^0), y - \hat{y}^k\rangle + \tfrac{\beta_k}{2}\|y - \hat{y}^k\|^2\right\}, \quad (39)\\
\text{or}\quad &\tilde{y}^{k+1} := \operatorname*{argmin}_y\left\{g(y) + \langle\nabla_y\psi_{\rho_k}(x^{k+1}, \hat{y}^k;\lambda^0), y - \hat{y}^k\rangle + \tfrac{\beta_k\tau_k}{2}\|y - \tilde{y}^k\|^2\right\},
\end{aligned}$$

Our strategy is to frequently update λ^0 and restart the algorithms as follows. We perform k_s steps (e.g., $k_s = 100$) starting from $k := 0$ to $k := k_s - 1$, and restart the variables by resetting:

$$\rho_{k_s} := \rho_0, \quad \tau_{k_s} := 1, \quad \hat{y}^{k_s} := y^{k_s}, \text{ and } \lambda^0 := \lambda^0 + \nabla\varphi_{\rho_{k_s}}(Ax^{k_s+1} + B\hat{y}^{k_s} - c; \lambda^0).$$

Since proving the convergence of this variant is out of scope of this paper, we refer to our forthcoming work [40] for the full theory of restarting.

5 Numerical experiments

In the following numerical examples, we focus on the following problem template:

$$F^\star := \min_{y\in\mathbb{R}^p} \left\{F(y) := f(By) + g(y) + h(y)\right\}, \tag{40}$$

where f and g are convex and possibly nonsmooth, h is convex and L_h-Lipschitz gradient, and B is a linear operator. If we introduce $x := By$ and let $h = 0$, then the objective of (40) becomes $F(z) := f(x) + g(y)$ with an additional constraint $-x + By = 0$. Hence, (40) can be converted into (1). Otherwise, it becomes (35). We implement 9 algorithms to solve (40) as follows:

- Algorithm 1, denoted by `PAPA`, and its restarting variant, called `PAPA-rs`.
- Algorithm 2, denoted by `scvx-PAPA` and its restarting variant, called `scvx-PAPA-rs`.
- Algorithm 1 in [42], `ASGARD`, and its restarting variant, denoted by `ASGARD-rs`.
- The Chambolle-Pock algorithm in [11] and Vu-Condat's method in [13,49].
- The accelerated proximal gradient method, denoted by `AcProxGrad`, in [4,33].

These algorithms are implemented in Matlab (R2014b), running on a MacBook Pro. Laptop with 2.7 GHz Intel Core i5, and 16GB memory. Note that the per-iteration complexity of Algorithm 1, Algorithm 2, `ASGARD`, Chambolle-Pock's algorithm, and Vu-Condat's algorithm is essentially the same. For a thorough comparison to between `ASGARD` and other methods, including ADMM, we refer to [42].

For configuration of Algorithms 1 and 2, we choose $\rho_0 := \frac{1}{\|B\|}$ in Algorithm 1 and its variants. We choose $\rho_0 := \frac{\mu_g}{2\|B\|^2}$ in Algorithm 2 and its variants. However, if μ_g is unknown (e.g., problem may not be strongly convex), we choose $\mu_g := 0.1$. Since we consider the case $A = \mathbb{I}$, we set $\gamma_0 := 0$ in all variants of `PAPA`. For restarting variants, we restart `PAPA` after each 50 iterations and `scvx-PAPA` after each 100 iterations as described in Subsection 4.5. For `ASGARD`, we use the same setting as in [42], and for Chambolle-Pock's and Vu-Condat's algorithm, we choose the parameters as suggested in [11,13,49] for both the strongly and nonstrongly convex cases. We also restart `ASGARD` after every each 50 iterations. Our Matlab code is available online at `https://github.com/quoctd/PAPA-s1.0`.

5.1 Dense convex quadratic programs

We consider the following convex quadratic programming problem:

$$g^\star := \min_{y\in\mathbb{R}^{p_2}} \left\{g(y) := \tfrac{1}{2}y^\top Qy + q^\top y \mid a \le By \le b\right\}, \tag{41}$$

where $Q \in \mathbb{R}^{p_2\times p_2}$ is a symmetric positive [semi]definite matrix, $q \in \mathbb{R}^{p_2}$, $B \in \mathbb{R}^{n\times p_2}$ and $a, b \in \mathbb{R}^n$ such that $a \le b$. We assume that both Q and B are dense.

This problem can be reformulated into (1) by introducing a new variable $x := By$ to form the linear constraint $x - By = 0$ and an additional objective term $f(x) := \delta_{[a,b]}(x)$. In this case, we have $\mathcal{K} = \{0\}$.

The main step of both Algorithms 1 and 2 is to solve the two subproblems at Step 5. For (41), these two problems can be solved explicitly as

$$\begin{cases} x^{k+1} := \mathrm{proj}_{[a,b]}\left(B\hat{y}^k\right), \\ y^{k+1} := (\rho_k\|B\|^2\mathbb{I} + Q)^{-1}\Big(\rho_k\|B\|^2\hat{y}^k - \rho_k B^\top(B\hat{y}^k - x^{k+1}) - q\Big). \end{cases}$$

For Algorithm 2, we change from y^{k+1} to $\tilde{y}^{k+1}$, from $\hat{y}^k$ to $\tilde{y}^k$, and from $\rho_k\|B\|^2$ to $\tau_k\rho_k\|B\|^2$ in the second line.

Note that we can write (41) into the following form

$$g^\star := \min_y \left\{ G(y) := \tfrac{1}{2}y^\top Q y + q^\top y + f(By) \right\},$$

where $f(x) := \delta_{[a,b]}(x)$ is the indicator function of the box $[a,b]$. Hence, we can apply the Chambolle-Pock primal-dual algorithm [11] to solve (41).

We test the first 7 algorithms mentioned above on some synthetic data generated as follows. We randomly generate $R \in \mathbb{R}^{p_2\times m}$, $q \in \mathbb{R}^{p_2}$, and $B \in \mathbb{R}^{n\times p_2}$ using the standard Gaussian distribution, where $m = \lfloor p_2/2 \rfloor + 1$. To avoid large magnitudes, we normalize R by $\frac{1}{\sqrt{m}}R$, and B by $\frac{1}{\sqrt{n}}B$. We then define $Q := RR^T + \mu_g\mathbb{I}$, where $\mu_g = 0$ for the nonstrongly convex case and $\mu_g = 1$ for the strongly convex case. We generate a random vector $y^\natural$ using again the standard Gaussian distribution, and define $a := By - \texttt{rand}(n,1)$ and $b := By + \texttt{rand}(n,1)$ to make sure that the problem is feasible, where $\texttt{rand}(n,1)$ is a uniform random vector in $(0,1)^n$.

Figure 1 shows the convergence of 7 algorithms on a strongly convex instance of (41), where $p_2 = 2000$ and $n = 2000$. The left plot shows the convergence of the relative objective residual $\frac{|g(y^k)-g^\star|}{|g^\star|}$, where $g^\star$ is computed by CVX [22] using Mosek with the best accuracy. The right plot reveals the relative feasibility violation $\frac{\|\max\{By^k-b,0\}\|+\|\min\{By^k-a,0\}\|}{\max\{\|a\|,\|b\|\}}$.

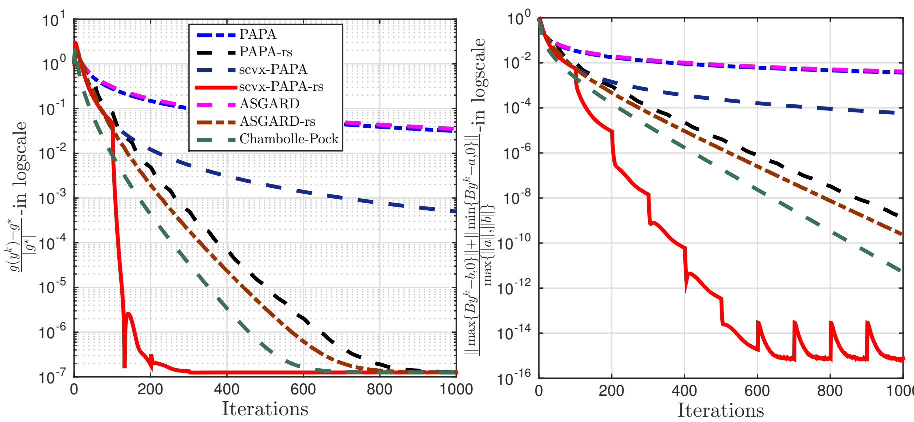

Fig. 1 A comparison of 7 algorithms on a **strongly convex** problem instance of (41) after 1000 iterations. The problem size is ($p_2 = 2000, n = 2000$). Left: The relative objective residual, Right: The relative feasibility violation. Due to Mosek's solution, the relative objective residual is saturated at a 10^{-7} accuracy, while the relative feasibility can reach a 10^{-15} accuracy.

Since the problem is strongly convex, Algorithm 2 shows its $\mathcal{O}\left(\frac{1}{k^2}\right)$ convergence rate as predicted by the theory (Theorem 2), while Algorithm 1 and `ASGARD` still show their $\mathcal{O}\left(\frac{1}{k}\right)$ convergence rate. The Chambolle-Pock algorithm using strong convexity works really well and exhibits beyond the theoretical $\mathcal{O}\left(\frac{1}{k^2}\right)$-rate. The restarting variant of Algorithm 2 completely outperforms the other methods, although the restarting variants of `PAPA` as well as `ASGARD` work well.

Next, we test these algorithms on a nonstrongly convex instance of (41) by setting $\mu_g := 0$. The convergence behavior of these algorithms is plotted in Figure 2.

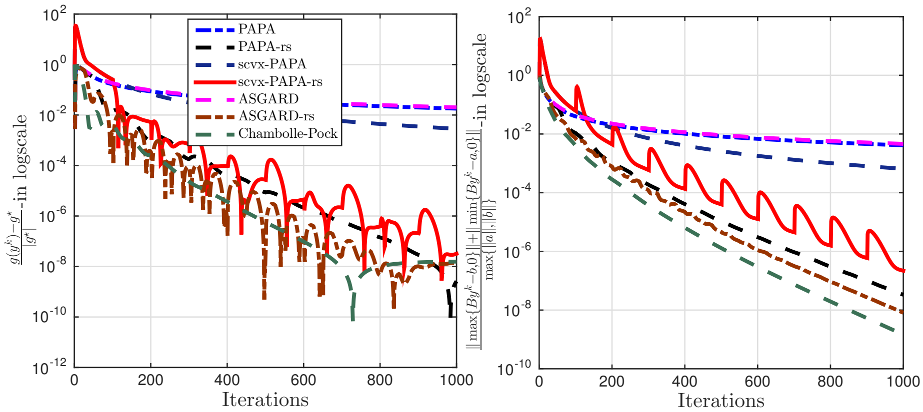

Fig. 2 A comparison of 7 algorithms on a **nonstrongly convex** problem instance of (41) after 1000 iterations. The problem size is $(p_2 = 2000, n = 2000)$. Left: The relative objective residual, Right: The relative feasibility violation.

Since the problem is no longer strongly convex, Algorithm 2 does not guarantee its $\mathcal{O}\left(\frac{1}{k^2}\right)$-rate, but Algorithm 1 still has its $\mathcal{O}\left(\frac{1}{k}\right)$-rate. The restarting variant of Algorithm 2 still improves its theoretical performance, but becomes worse than other restart variants and the Chambolle-Pock method. In this particular instance, `ASGARD` with restarting still works well.

Finally, we verify the restarting variants on the strongly convex problem instance of (41) by choosing different frequencies: $s = 50$ and $s = 100$. The convergence result of this run is plotted in Figure 3.

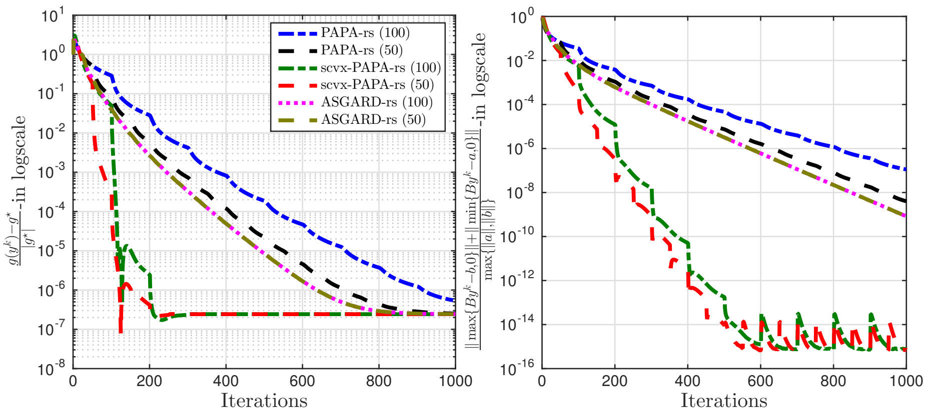

Fig. 3 A comparison of the restarting variants with two different frequencies on a **strongly convex** problem instance of (41) after 1000 iterations. The problem size is $(p_2 = 2000, n = 2000)$. Left: The relative objective residual, Right: The relative feasibility violation.

Figure 3 shows that these two frequencies seem not significantly affecting the performance of the restarting algorithms. For `PAPA`, $s = 50$ slightly works better than $s = 100$. However, we have observed that if we set the frequency s too small, e.g., $s = 10$, then the restarting variants are highly oscillated. If we set it too big, then it does not improve the performance and we need to run with a large number of iterations. In [40], we provide a full theory how to adaptively choose the frequency to guarantee the convergence of restarting `ASGARD` methods.

5.2 The elastic net problem with square-root loss

In this example, we consider the common elastic net LASSO problem studied in [51] but with a square-root loss as follows:

$$F^\star := \min_{y\in\mathbb{R}^{p_2}} \left\{ F(y) := \|By - c\|_2 + \tfrac{\kappa_1}{2}\|y\|^2 + \kappa_2\|y\|_1 \right\}, \tag{42}$$

where $\kappa_1 > 0$ and $\kappa_2 > 0$ are two regularization parameters. Due to the nonsmoothness of the square-root loss $\|By - c\|_2$, this problem is harder to solve than the standard elastic net in [51], and algorithms such as FISTA [4] is not applicable.

By introducing $x := By - c$, we can reformulate (42) into (1) as

$$F^\star := \min_{x,y} \left\{ F(x,y) := \|x\|_2 + \tfrac{\kappa_1}{2}\|y\|^2 + \kappa_2\|y\|_1 \mid -x + By = c \right\}.$$

Since $g(y) := \frac{\kappa_1}{2}\|y\|^2 + \kappa_2\|y\|_1$ is strongly convex, we can apply Algorithm 2 to solve it. By choosing $\gamma_0 = 0$, the two subproblems at Step 6 of Algorithm 2 become:

$$x^{k+1} := \mathrm{prox}_{\|\cdot\|_2/\rho_k}\big(B\hat{y}^k - b\big) \quad\text{and}\quad \tilde{y}^{k+1} := \mathrm{prox}_{\sigma_k\|\cdot\|_1}\big(u^k\big),$$

where $\sigma_k := \frac{\kappa_2}{\kappa_1 + \rho_k\|B\|^2}$ and $u^k := \frac{\|B\|^2\hat{y}^k - B^\top(B\hat{y}^k - x^{k+1} - c)}{\kappa_1/\rho_k + \|B\|^2}$.

In order to apply the Chambolle-Pock method in [11, Algorithm 2], we define $F(By) := \|By - c\|_2$ and $G(y) := \frac{\kappa_1}{2}\|y\|^2 + \kappa_2\|y\|_1$. In this case, G is strongly convex with the parameter $\mu_g = \kappa_2$. Hence, we choose the parameters as suggested in [11, Algorithm 2]. When $\kappa_2 = 0$, i.e., G is non-strongly convex, we use again [11, Algorithm 1] with the parameters $\sigma = \tau = \frac{1}{2\|B\|^2}$ and $\theta = 1$.

We compare again the first 7 algorithms discussed above to solve (42). We generate the data as follows. Matrix $B \in \mathbb{R}^{n\times p_2}$ is generated randomly using standard Gaussian distribution $\mathcal{N}(0,1)$, and then is normalized by $\frac{1}{\sqrt{n}}$, i.e., $B := \frac{1}{\sqrt{n}}\texttt{randn}(n, p_2)$. We generate a sparse vector $y^\natural$ with s-nonzero entries sampling from the standard Gaussian distribution as the true parameter vector. Then, we generate the observed measurement as $c = By^\natural + \sigma\mathcal{N}(0,1)$, where $\sigma = 0$ in the noiseless case, and $\sigma = 10^{-3}$ in the noisy case. We choose $\kappa_1 = 0.1$ and $\kappa_2 = 0.01$ for our test. In this case, we obtain solutions with approximately 2% sparsity.

Figure 4 shows the actual convergence behavior of the two instances of (42) with noise and without noise respectively, in terms of the relative objective residual $\frac{F(y^k) - F^\star}{|F^\star|}$ of (42), where the optimal value $F^\star$ is computed via CVX [22] using Mosek with the best precision.

The theoretical algorithms, i.e., `PAPA` and `ASGARD` [42], still show the $\mathcal{O}\left(\frac{1}{k}\right)$-rate on the original objective residual. But their restarting variants exhibit a much better convergence rate without employing the strong convexity. `ASGARD` with restart performs worse than `PAPA-rs` in this example. If we exploit the convexity

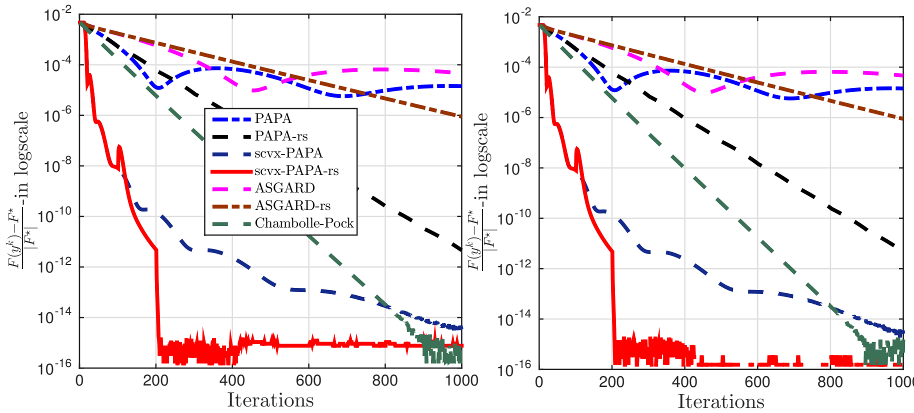

Fig. 4 A comparison of 7 algorithms on the original objective residual $\frac{F(y^k)-F^\star}{|F^\star|}$ of (42) after 1000 iterations. The problem size is $(p_2 = 5000, n = 1750, s = 500)$. Left: **without noise**; Right: **with Gaussian noise** (with variance $\sigma = 10^{-3}$).

as in Algorithm 2, then this algorithm and its restart variant completely outperform other methods. The theoretical version of Algorithm 2 performs significantly well in this example, beyond the theoretical $\mathcal{O}\left(\frac{1}{k^2}\right)$-rate. It even performs better than Chambolle-Pock's method with the strong convexity [11, Algorithm 2]. The restarting variant requires approximately 200 iterations to achieve up to the 10^{-15} accuracy.

5.3 Square-root Lasso

We now show that Algorithm 2 still works well even when the problem is not strongly convex using again (42). In this test, we set $\kappa_1 = 0$, and problem (42) reduces to the common square-root LASSO problem [6]. We test 3 algorithms as above on a new instance of (42) with the size $(p_2 = 5000, n = 1750, s = 500)$, and noise. Since $\kappa_1 = 0$, we do not know if the problem is strongly convex or not. Hence, we select three different values of μ_g in Algorithm 2 and the Chambolle-Pock method as $\mu_g = 1$, $\mu_g = 0.1$ and $\mu_g = 0.01$. Figure 5 shows the result of this test when we restart after each 100 iterations (left), and after 50 iterations (right).

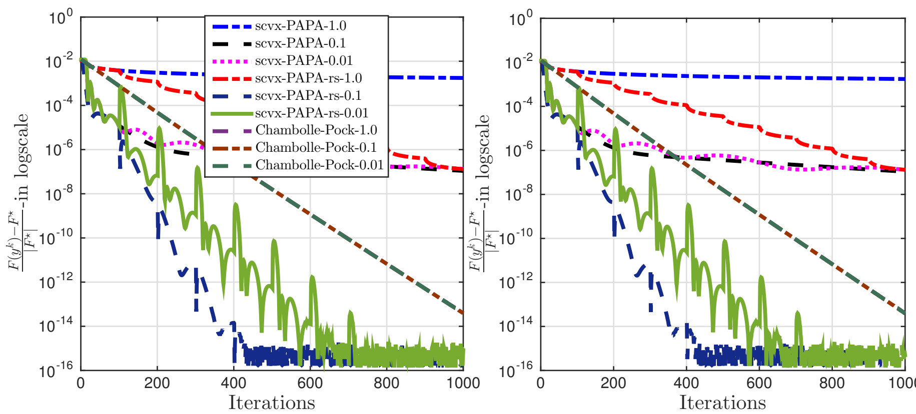

Fig. 5 A comparison of 9 algorithmic variants on the square-root LASSO problem (42) (i.e., $\kappa_1 = 0$) after 1000 iterations. The problem size is $(p_2 = 5000, n = 1750, s = 500)$. Left: Restarting after each 100 iterations; Right: Restarting after each 50 iterations.

Figure 5 shows that Algorithm 2 still has the $\mathcal{O}\left(\frac{1}{k^2}\right)$-rate. The restarting Algorithm 2 with $\mu_g = 0.1$ still outperforms Algorithm 1, and the Chambolle-Pock method with strong convexity. When $\mu_g = 0.01$, it still performs well compared to the Chambolle-Pock method, but if $\mu_g = 1$, then it becomes worse. This is affected by the choice of the initial value $\rho_0 = \frac{\mu_g}{2\|B\|^2}$, which is inappropriate.

5.4 Image reconstruction with low sampling rate

We consider an image reconstruction problem using low sampling rates as:

$$F^\star := \min_{Y\in\mathbb{R}^{m_1\times m_2}} \left\{ F(Y) := \tfrac{1}{2}\|\mathcal{A}(Y) - b\|_F^2 + \kappa\|Y\|_{\mathrm{TV}} \right\}, \tag{43}$$

where $\mathcal{A}$ is a linear operator, b is a measurement vector, $\|\cdot\|_F$ is the Frobenius norm, $\kappa > 0$ is a regularization parameter, and $\|Y\|_{\mathrm{TV}}$ is the total-variation norm.

To apply our methods, we use $\|Y\|_{\mathrm{TV}} = \|D(Y)\|_1$ and reformulate this problem into (35) as

$$\min_{X,Y} \left\{ \kappa\|X\|_1 + \tfrac{1}{2}\|\mathcal{A}(Y) - b\|_F^2 \mid X - D(Y) = 0 \right\},$$

where we choose $f(X) := \kappa\|X\|_1$, $g(Y) := 0$, and $h(Y) := \frac{1}{2}\|\mathcal{A}(Y) - b\|_F^2$ which is Lipschitz gradient continuous with $L_h := \|\mathcal{A}^*\mathcal{A}\|^2$. Although, h and g may not be quasi-strongly convex, we still apply Algorithm 2(b) in Subsection 4.4 to solve it.

We also implement Vu-Condat's algorithm [13,49] and FISTA [4,33] to directly solve (43). For Vu-Condat's algorithm, we implement the following scheme:

$$\begin{cases} \tilde{X}^k & := \mathrm{prox}_{\tau g}\big(X^k - \tau(\nabla h(X^k) + D^*(X^k))\big) \\ \tilde{X}^k & := \mathrm{prox}_{\sigma f}\big(Y^k + \sigma D(2\tilde{X}^k - X^k)\big) \\ (X^{k+1}, Y^{k+1}) & := (1-\theta)(X^k, Y^k) + \theta(\tilde{X}^k, \tilde{Y}^k), \end{cases} \tag{44}$$

where D^* is the adjoint operator of D, $\theta := 1$, and $\tau > 0$ and $\sigma > 0$ satisfying $\frac{1}{\tau} - \sigma\|D\|^2 \geq \frac{L_h}{2}$. The last condition leads $0 < \tau < \frac{2}{L_h}$ and $0 < \sigma \leq \frac{1}{\|D\|^2}\left(\frac{1}{\tau} - \frac{L_h}{2}\right)$. We test Vu-Condat's algorithm using $(\tau, \sigma) := \left(\frac{0.089}{L_h}, \frac{1}{\|D\|^2}\left(\frac{1}{\tau} - \frac{L_h}{2}\right)\right)$ after carefully tuning these parameters.

For Algorithm 1, we set $\rho_0 := \frac{1}{2\|D\|}$, and for Algorithm 2, we set $\rho_0 := \frac{1}{4\|D\|^2}$. We also implement the restarting variants of both algorithms with a frequency of $s = 50$ iterations. For FISTA, we compute the proximal operator $\mathrm{prox}_{\|\cdot\|_{\mathrm{TV}}}$ using a primal-dual method as in [11] by setting the number of iterations at 25 and 50, respectively, and also use a fixed restarting strategy after each 50 iterations [35].

We test these algorithms on 6 MRI images of different sizes downloaded from different websites. We generate the observed measurement b by using subsampling -FFT transformation at the rate of 20%. After tuning the regularization parameter κ, we fix it at $\kappa = 4.0912 \times 10^{-4}$ for all the experiment. Table 1 shows the results of 8 algorithms on these MRI images after 200 iterations in terms of the objective values, computational time, and PSNR (Peak signal-to-noise ratio) [11].

Table 1 shows that our algorithms and Vu-Condat's method outperform FISTA in terms of computational time. This is not surprised since FISTA requires to evaluate an expensive proximal operator of the TV-norm at each iteration. However, it gives a slightly better PSNR while produces a worse objective value than our methods. Vu-Condat's algorithm with tuned parameters has a similar performance

Table 1 The results and performance of 8 algorithms on 6 MRI images

	`Hip` (798 × 802)			`Knee` (779 × 693)			`Brain-tomor` (650 × 650)		
Algorithms	$F(Y^k)$	PSNR	Time[s]	$F(Y^k)$	PSNR	Time[s]	$F(Y^k)$	PSNR	Time[s]
`PAPA`	0.01070	81.56	57.97	0.00840	79.62	48.69	0.01050	77.82	34.63
`PAPA-rs`	0.01056	81.35	57.57	0.00828	79.51	48.69	0.01039	77.72	34.36
`scvx-PAPA`	0.01034	81.24	64.34	0.00805	79.44	53.81	0.01025	77.70	37.98
`scvx-PAPA-rs`	0.01035	81.23	67.47	0.00807	79.45	53.48	0.01026	77.69	37.93
`Vu-Condat-tuned`	0.01030	81.23	56.02	0.00801	79.45	45.92	0.01023	77.70	31.71
`AcProxGrad-25`	0.01179	82.25	1055.94	0.00917	79.78	844.47	0.01133	78.74	674.08
`AcProxGrad-rs`	0.01179	82.25	1052.23	0.00917	79.78	860.97	0.01133	78.74	652.91
`AcProxGrad-50`	0.01104	82.30	2052.68	0.00865	79.83	1652.81	0.01079	78.80	1264.33
	`Body` (895 × 320)			`Confocal` (370 × 370)			`Leg` (588 × 418)		
Algorithms	$F(Y^k)$	PSNR	Time[s]	$F(Y^k)$	PSNR	Time[s]	$F(Y^k)$	PSNR	Time[s]
`PAPA`	0.01674	66.92	22.80	0.02539	67.58	12.12	0.01050	74.50	22.30
`PAPA-rs`	0.01664	66.96	22.70	0.02534	67.60	11.80	0.01040	74.37	22.81
`scvx-PAPA`	0.01653	66.98	25.09	0.02528	67.67	13.28	0.01030	74.36	25.22
`scvx-PAPA-rs`	0.01664	66.98	25.15	0.02529	67.61	13.41	0.01030	74.34	25.83
`Vu-Condat-tuned`	0.01652	66.99	22.99	0.02527	67.74	10.84	0.01028	74.38	20.70
`AcProxGrad-25`	0.01728	67.35	400.36	0.02652	68.97	136.46	0.01104	75.23	361.63
`AcProxGrad-rs`	0.01728	67.35	431.07	0.02652	68.97	132.13	0.01104	75.23	366.83
`AcProxGrad-50`	0.01697	67.39	817.63	0.02639	68.97	256.97	0.01074	75.21	700.66

as our methods. Unfortunately, the restarting variants with a fixed frequency, e.g., $s = 50$, do not significantly improve the performance of all methods in this example. This happens perhaps due to the nonstrong convexity of the problem.

In order to observe the quality of reconstruction, we plot the result of 8 algorithm in Figure 6 for one MRI image (`Hip`) of the size 798×802 (i.e., $p_2 = 639,996$). Clearly, we can see that the quality of the reconstruction is still acceptable with only 20% of the measurement.

Acknowledgments: This work is partly supported by the NSF-grant, DMS-1619884, USA.

A Appendix: The proof of technical results in the main text

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

A.1 Properties of the distance function $\mathrm{dist}_{\mathcal{K}}(\cdot)$.

Our aim is to investigate some necessary properties of ψ defined by (7) to analyze the convergence of Algorithms 1 and 2. We first consider the following distance function:

$$\varphi(u) := \tfrac{1}{2}\mathrm{dist}_{\mathcal{K}}\big(u\big)^2 = \min_{r\in\mathcal{K}} \tfrac{1}{2}\|r-u\|^2 = \tfrac{1}{2}\|r^*(u)-u\|^2 = \tfrac{1}{2}\|\mathrm{proj}_{\mathcal{K}}(u)-u\|^2, \quad (45)$$

where $r^*(u) := \mathrm{proj}_{\mathcal{K}}(u)$ is the projection of u onto $\mathcal{K}$. Clearly, we can write φ in (45) as follows:

$$\varphi(u) = \max_{\lambda\in\mathbb{R}^n}\min_{r\in\mathcal{K}}\left\{\langle u-r,\lambda\rangle - \tfrac{1}{2}\|\lambda\|^2\right\} = \max_{\lambda\in\mathbb{R}^n}\left\{\langle u,\lambda\rangle - s_{\mathcal{K}}(\lambda) - \tfrac{1}{2}\|\lambda\|^2\right\}, \quad (46)$$

where $s_{\mathcal{K}}(\lambda) := \sup_{r\in\mathcal{K}}\langle\lambda, r\rangle$ is the support function of $\mathcal{K}$.

The function φ is convex and differentiable. Its gradient is given by

$$\nabla\varphi(u) = u - \mathrm{proj}_{\mathcal{K}}(u) = \mathrm{proj}_{\mathcal{K}^\circ}(u), \quad (47)$$

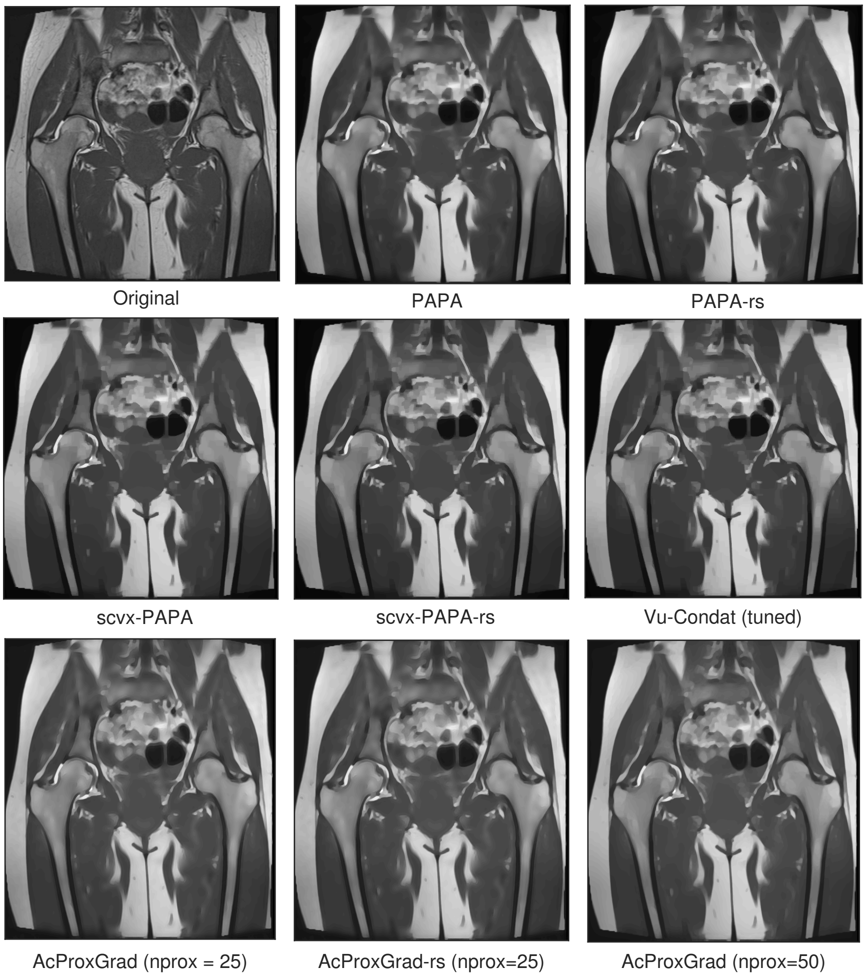

Fig. 6 The original image, and its reconstructions from 8 algorithms using 20% of measurement. Here, `nprox` is the number of iterations required to evaluate the proximal operator of the TV-norm, and Vu-Condat (tuned) is Vu-Condat's method [13, 49] using tuned parameters.

where $\mathcal{K}^{\circ} := \{v \in \mathbb{R}^n \mid \langle u, v\rangle \leq 1,\ u \in \mathcal{K}\}$ is the polar set of $\mathcal{K}$. If $\mathcal{K}$ is a cone, then $\nabla\varphi(u) = \mathrm{proj}_{\mathcal{K}^{\circ}}(u) = \mathrm{proj}_{-\mathcal{K}^*}(u)$, where $\mathcal{K}^* := \{v \in \mathbb{R}^n \mid \langle u, v\rangle \geq 0,\ u \in \mathcal{K}\}$ is the dual cone of $\mathcal{K}$.

By using the property of $\mathrm{proj}_{\mathcal{K}}(\cdot)$, it is easy to prove that $\nabla\varphi(\cdot)$ is Lipschitz continuous with the Lipschitz constant $L_{\varphi} = 1$. Hence, for any $u, v \in \mathbb{R}^n$, we have (see [31]):

$$\begin{aligned} \varphi(u) &+ \langle \nabla\varphi(u), v - u\rangle + \tfrac{1}{2}\|\nabla\varphi(v) - \nabla\varphi(u)\|^2 \leq \varphi(v), \\ \varphi(v) &\leq \varphi(u) + \langle \nabla\varphi(u), v - u\rangle + \tfrac{1}{2}\|v - u\|^2. \end{aligned} \tag{48}$$

Let us recall ψ defined by (7) as

$$\psi(x, y) := \varphi(Ax + By - c) = \tfrac{1}{2}\mathrm{dist}_{\mathcal{K}}\big(Ax + By - c\big)^2. \tag{49}$$

Then, ψ is also convex and differentiable, and its gradient is given by

$$\begin{aligned}\nabla_x\psi(x,y) &= A^\top\left(Ax+By-c-\mathrm{proj}_{\mathcal{K}}\left(Ax+By-c\right)\right),\\ \nabla_y\psi(x,y) &= B^\top\left(Ax+By-c-\mathrm{proj}_{\mathcal{K}}\left(Ax+By-c\right)\right).\end{aligned}\tag{50}$$

These partial gradients satisfy (see [31]):

$$\begin{aligned}\psi(\hat{x},y) &\leq \psi(x,y)+\langle\nabla_x\psi(x,y),\hat{x}-x\rangle+\tfrac{1}{2}\|A(\hat{x}-x)\|^2,\\ \psi(x,\hat{y}) &\leq \psi(x,y)+\langle\nabla_y\psi(x,y),\hat{y}-y\rangle+\tfrac{1}{2}\|B(\hat{y}-y)\|^2.\end{aligned}\tag{51}$$

For given $x_+\in\mathbb{R}^{p_1}$ and $\hat{y}\in\mathbb{R}^{p_2}$, let us define the following linear function

$$\ell(x,y;x_+,\hat{y}) := \psi(x_+,\hat{y})+\langle\nabla_x\psi(x_+,\hat{y}),x-x_+\rangle+\langle\nabla_y\psi(x_+,\hat{y}),y-\hat{y}\rangle.\tag{52}$$

Then, the following lemma provides two properties of $\ell(\cdot)$.

Lemma 4 *Let $x^\star\in\mathbb{R}^{p_1}, y^\star\in\mathbb{R}^{p_2}$ be such that $Ax^\star+By^\star-c\in\mathcal{K}$. Then, for any $x\in\mathbb{R}^{p_1}, y\in\mathbb{R}^{p_2}$, ℓ defined by* (52) *and ψ defined by* (49) *satisfy*

$$\begin{aligned}\ell(x^\star,y^\star;x_+,\hat{y}) &\leq -\tfrac{1}{2}\|\hat{s}_+\|^2,\\ \ell(x,y;x_+,\hat{y}) &\leq \psi(x,y)-\tfrac{1}{2}\|s-\hat{s}_+\|^2,\end{aligned}\tag{53}$$

where $\hat{s}_+ := Ax_+ + B\hat{y}-c-\mathrm{proj}_{\mathcal{K}}\left(Ax_++B\hat{y}-c\right)$ and $s := Ax+By-c-\mathrm{proj}_{\mathcal{K}}\left(Ax+By-c\right)$.

Proof Let us denote by $\hat{u} := Ax_++B\hat{y}-c\in\mathbb{R}^n$. Since $Ax^\star+By^\star-c\in\mathcal{K}$, if we define $r^\star := Ax^\star+By^\star-c$, then $r^\star\in\mathcal{K}$. Hence, we have

$$\begin{aligned}\ell(x^\star,y^\star;x_+,\hat{y}) &:= \psi(x_+,\hat{y})+\langle\nabla_x\psi(x_+,\hat{y}),x^\star-x_+\rangle+\langle\nabla_y\psi(x_+,\hat{y}),y^\star-\hat{y}\rangle\\ &\overset{(49)}{=} \tfrac{1}{2}\|\hat{u}-\mathrm{proj}_{\mathcal{K}}\left(\hat{u}\right)\|^2+\langle\hat{u}-\mathrm{proj}_{\mathcal{K}}\left(\hat{u}\right),A(x^\star-x_+)+B(y^\star-\hat{y})\rangle\\ &= \|\hat{u}-\mathrm{proj}_{\mathcal{K}}\left(\hat{u}\right)\|^2-\langle\hat{u}-\mathrm{proj}_{\mathcal{K}}\left(\hat{u}\right),\hat{u}-r^\star\rangle-\tfrac{1}{2}\|\hat{u}-\mathrm{proj}_{\mathcal{K}}\left(\hat{u}\right)\|^2\\ &= \langle\hat{u}-\mathrm{proj}_{\mathcal{K}}\left(\hat{u}\right),r^\star-\mathrm{proj}_{\mathcal{K}}\left(\hat{u}\right)\rangle-\tfrac{1}{2}\|\hat{u}-\mathrm{proj}_{\mathcal{K}}\left(\hat{u}\right)\|^2\\ &\leq -\tfrac{1}{2}\|\hat{u}-\mathrm{proj}_{\mathcal{K}}\left(\hat{u}\right)\|^2,\end{aligned}$$

which is the first inequality of (53). Here, we use the property $\langle\hat{u}-\mathrm{proj}_{\mathcal{K}}\left(\hat{u}\right),r^\star-\mathrm{proj}_{\mathcal{K}}\left(\hat{u}\right)\rangle\leq 0$ for any $r^\star\in\mathcal{K}$ of the projection $\mathrm{proj}_{\mathcal{K}}\left(\cdot\right)$. The second inequality of (53) follows directly from (48) and the definition of ψ in (49). □

A.2 The proof of Lemma 5: Descent property of the alternating scheme (10)-(11).

In the sequel, we will use the following lemma for our analysis.

Lemma 5 *Let $z_+ := (x_+,y_+)$ be generated by* (10) *and* (11). *Let $\ell(\cdot)$ be defined by* (52) *and $\Phi_\rho(\cdot)$ be defined by* (7). *Then, for any $x\in\mathbb{R}^{p_1}$ and $y\in\mathbb{R}^{p_2}$, we have*

$$\begin{aligned}\Phi_\rho(z_+) \leq\ & f(x)+g(y)+\rho\ell(x,y;x_+,\hat{y})+\gamma\langle x_+-\hat{x},x-\hat{x}\rangle\\ &+\beta\langle y_+-\tilde{y},y-\tilde{y}\rangle-\gamma\|x_+-\hat{x}\|^2-\beta\|y_+-\tilde{y}\|^2-\tfrac{\rho\|B\|^2}{2}\|y_+-\hat{y}\|^2\\ &-\tfrac{\mu_f}{2}\|x_+-x\|^2-\tfrac{\mu_g}{2}\|y_+-y\|^2.\end{aligned}\tag{54}$$

Proof We first write the optimality condition of (10) and (11) as follows:

$$\begin{cases} 0 = \nabla f(x_+) + \rho\nabla_x\psi(x_+,\hat{y}) + \gamma(x_+ - \hat{x}), & \nabla f(x_+) \in \partial f(x_+),\\ 0 = \nabla g(y_+) + \rho\nabla_y\psi(x_+,\hat{y}) + \beta(y_+ - \tilde{y}), & \nabla g(y_+) \in \partial g(y_+). \end{cases} \tag{55}$$

Next, using the μ_f-strong convexity of f and the μ_g-strong convexity of g, for any $x \in \mathbb{R}^{p_1}$ and $y \in \mathbb{R}^{p_2}$, we have

$$\begin{aligned} f(x_+) &\leq f(x) + \langle \nabla f(x_+), x_+ - x\rangle - \tfrac{\mu_f}{2}\|x_+ - x\|^2, \quad \nabla f(x_+) \in \partial f(x_+),\\ g(y_+) &\leq g(y) + \langle \nabla g(y_+), y_+ - y\rangle - \tfrac{\mu_g}{2}\|y_+ - y\|^2, \quad \nabla g(y_+) \in \partial g(y_+). \end{aligned} \tag{56}$$

Using the second inequality of (51), we have

$$\psi(x_+,y_+) \leq \psi(x_+,\hat{y}) + \langle \nabla_y\psi(x_+,\hat{y}), y_+ - \hat{y}\rangle + \tfrac{\|B\|^2}{2}\|y_+ - \hat{y}\|^2.$$

Multiplying this by ρ, and adding the result to (56), then using (7), we get

$$\begin{aligned} \Phi_\rho(z_+) &\overset{(7)}{=} f(x_+) + g(y_+) + \rho\psi(x_+,y_+)\\ &\leq f(x) + g(y) + \langle \nabla f(x_+) + \rho\nabla_x\psi(x_+,\hat{y}), x_+ - x\rangle - \tfrac{\mu_f}{2}\|x_+ - x\|^2\\ &+ \langle \nabla g(y_+) + \rho\nabla_y\psi(x_+,\hat{y}), y_+ - y\rangle - \tfrac{\mu_g}{2}\|y_+ - y\|^2 + \tfrac{\rho\|B\|^2}{2}\|y_+ - \hat{y}\|^2\\ &+ \rho\psi(x_+,\hat{y}) + \rho\langle \nabla_x\psi(x_+,\hat{y}), x - x_+\rangle + \rho\langle \nabla_y\psi(x_+,\hat{y}), y - \hat{y}\rangle\\ &\overset{(55)+(52)}{=} f(x) + g(y) + \rho\ell(x,y;x_+,\hat{y}) - \tfrac{\mu_f}{2}\|x_+ - x\|^2 - \tfrac{\mu_g}{2}\|y_+ - y\|^2\\ &+ \gamma\langle \hat{x} - x_+, x_+ - x\rangle + \beta\langle \tilde{y} - y_+, y_+ - y\rangle + \tfrac{\rho\|B\|^2}{2}\|y_+ - \hat{y}\|^2\\ &= f(x) + g(y) + \rho\ell(x,y;x_+,\hat{y}) - \tfrac{\mu_f}{2}\|x_+ - x\|^2 - \tfrac{\mu_g}{2}\|y_+ - y\|^2 + \tfrac{\rho\|B\|^2}{2}\|y_+ - \hat{y}\|^2\\ &+ \gamma\langle \hat{x} - x_+, \hat{x} - x\rangle - \gamma\|x_+ - \hat{x}\|^2 + \beta\langle \tilde{y} - y_+, \tilde{y} - y\rangle - \beta\|y_+ - \tilde{y}\|^2, \end{aligned}$$

which is exactly (54). □

A.3 The proof of Lemma 2: The key estimate for Algorithm 1

Using the fact that $\tau_k = \frac{1}{k+1}$, we have $\frac{\tau_{k+1}(1-\tau_k)}{\tau_k} = \frac{k}{k+2}$. Hence, the third line of Step 5 of Algorithm 1 can be written as

$$(\hat{x}^{k+1}, \hat{y}^{k+1}) = (x^{k+1}, y^{k+1}) + \tfrac{\tau_{k+1}(1-\tau_k)}{\tau_k}(x^{k+1} - x^k, y^{k+1} - y^k).$$

This step can be split into two steps with $(\hat{x}^k, \hat{y}^k)$ and $(\tilde{x}^k, \tilde{y}^k)$ as in (15), which is standard in accelerated gradient methods [4,31]. We omit the detail derivation.

Next, we prove (16). We first define the following two quantities and function:

$$\begin{cases} s^k &:= Ax^k + By^k - c - \mathrm{proj}_{\mathcal{K}}\big(Ax^k + By^k - c\big),\\ \hat{s}^{k+1} &:= Ax^{k+1} + B\hat{y}^k - c - \mathrm{proj}_{\mathcal{K}}\big(Ax^{k+1} + B\hat{y}^k - c\big),\\ \ell_k(z) &:= \psi(x^{k+1},\hat{y}^k) + \langle \nabla_x\psi(x^{k+1},\hat{y}^k), x - x^{k+1}\rangle + \langle \nabla_y\psi(x^{k+1},\hat{y}^k), y - \hat{y}^k\rangle. \end{cases} \tag{57}$$

Using (53), and s^k and $\hat{s}^{k+1}$ in (57), we have

$$\ell_k(z^k) \leq \psi(x^k,y^k) - \tfrac{1}{2}\|s^k - \hat{s}^{k+1}\|^2, \quad \text{and} \quad \ell_k(z^\star) \leq -\tfrac{1}{2}\|\hat{s}^{k+1}\|^2. \tag{58}$$

Using (54) with $x_+ := x^{k+1}$, $y_+ := y^{k+1}$, $\hat{x} := \hat{x}^k$, $\hat{y} := \hat{y}^k$, $\tilde{y} := \hat{y}^k$, $\rho := \rho_k$, $\gamma := \gamma_k$ and $\beta := \beta_k$, and then combining the result with the first and the second inequality of (58) at $(x,y) = (x^k, y^k)$ and $(x,y) = (x^\star, y^\star)$ respectively, we obtain

$$
\begin{aligned}
\Phi_{\rho_k}(z^{k+1}) &\le \Phi_{\rho_k}(z^k) + \gamma_k\langle \hat{x}^k - x^{k+1}, \hat{x}^k - x^k\rangle - \gamma_k\|\hat{x}^k - x^{k+1}\|^2 \\
&\quad + \beta_k\langle \hat{y}^k - y^{k+1}, \hat{y}^k - y^k\rangle - \tfrac{(2\beta_k - \|B\|^2\rho_k)}{2}\|\hat{y}^k - y^{k+1}\|^2 \\
&\quad - \tfrac{\rho_k}{2}\|s^k - \hat{s}^{k+1}\|^2 - \tfrac{\mu_f}{2}\|x^{k+1} - x^k\|^2 - \tfrac{\mu_g}{2}\|y^{k+1} - y^k\|^2. \\
\Phi_{\rho_k}(z^{k+1}) &\le f(x^\star) + g(y^\star) - \tfrac{\rho_k}{2}\|\hat{s}^{k+1}\|^2 + \gamma_k\langle \hat{x}^k - x^{k+1}, \hat{x}^k - x^\star\rangle \\
&\quad + \beta_k\langle \hat{y}^k - y^{k+1}, \hat{y}^k - y^\star\rangle - \gamma_k\|\hat{x}^k - x^{k+1}\|^2 \\
&\quad - \tfrac{(2\beta_k - \|B\|^2\rho_k)}{2}\|\hat{y}^k - y^{k+1}\|^2 - \tfrac{\mu_f}{2}\|x^{k+1} - x^\star\|^2 - \tfrac{\mu_g}{2}\|y^{k+1} - y^\star\|^2.
\end{aligned}
\tag{59}
$$

Multiplying the first inequality of (59) by $1 - \tau_k \in [0,1]$ and the second one by $\tau_k \in [0,1]$, and summing up the results, then using $\hat{x}^k - (1-\tau_k)x^k = \tau_k\tilde{x}^k$ and $\hat{y}^k - (1-\tau_k)y^k = \tau_k\tilde{y}^k$ from Step 5 of Algorithm 1, we obtain

$$
\begin{aligned}
\Phi_{\rho_k}(z^{k+1}) &\le (1-\tau_k)\Phi_{\rho_k}(z^k) + \tau_k F(z^\star) + \gamma_k\tau_k\langle \hat{x}^k - x^{k+1}, \tilde{x}^k - x^\star\rangle \\
&\quad - \gamma_k\|x^{k+1} - \hat{x}^k\|^2 + \beta_k\tau_k\langle \hat{y}^k - y^{k+1}, \tilde{y}^k - y^\star\rangle - \tfrac{(2\beta_k - \|B\|^2\rho_k)}{2}\|y^{k+1} - \hat{y}^k\|^2 \\
&\quad - \tfrac{\rho_k\tau_k}{2}\|\hat{s}^{k+1}\|^2 - \tfrac{(1-\tau_k)\rho_k}{2}\|s^k - \hat{s}^{k+1}\|^2 - \tfrac{(1-\tau_k)\mu_f}{2}\|x^{k+1} - x^k\|^2 \\
&\quad - \tfrac{(1-\tau_k)\mu_g}{2}\|y^{k+1} - y^k\|^2 - \tfrac{\tau_k\mu_f}{2}\|x^{k+1} - x^\star\|^2 - \tfrac{\tau_k\mu_g}{2}\|y^{k+1} - y^\star\|^2.
\end{aligned}
\tag{60}
$$

By the update rule at Step 5 of Algorithm 1, we can show that

$$
\begin{aligned}
2\tau_k\langle \hat{x}^k - x^{k+1}, \tilde{x}^k - x^\star\rangle &= \tau_k^2\|\tilde{x}^k - x^\star\|^2 - \tau_k^2\|\tilde{x}^{k+1} - x^\star\|^2 + \|x^{k+1} - \hat{x}^k\|^2, \\
2\tau_k\langle \hat{y}^k - y^{k+1}, \tilde{y}^k - y^\star\rangle &= \tau_k^2\|\tilde{y}^k - y^\star\|^2 - \tau_k^2\|\tilde{y}^{k+1} - y^\star\|^2 + \|y^{k+1} - \hat{y}^k\|^2.
\end{aligned}
$$

Using this relation and $\Phi_{\rho_k}(z^k) = \Phi_{\rho_{k-1}}(z^k) + \frac{(\rho_k - \rho_{k-1})}{2}\|s^k\|^2$ into (60), we get

$$
\begin{aligned}
\Phi_{\rho_k}(z^{k+1}) &\le (1-\tau_k)\Phi_{\rho_{k-1}}(z^k) + \tau_k F(z^\star) + \gamma_k\tau_k^2\left[\|\tilde{x}^k - x^\star\|^2 - \|\tilde{x}^{k+1} - x^\star\|^2\right] \\
&\quad - \tfrac{\gamma_k}{2}\|\hat{x}^k - x^{k+1}\|^2 + \tfrac{\beta_k\tau_k^2}{2}\left[\|\tilde{y}^k - y^\star\|^2 - \|\tilde{y}^{k+1} - y^\star\|^2\right] \\
&\quad - \tfrac{(\beta_k - \|B\|^2\rho_k)}{2}\|y^{k+1} - \hat{y}^k\|^2 - \tfrac{(1-\tau_k)\mu_f}{2}\|x^{k+1} - x^k\|^2 \\
&\quad - \tfrac{(1-\tau_k)\mu_g}{2}\|y^{k+1} - y^k\|^2 - \tfrac{\tau_k\mu_f}{2}\|x^{k+1} - x^\star\|^2 - \tfrac{\tau_k\mu_g}{2}\|y^{k+1} - y^\star\|^2 - R_k,
\end{aligned}
\tag{61}
$$

where $R_k := \frac{(1-\tau_k)\rho_k}{2}\|s^k - \hat{s}^{k+1}\|^2 + \frac{\rho_k\tau_k}{2}\|\hat{s}^{k+1}\|^2 - \frac{(1-\tau_k)(\rho_k - \rho_{k-1})}{2}\|s^k\|^2$. It is easy to show that

$$
R_k \ge \tfrac{(1-\tau_k)}{2}\left[\rho_{k-1} - \rho_k(1-\tau_k)\right]\|s^k\|^2. \tag{62}
$$

By the convexity of $\|\cdot\|^2$ and Step 5 of Algorithm 1, we have

$$
\begin{aligned}
(1-\tau_k)\|x^{k+1} - x^k\|^2 + \tau_k\|x^{k+1} - x^\star\|^2 &\ge \tau_k^2\|\tilde{x}^{k+1} - x^\star\|^2, \\
(1-\tau_k)\|y^{k+1} - y^k\|^2 + \tau_k\|y^{k+1} - y^\star\|^2 &\ge \tau_k^2\|\tilde{y}^{k+1} - y^\star\|^2.
\end{aligned}
$$

Using these expressions and (62) into (61), we obtain (16). □

A.4 The proof of Lemma 3: The key estimate for Algorithm 2

The proof of (19) is similar to the proof of (15), and we skip its detail here.

Next, for simplicity of notation, let us define the following quantities:

$$\begin{cases} \breve{y}^{k+1} & := (1-\tau_k)y^k + \tau_k\tilde{y}^{k+1},\\ Q_k(y) & := \psi(x^{k+1},\hat{y}^k) + \langle \nabla_y\psi(x^{k+1},\hat{y}^k), y-\hat{y}^k\rangle + \frac{\|B\|^2}{2}\|y-\hat{y}^k\|^2,\\ \hat{\Phi}_{k+1} & := f(x^{k+1}) + g(\breve{y}^{k+1}) + \rho_kQ_k(\breve{y}^{k+1}). \end{cases} \tag{63}$$

From the first line of Step 6 of Algorithm 2 and $\breve{y}^{k+1}$ in (63), we have $\breve{y}^{k+1} - \hat{y}^k = \tau_k(\tilde{y}^{k+1} - \tilde{y}^k)$. Moreover, from (19), we have $x^{k+1} - (1-\tau_k)x^k - \tau_k\tilde{x}^{k+1} = 0$. Hence, using these expressions, the convexity of ψ in y, and the definition of $\ell_k(\cdot)$ in (57), we can derive

$$\begin{aligned} Q_k(\breve{y}^{k+1}) &\overset{(63)}{=} \psi(x^{k+1},\hat{y}^k) + \langle\nabla_y\psi(x^{k+1},\hat{y}^k), \breve{y}^{k+1}-\hat{y}^k\rangle + \tfrac{\|B\|^2}{2}\|\breve{y}^{k+1}-\hat{y}^k\|^2\\ &\overset{(57)}{=} (1-\tau_k)\ell_k(z^k) + \tau_k\ell_k(\tilde{z}^{k+1}) + \frac{\|B\|^2\tau_k^2}{2}\|\tilde{y}^{k+1}-\tilde{y}^k\|^2\\ &+ \langle\nabla_x\psi(x^{k+1},\hat{y}^k), x^{k+1}-(1-\tau_k)x^k - \tau_k\tilde{x}^{k+1}\rangle\\ &= (1-\tau_k)\ell_k(z^k) + \tau_k\ell_k(\tilde{z}^{k+1}) + \frac{\|B\|\tau_k^2}{2}\|\tilde{y}^{k+1}-\tilde{y}^k\|^2. \end{aligned} \tag{64}$$

By the convexity of f, and $\tau_k\tilde{x}^{k+1} = x^{k+1} - (1-\tau_k)x^k$ from (19), we can derive

$$f(x^{k+1}) \leq (1-\tau_k)f(x^k) + \tau_kf(x^\star) + \tau_k\langle\nabla f(x^{k+1}), \tilde{x}^{k+1}-x^\star\rangle - \frac{\mu_f\tau_k^2}{2}\|\tilde{x}^{k+1}-x^\star\|^2, \tag{65}$$

where $\nabla f(x^{k+1}) \in \partial f(x^{k+1})$, and $\mu_f \geq 0$ is the strong convexity parameter of f.

By the convexity of g and $\breve{y}^{k+1} = (1-\tau_k)y^k + \tau_k\tilde{y}^{k+1}$ in (63), we also derive

$$g(\breve{y}^{k+1}) \leq (1-\tau_k)g(y^k) + \tau_kg(y^\star) + \tau_k\langle\nabla g(\tilde{y}^{k+1}), \tilde{y}^{k+1}-y^\star\rangle - \frac{\tau_k\mu_g}{2}\|\tilde{y}^{k+1}-y^\star\|^2, \tag{66}$$

for $\nabla g(\tilde{y}^{k+1}) \in \partial g(\tilde{y}^{k+1})$, where $\mu_g > 0$ is the strong convexity parameter of g.

From the definition of $\ell_k(\cdot)$ in (57), and the relation (53), we can write

$$\begin{aligned} \ell_k(\tilde{z}^{k+1}) &= \ell_k(z^\star) + \langle\nabla_x\psi(x^{k+1},\hat{y}^k), \tilde{x}^{k+1}-x^\star\rangle + \langle\nabla_y\psi(x^{k+1},\hat{y}^k), \tilde{y}^{k+1}-y^\star\rangle\\ &\overset{(53)}{\leq} \langle\nabla_x\psi(x^{k+1},\hat{y}^k), \tilde{x}^{k+1}-x^\star\rangle + \langle\nabla_y\psi(x^{k+1},\hat{y}^k), \tilde{y}^{k+1}-y^\star\rangle - \tfrac{1}{2}\|\hat{s}^{k+1}\|^2. \end{aligned}$$

Combining this, (64), (65) and (66) and then using $\hat{\Phi}_k$ defined in (63), we have

$$\begin{aligned} \hat{\Phi}_{k+1} &\overset{(63)}{=} f(x^{k+1}) + g(\breve{y}^{k+1}) + \rho_kQ_k(\breve{y}^{k+1})\\ &\overset{(64),(65),(66)}{\leq} (1-\tau_k)\left[f(x^k) + g(y^k) + \rho_k\ell_k(z^k)\right] + \tau_k\left[f(x^\star) + g(y^\star)\right]\\ &+ \tau_k\langle\nabla f(x^{k+1}), \tilde{x}^{k+1}-x^\star\rangle + \tau_k\langle\nabla g(\tilde{y}^{k+1}), \tilde{y}^{k+1}-y^\star\rangle + \rho_k\tau_k\ell_k(\tilde{z}^{k+1})\\ &+ \tfrac{\|B\|^2\rho_k\tau_k^2}{2}\|\tilde{y}^{k+1}-\tilde{y}^k\|^2 - \tfrac{\mu_f\tau_k^2}{2}\|\tilde{x}^{k+1}-x^\star\|^2 - \tfrac{\tau_k\mu_g}{2}\|\tilde{y}^{k+1}-y^\star\|^2\\ &\leq (1-\tau_k)\left[f(x^k) + g(y^k) + \rho_k\ell_k(z^k)\right] + \tau_k\left[f(x^\star) + g(y^\star)\right]\\ &- \tfrac{\rho_k\tau_k}{2}\|\hat{s}^{k+1}\|^2 + \tau_k\langle\nabla f(x^{k+1}) + \rho_k\nabla_x\psi(x^{k+1},\hat{y}^k), \tilde{x}^{k+1}-x^\star\rangle\\ &+ \tau_k\langle\nabla g(\tilde{y}^{k+1}) + \rho_k\nabla_y\psi(x^{k+1},\hat{y}^k), \tilde{y}^{k+1}-y^\star\rangle + \tfrac{\|B\|^2\rho_k\tau_k^2}{2}\|\tilde{y}^{k+1}-\tilde{y}^k\|^2\\ &- \tfrac{\mu_f\tau_k^2}{2}\|\tilde{x}^{k+1}-x^\star\|^2 - \tfrac{\tau_k\mu_g}{2}\|\tilde{y}^{k+1}-y^\star\|^2. \end{aligned} \tag{67}$$

Next, by the optimality condition of (10) and (11), we have

$$\begin{cases} 0 = \nabla f(x^{k+1}) + \rho_k \nabla_x \psi(x^{k+1}, \hat{y}^k) + \gamma_k(x^{k+1} - \hat{x}^k), \quad \nabla f(x^{k+1}) \in \partial f(x^{k+1}), \\ 0 = \nabla g(\tilde{y}^{k+1}) + \rho_k \nabla_y \psi(x^{k+1}, \hat{y}^k) + \beta_k\tau_k(\tilde{y}^{k+1} - \tilde{y}^k), \, \nabla g(\tilde{y}^{k+1}) \in \partial g(\tilde{y}^{k+1}). \end{cases} \tag{68}$$

In addition, using (19), we also have

$$\begin{aligned} 2\tau_k\langle \hat{x}^k - x^{k+1}, \tilde{x}^k - x^\star\rangle &= \tau_k^2\|\tilde{x}^k - x^\star\|^2 - \tau_k^2\|\tilde{x}^{k+1} - x^\star\|^2 + \|x^{k+1} - \hat{x}^k\|^2, \\ 2\langle \tilde{y}^k - \tilde{y}^{k+1}, \tilde{y}^{k+1} - y^\star\rangle &= \|\tilde{y}^k - y^\star\|^2 - \|\tilde{y}^{k+1} - y^\star\|^2 + \|\tilde{y}^{k+1} - \tilde{y}^k\|^2. \end{aligned} \tag{69}$$

Using (68) and (69) into (67) we obtain

$$\begin{aligned} \hat{\Phi}_{k+1} &\leq (1-\tau_k)\left[f(x^k) + g(y^k) + \rho_k\ell_k(z^k)\right] + \tau_k\left[f(x^\star) + g(y^\star)\right] \\ &\quad + \gamma_k\tau_k\langle \hat{x}^k - x^{k+1}, \tilde{x}^{k+1} - x^\star\rangle + \beta_k\tau_k^2\langle \tilde{y}^k - \tilde{y}^{k+1}, \tilde{y}^{k+1} - y^\star\rangle - \tfrac{\rho_k\tau_k}{2}\|\hat{s}^{k+1}\|^2 \\ &\quad + \tfrac{\|B\|^2\rho_k\tau_k^2}{2}\|\tilde{y}^{k+1} - \tilde{y}^k\|^2 - \tfrac{\mu_f\tau_k^2}{2}\|\tilde{x}^{k+1} - x^\star\|^2 - \tfrac{\tau_k\mu_g}{2}\|\tilde{y}^{k+1} - y^\star\|^2 \\ &= (1-\tau_k)\left[f(x^k) + g(y^k) + \rho_k\ell_k(z^k)\right] + \tau_k F(z^\star) - \tfrac{\rho_k\tau_k}{2}\|\hat{s}^{k+1}\|^2 \\ &\quad + \tfrac{\gamma_k\tau_k^2}{2}\|\tilde{x}^k - x^\star\|^2 - \tfrac{(\gamma_k+\mu_f)\tau_k^2}{2}\|\tilde{x}^{k+1} - x^\star\|^2 - \tfrac{\gamma_k}{2}\|x^{k+1} - \hat{x}^k\|^2 \\ &\quad + \tfrac{\|B\|^2\rho_k\tau_k^2}{2}\|\tilde{y}^k - y^\star\|^2 - \tfrac{(\|B\|^2\rho_k\tau_k^2 + \mu_g\tau_k)}{2}\|\tilde{y}^{k+1} - y^\star\|^2 \\ &\quad - \tfrac{(\beta_k - \rho_k\|B\|^2)\tau_k^2}{2}\|\tilde{y}^{k+1} - \tilde{y}^k\|^2. \end{aligned} \tag{70}$$

From the definition of ψ in (49) and (53), we have

$$\begin{aligned} \ell_k(z^k) &\leq \psi(x^k, y^k) - \tfrac{1}{2}\|s^k - \hat{s}^{k+1}\|^2, \\ \Phi_{\rho_k}(z^k) &= \Phi_{\rho_{k-1}}(z^k) + \tfrac{(\rho_k - \rho_{k-1})}{2}\psi(x^k, y^k) = \Phi_{\rho_{k-1}}(z^k) + \tfrac{(\rho_k - \rho_{k-1})}{2}\|s^k\|^2. \end{aligned}$$

Using these expressions into (70), we obtain

$$\begin{aligned} \hat{\Phi}_{k+1} &\leq (1-\tau_k)\Phi_{\rho_{k-1}}(z^k) + \tau_k F(z^\star) + \tfrac{\gamma_k\tau_k^2}{2}\|\tilde{x}^k - x^\star\|^2 - \tfrac{(\gamma_k+\mu_f)\tau_k^2}{2}\|\tilde{x}^{k+1} - x^\star\|^2 \\ &\quad - \tfrac{\gamma_k}{2}\|x^{k+1} - \hat{x}^k\|^2 + \tfrac{\|B\|^2\rho_k\tau_k^2}{2}\|\tilde{y}^k - y^\star\|^2 - \tfrac{(\|B\|^2\rho_k\tau_k^2 + \mu_g\tau_k)}{2}\|\tilde{y}^{k+1} - y^\star\|^2 \\ &\quad - \tfrac{(\beta_k - \|B\|^2\rho_k)\tau_k^2}{2}\|\tilde{y}^{k+1} - \tilde{y}^k\|^2 - R_k, \end{aligned} \tag{71}$$

where R_k is defined as $R_k := \frac{\rho_k\tau_k}{2}\|\hat{s}^{k+1}\|^2 + \frac{(1-\tau_k)\rho_k}{2}\|s^k - \hat{s}^{k+1}\|^2 - \frac{(\rho_k - \rho_{k-1})(1-\tau_k)}{2}\|s^k\|^2$.

Let consider two cases:

- For **Choice 1** at Step 7 of Algorithm 2, we have $y^{k+1} = \breve{y}^{k+1}$. Hence, using (51), we get

$$\Phi_{\rho_k}(z^{k+1}) = f(x^{k+1}) + g(y^{k+1}) + \rho_k\psi(x^{k+1}, y^{k+1}) \leq \hat{\Phi}_{k+1}. \tag{72}$$

- For **Choice 2** at Step 7 of Algorithm 2, we have

$$\begin{aligned} \Phi_{\rho_k}(z^{k+1}) &\leq g(y^{k+1}) + \rho_k Q_k(y^{k+1}) = \min_y \{g(y) + \rho_k Q_k(y)\} \\ &\leq g(\breve{y}^{k+1}) + \rho_k Q_k(\breve{y}^{k+1}) = \hat{\Phi}_{k+1}. \end{aligned} \tag{73}$$

Using either (72) or (73) into (71), we obtain

$$\begin{aligned}\Phi_{\rho_k}(z^{k+1}) \le (1-\tau_k)\Phi_{\rho_{k-1}}(z^k) + \tau_k F(z^\star) + \tfrac{\gamma_k\tau_k^2}{2}\|\tilde{x}^k - x^\star\|^2 - \tfrac{(\gamma_k+\mu_f)\tau_k^2}{2}\|\tilde{x}^{k+1} - x^\star\|^2 \\ - \tfrac{\gamma_k}{2}\|x^{k+1} - \hat{x}^k\|^2 + \tfrac{\|B\|^2\rho_k\tau_k^2}{2}\|\tilde{y}^k - y^\star\|^2 - \tfrac{(\|B\|^2\rho_k\tau_k^2 + \mu_g\tau_k)}{2}\|\tilde{y}^{k+1} - y^\star\|^2 \\ - \tfrac{(\beta_k - \|B\|^2\rho_k)\tau_k^2}{2}\|\tilde{y}^{k+1} - \tilde{y}^k\|^2 - R_k,\end{aligned}$$

Using the lower bound (62) of R_k into this inequality, we obtain (20). □

A.5 The proof of Remark 1: The $\mathcal{O}\left(\frac{1}{k}\right)$-convergence rate

We show that even with the tightest condition for τ_k and ρ_k in Algorithm 1, the $\mathcal{O}\left(\frac{1}{k}\right)$ convergence rate cannot be improved as stated in Remark 1. Indeed, the tightest condition for τ_k and ρ_k derived from (16) of Lemma 2 is

$$\|B\|^2\rho_k\tau_k^2 = (\|B\|^2\rho_{k-1} + \mu_g)\tau_{k-1}^2(1-\tau_k) \quad \text{and} \quad \rho_k(1-\tau_k) = \rho_{k-1}.$$

Since we need to keep the rate of τ_k at $\mathcal{O}\left(\frac{1}{k}\right)$, without loss of generality, we can assume that $\tau_k = \frac{c}{k+1}$ for some $c > 0$. Combining both conditions above and using $\tau_k = \frac{c}{k+1}$, we derive $\frac{\tau_k^2}{\tau_{k-1}^2(1-\tau_k)^2} - 1 = \frac{\mu_g}{\|B\|^2\rho_{k-1}}$, or $\frac{k^2}{(k+1-c)^2} - 1 = \frac{\mu_g}{\|B\|^2\rho_{k-1}}$. This leads to $\rho_{k-1} = \frac{\mu_g(k+1-c)^2}{(c-1)(2k-c+1)\|B\|^2}$. If $c = 1$, then $\tau_k = \frac{1}{k+1}$, which we obtain the rate as in Theorem 1. If $c > 1$, then $\rho_{k-1} = \frac{\mu_g(k+1-c)^2}{(c-1)(2k-c+1)\|B\|^2}$ shows that $\rho_k = \mathcal{O}\left(\frac{1}{k}\right)$, which does not improve the $\mathcal{O}\left(\frac{1}{k}\right)$ rate in Theorem 1. If $c < 1$, then $\omega_k := \prod_{i=1}^k(1-\tau_i) = \mathcal{O}\left(\frac{1}{k^c}\right)$ [43], which is worse than $\mathcal{O}\left(\frac{1}{k}\right)$. Hence, we can conclude that the rate in Theorem 1 is not better than $\mathcal{O}\left(\frac{1}{k}\right)$. □

A.6 The proof of Corollary 1: Application to composite convex minimization

By the L_f-Lipschitz continuity of f and Lemma 1, we have

$$\begin{aligned}0 \le P(y^k) - P^\star = f(y^k) + g(y^k) - P^\star &\le f(x^k) + g(y^k) + |f(y^k) - f(x^k)| - P^\star \\ &\le f(x^k) + g(y^k) - P^\star + L_f\|x^k - y^k\| \\ &\overset{(8)}{\le} S_{\rho_k}(z^k) - \tfrac{\rho_k}{2}\|x^k - y^k\|^2 + L_f\|x^k - y^k\|,\end{aligned} \tag{74}$$

where $S_\rho(z) := \Phi_\rho(z) - P^\star$. This inequality also leads to

$$\|x^k - y^k\| \le \frac{1}{\rho_k}\left(L_f + \sqrt{L_f^2 + 2\rho_k S_{\rho_k}(z^k)}\right) \le \frac{1}{\rho_k}\left(2L_f + \sqrt{2\rho_k S_{\rho_k}(z^k)}\right). \tag{75}$$

Since using (24) is equivalent to applying Algorithm 1 to its constrained reformulation. In this case, by (17), we have

$$S_{\rho_k}(z^k) \le \frac{\rho_0\|y^0 - y^\star\|^2}{2k} \quad \text{and} \quad \rho_k = (k+1)\rho_0.$$

Using this into (75) we get

$$\|x^k - y^k\| \le \frac{1}{\rho_0(k+1)}\left(2L_f + \sqrt{2\rho_0^2\|y^0 - y^\star\|^2}\right) = \frac{2L_f + \sqrt{2}\rho_0\|y^0 - y^\star\|}{\rho_0(k+1)}.$$

Substituting this into (74) and using the bound of S_{ρ_k}, we obtain (27).

Now, if we use (26), then it is equivalent to applying Algorithm 2 with **Choice 1** to solve its constrained reformulation. In this case, from the proof of Theorem 2, we can derive

$$S_{\rho_k}(z^k) \leq \frac{2\rho_0\|y^0 - y^\star\|^2}{(k+1)^2} \quad \text{and} \quad \frac{(k+2)^2}{4}\rho_0 \leq \rho_k \leq (k+1)^2\rho_0.$$

Combining these estimates and (75), we have $\|x^k - y^k\| \leq \frac{4(2L_f + \sqrt{2}\rho_0\|y^0 - y^\star\|)}{\rho_0(k+2)^2}$. Substituting this into (74) and using the bound of S_{ρ_k} we obtain (28). □

A.7 The proof of Corollary 2: Extension to the sum of three objectives

By the Lipschitz gradient of h, using [31, Theorem 2.1.5], we can show that

$$\begin{aligned} h(y^{k+1}) &\leq h(\hat{y}^k) + \langle \nabla h(\hat{y}^k), y^{k+1} - \hat{y}^k \rangle + \tfrac{L_h}{2}\|y^{k+1} - \hat{y}^k\|^2 \\ &\leq h(\hat{y}^k) + \langle \nabla h(\hat{y}^k), y - \hat{y}^k \rangle + \langle \nabla h(\hat{y}^k), y^{k+1} - y \rangle + \tfrac{L_h}{2}\|y^{k+1} - \hat{y}^k\|^2. \end{aligned}$$

In addition, the optimality condition of (36) is

$$0 = \nabla g(y^{k+1}) + \nabla h(\hat{y}^k) + \rho_k \nabla_y \psi(x^{k+1}, \hat{y}^k) + \hat{\beta}_k(y^{k+1} - \hat{y}^k),\ \nabla g(y^{k+1}) \in \partial g(y^{k+1}).$$

Using these expressions and the same argument as the proof of Lemma 5, we derive

$$\begin{aligned} \Phi_{\rho_k}(z^{k+1}) &\leq f(x) + g(y) + h(\hat{y}^k) + \langle \nabla h(\hat{y}^k, y - \hat{y}^k \rangle + \rho_k \ell_k(x, y) \\ &\quad + \gamma_k \langle \hat{x}^k - x^{k+1}, x^{k+1} - x \rangle + \hat{\beta}_k \langle \hat{y}^k - y^{k+1}, y^{k+1} - y \rangle \\ &\quad + \tfrac{\rho_k\|B\|^2 + L_h}{2}\|y^{k+1} - \hat{y}^k\|^2 - \tfrac{\mu_f}{2}\|x^{k+1} - x\|^2 - \tfrac{\mu_g}{2}\|y^{k+1} - y\|^2. \end{aligned} \tag{76}$$

Eventually, with the same proof as in (16), we can show that

$$\begin{aligned} \Phi_{\rho_k}(z^{k+1}) &\leq (1 - \tau_k)\Phi_{\rho_{k-1}}(z^k) + \tau_k F(z^\star) + \tfrac{\gamma_k \tau_k^2}{2}\|\tilde{x}^k - x^\star\|^2 \\ &\quad - \tfrac{(\gamma_k + \mu_f)\tau_k^2}{2}\|\tilde{x}^{k+1} - x^\star\|^2 + \tfrac{\hat{\beta}_k \tau_k^2}{2}\|\tilde{y}^k - y^\star\|^2 - \tfrac{(\hat{\beta}_k + \mu_g)\tau_k^2}{2}\|\tilde{y}^{k+1} - y^\star\|^2 \\ &\quad - \tfrac{(\hat{\beta}_k - \|B\|^2\rho_k - L_h)}{2}\|y^{k+1} - \hat{y}^k\|^2 - \tfrac{(1-\tau_k)}{2}\left[\rho_{k-1} - \rho_k(1 - \tau_k)\right]\|s^k\|^2, \end{aligned} \tag{77}$$

where $s^k := Ax^k + By^k - c - \mathrm{proj}_{\mathcal{K}}\big(Ax^k + By^k - c\big)$. Hence, we can choose $\hat{\beta}_k = \|B\|^2\rho_k + L_h$. Assume that $\mu_f = \mu_g = 0$, then we have a condition on ρ_k and τ_k as

$$\frac{(\|B\|^2\rho_k + L_h)\tau_k^2}{1 - \tau_k} \leq (\|B\|^2\rho_{k-1} + L_h)\tau_{k-1}^2 \ \text{ and } \ \rho_k = \frac{\rho_{k-1}}{1 - \tau_k}.$$

If we choose $\tau_k = \frac{1}{k+1}$, then $\rho_k = \rho_0(k+1)$. The first condition above becomes

$$\begin{aligned} &\tfrac{(\|B\|^2\rho_0(k+1) + L_h)}{k(k+1)} \leq \tfrac{(\|B\|^2\rho_0 k + L_h)}{k^2} \\ \Leftrightarrow\ &\|B\|^2\rho_0 k(k+1) + L_h k \leq \|B\|^2\rho_0 k(k+1) + L_h(k+1). \end{aligned}$$

which certainly holds.

The remaining proof of the first part in Corollary 2 is similar to the proof of Theorem 1, but with $R_p^2 := \frac{\gamma_0}{\rho_0}\|x^0 - x^\star\|^2 + (L_h + \rho_0\|B\|^2)\|y^0 - y^\star\|^2$ due to (77).

We now prove the second part of Corollary 2. For the case (i) with $\mu_g > 0$, the proof is very similar to the proof of Theorem 2, but β_k is changed to $\hat{\beta}_k$ and is updated as $\hat{\beta}_k = \rho_k\left\Vert B\right\Vert^2 + L_h$. We omit the detail of this analysis here. We only prove the second case (ii) when $L_h < 2\mu_h$.

Using the convexity and the Lipschitz gradient of h, we can derive

$$
\begin{aligned}
h(\breve{y}^{k+1}) &\leq (1-\tau_k)h(y^k) + \tau_k h(\tilde{y}^{k+1}) - \tfrac{\mu_h\tau_k(1-\tau_k)}{2}\Vert\tilde{y}^{k+1} - y^k\Vert^2\\
&\leq (1-\tau_k)h(y^k) + \tau_k h(\tilde{y}^k) + \tau_k\langle\nabla h(\tilde{y}^k), \tilde{y}^{k+1} - \tilde{y}^k\rangle + \tfrac{\tau_k L_h}{2}\Vert\tilde{y}^{k+1} - \tilde{y}^k\Vert^2\\
&\leq (1-\tau_k)h(y^k) + \tau_k h(y^\star) + \tau_k\langle\nabla h(\tilde{y}^k), \tilde{y}^{k+1} - y^\star\rangle\\
&+ \tfrac{\tau_k L_h}{2}\Vert\tilde{y}^{k+1} - \tilde{y}^k\Vert^2 - \tfrac{\tau_k\mu_h}{2}\Vert\tilde{y}^k - y^\star\Vert^2.
\end{aligned}
$$

Using this estimate, with a similar proof as of (67), we can derive

$$
\begin{aligned}
\breve{\Phi}_{k+1} &:= f(x^{k+1}) + g(\breve{y}^{k+1}) + h(\breve{y}^{k+1}) + \rho_k Q_k(\breve{y}^{k+1})\\
&\overset{(64),(65),(66)}{\leq} (1-\tau_k)\left[f(x^k) + g(y^k) + h(y^k) + \rho_k\ell_k(z^k)\right] + \tau_k\left[f(x^\star) + g(y^\star) + h(y^\star)\right]\\
&+ \tau_k\langle\nabla f(x^{k+1}), \tilde{x}^{k+1} - x^\star\rangle + \tau_k\langle\nabla g(\tilde{y}^{k+1}) + \nabla h(\tilde{y}^k), \tilde{y}^{k+1} - y^\star\rangle - \tfrac{\rho_k\tau_k}{2}\Vert\hat{s}^{k+1}\Vert^2\\
&+ \tfrac{(\Vert B\Vert^2\rho_k\tau_k^2 + L_h\tau_k)}{2}\Vert\tilde{y}^{k+1} - \tilde{y}^k\Vert^2 - \tfrac{\mu_f\tau_k^2}{2}\Vert\tilde{x}^{k+1} - x^\star\Vert^2 - \tfrac{\tau_k\mu_g}{2}\Vert\tilde{y}^{k+1} - y^\star\Vert^2 - \tfrac{\tau_k\mu_h}{2}\Vert\tilde{y}^k - y^\star\Vert^2\\
&\leq (1-\tau_k)\left[f(x^k) + g(y^k) + h(y^k) + \rho_k\ell_k(z^k)\right] + \tau_k\left[f(x^\star) + g(y^\star) + h(y^\star)\right]\\
&- \tfrac{\rho_k\tau_k}{2}\Vert\hat{s}^{k+1}\Vert^2 + \tau_k\gamma_k\langle\hat{x}^k - x^{k+1}, \tilde{x}^{k+1} - x^\star\rangle + \tau_k^2\hat{\beta}_k\langle\tilde{y}^k - \tilde{y}^{k+1}, \hat{y}^k), \tilde{y}^{k+1} - y^\star\rangle\\
&+ \tfrac{(\Vert B\Vert^2\rho_k\tau_k^2 + \tau_k L_h)}{2}\Vert\tilde{y}^{k+1} - \tilde{y}^k\Vert^2 - \tfrac{\mu_f\tau_k^2}{2}\Vert\tilde{x}^{k+1} - x^\star\Vert^2 - \tfrac{\tau_k\mu_g}{2}\Vert\tilde{y}^{k+1} - y^\star\Vert^2 - \tfrac{\tau_k\mu_h}{2}\Vert\tilde{y}^k - y^\star\Vert^2.
\end{aligned}
$$

Here, we use the optimality condition of (10) and (37) into the last inequality.

Using the same argument as the proof of (20), the last inequality above together with (38) leads to

$$
\begin{aligned}
D_{k+1}^\star &+ \tfrac{(\gamma_k + \mu_f)\tau_k^2}{2}\Vert\tilde{x}^{k+1} - x^\star\Vert^2 + \tfrac{\hat{\beta}_k\tau_k^2 + \mu_g\tau_k}{2}\Vert\tilde{y}^{k+1} - y^\star\Vert^2 \leq (1-\tau_k)D_k^\star\\
&+ \tfrac{\gamma_k\tau_k^2}{2}\Vert\tilde{x}^k - x^\star\Vert^2 + \tfrac{\hat{\beta}_k\tau_k^2 - \tau_k\mu_h}{2}\Vert\tilde{y}^k - y^\star\Vert^2\\
&- \tfrac{(\hat{\beta}_k\tau_k^2 - \rho_k\Vert B\Vert^2\tau_k^2 - \tau_k L_h)}{2}\Vert\tilde{y}^{k+1} - \tilde{y}^k\Vert^2 - \tfrac{(1-\tau_k)}{2}\left[\rho_{k-1} - \rho_k(1-\tau_k)\right]\Vert s^k\Vert^2,
\end{aligned}
\tag{78}
$$

where $D_k^\star := \Phi_{\rho_{k-1}}(z^k) - F^\star$ and $s^k := Ax^k + By^k - c - \mathrm{proj}_{\mathcal{K}}\big(Ax^k + By^k - c\big)$. We still choose the update rule for τ_k, ρ_k and γ_k as in Algorithm 2. Then, in order to telescope this inequality, we impose the following conditions:

$$
\hat{\beta}_k = \rho_k\Vert B\Vert^2 + \tfrac{L_h}{\tau_k}, \quad \text{and} \quad \hat{\beta}_k\tau_k^2 - \mu_h\tau_k \leq (1-\tau_k)(\hat{\beta}_{k-1}\tau_{k-1}^2 + \mu_g\tau_{k-1}).
$$

Using the first condition into the second one and note that $1 - \tau_k = \frac{\tau_k^2}{\tau_{k-1}^2}$ and $\rho_k = \frac{\rho_0}{\tau_k^2}$, we obtain $\rho_0\left\Vert B\right\Vert^2 + L_h - \mu_h \leq \frac{\tau_k}{\tau_{k-1}}(L_h + \mu_g)$. This condition holds if $\rho_0 \leq \frac{\mu_g + 2\mu_h - L_h}{2\Vert B\Vert^2} > 0$. Using (78) we have the same conclusion as in Theorem 2. □

References

1. H. Attouch, Z. Chbani, and H. Riahi. Rate of convergence of the nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$. *arXiv preprint arXiv:1706.05671*, 2017.
2. A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.*, 16(3):697–725, 2006.
3. H.H. Bauschke and P. Combettes. *Convex analysis and monotone operators theory in Hilbert spaces.* Springer-Verlag, 2011.

4. A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding agorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.
5. S. Becker, E. J. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Compt.*, 3(3):165–218, 2011.
6. A. Belloni, V. Chernozhukov, and L. Wang. Square-root LASSO: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 94(4):791–806, 2011.
7. A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization: Analysis, algorithms, and engineering applications*, volume 3 of *MPS/SIAM Series on Optimization*. SIAM, 2001.
8. D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.
9. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
10. S. Boyd and L. Vandenberghe. *Convex Optimization*. University Press, Cambridge, 2004.
11. A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.
12. G. Chen and M. Teboulle. A proximal-based decomposition method for convex minimization problems. *Math. Program.*, 64:81–101, 1994.
13. L. Condat. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J. Optim. Theory Appl.*, 158:460–479, 2013.
14. D. Davis. Convergence rate analysis of primal-dual splitting schemes. *SIAM J. Optim.*, 25(3):1912–1943, 2015.
15. D. Davis. Convergence rate analysis of the forward-Douglas-Rachford splitting scheme. *SIAM J. Optim.*, 25(3):1760–1786, 2015.
16. D. Davis and W. Yin. Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions. *Mathematics of Operations Research*, 2014.
17. J. Eckstein and D. Bertsekas. On the Douglas - Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Program.*, 55:293–318, 1992.
18. E. Esser, X. Zhang, and T. Chan. A general framework for a class of first order primal-dual algorithms for TV-minimization. *SIAM J. Imaging Sciences*, 3(4):1015–1046, 2010.
19. R. Fletcher. *Practical Methods of Optimization*. Wiley, Chichester, 2nd edition, 1987.
20. D. Goldfarb, S. Ma, and K. Scheinberg. Fast alternating linearization methods of minimization of the sum of two convex functions. *Math. Program., Ser. A*, 141(1):349–382, 2012.
21. T. Goldstein, B. ODonoghue, and S. Setzer. Fast Alternating Direction Optimization Methods. *SIAM J. Imaging Sci.*, 7(3):1588–1623, 2012.
22. M. Grant, S. Boyd, and Y. Ye. Disciplined convex programming. In L. Liberti and N. Maculan, editors, *Global Optimization: From Theory to Implementation*, Nonconvex Optimization and its Applications, pages 155–210. Springer, 2006.
23. B.S. He and X.M. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM J. Numer. Anal.*, 50:700–709, 2012.
24. M. Jaggi. Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. *JMLR W&CP*, 28(1):427–435, 2013.
25. R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 315–323, 2013.
26. G. Lan and R.D.C. Monteiro. Iteration complexity of first-order penalty methods for convex programming. *Math. Program.*, 138(1):115–139, 2013.
27. I. Necoara and J.A.K. Suykens. Application of a Smoothing Technique to Decomposition in Convex Optimization. *IEEE Tran. Automatic Control*, 53(11):2674–2679, 2008.
28. I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *arXiv preprint arXiv:1504.06298*, 2015.
29. I. Necoara, A. Patrascu, and F. Glineur. Complexity certifications of first order inexact Lagrangian and penalty methods for conic convex programming. *arXiv preprint arXiv:1506.05320*, 2015.
30. 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.
31. Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, 2004.

32. Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.*, 22(2):341–362, 2012.
33. Y. Nesterov. Gradient methods for minimizing composite objective function. *Math. Program.*, 140(1):125–161, 2013.
34. J. Nocedal and S.J. Wright. *Numerical Optimization.* Springer Series in Operations Research and Financial Engineering. Springer, 2 edition, 2006.
35. B. O'Donoghue and E. Candes. Adaptive Restart for Accelerated Gradient Schemes. *Found. Comput. Math.*, 15:715–732, 2015.
36. Y. Ouyang, Y. Chen, G. Lan, and E. JR. Pasiliao. An accelerated linearized alternating direction method of multiplier. *SIAM J. Imaging Sci.*, 8(1):644–681, 2015.
37. R. T. Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematics Series.* Princeton University Press, 1970.
38. R. Shefi and M. Teboulle. On the rate of convergence of the proximal alternating linearized minimization algorithm for convex problems. *EURO J. Comput. Optim.*, 4(1):27–46, 2016.
39. S. Sra, S. Nowozin, and S. J. Wright. *Optimization for machine learning.* MIT Press, 2012.
40. Q. Tran-Dinh, A. Alacaoglu, O. Fercoq, and V. Cevher. Self-adaptive restarting smoothed gap reduction algorithms for nonsmooth convex optimization. *Tech. Report (STOR-UNC-Chapel Hill)*, 2017.
41. Q. Tran-Dinh and V. Cevher. A primal-dual algorithmic framework for constrained convex minimization. *Tech. Report., LIONS*, pages 1–54, 2014.
42. Q. Tran-Dinh, O. Fercoq, and V. Cevher. A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *To appear on SIAM J. Optim.*, pages 1–29, 2017.
43. Q. Tran-Dinh, I. Necoara, and M. Diehl. Fast inexact decomposition algorithms for large-scale separable convex optimization. *Optimization*, 66(2): 325–356, 2016.
44. Q. Tran-Dinh. Construction and iteration-complexity of primal sequences in alternating minimization algorithms. *Tech. Report, UNC-STOR-2015, arXiv preprint arXiv:1511.03305*, 2015.
45. Q. Tran-Dinh. Adaptive Smoothing Algorithms for Nonsmooth Composite Convex Minimization. *Comput. Optim. Appl.*, 66(3):425–451, 2016.
46. P. Tseng. Applications of splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control Optim.*, 29:119–138, 1991.
47. P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Submitted to SIAM J. Optim.*, 2008.
48. Q. Van Nguyen, O. Fercoq, and V. Cevher. Smoothing technique for nonsmooth composite minimization with linear operator. *arXiv preprint arXiv:1706.05837*, 2017.
49. B. Cong Vu. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.
50. Y. Xu. Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming. *SIAM J. Optim.*, 27(3):1459–1484, 2017.
51. H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.