ORIGINAL PAPER

A unified convergence rate analysis of the accelerated smoothed gap reduction algorithm

Quoc Tran-Dinh[1]

Abstract

In this paper, we develop a unified convergence analysis framework for the *Accelerated Smoothed GAp ReDuction algorithm* (ASGARD) introduced in Tran-Dinh et al. (SIAM J Optim 28(1):96–134, 2018). Unlike Tran-Dinh et al. (SIAM J Optim 28(1):96–134, 2018), the new analysis covers three settings in a single algorithm: general convexity, strong convexity, and strong convexity and smoothness. Moreover, we establish the convergence guarantees on three criteria: (i) gap function, (ii) primal objective residual, and (iii) dual objective residual. Our convergence rates are optimal (up to a constant factor) in all cases. While the convergence rate on the primal objective residual for the general convex case has been established in Tran-Dinh et al. (SIAM J Optim 28(1):96–134, 2018), we prove additional convergence rates on the gap function and the dual objective residual. The analysis for the last two cases is completely new. Our results provide a complete picture on the convergence guarantees of ASGARD. Finally, we present four different numerical experiments on a representative optimization model to verify our algorithm and compare it with the well-known Nesterov's smoothing algorithm.

Keywords Accelerated smoothed gap reduction · Primal–dual algorithm · Nesterov's smoothing technique · Convex–concave saddle-point problem

1 Introduction

We consider the following classical convex–concave saddle-point problem:

$$\min_{x \in \mathbb{R}^p} \max_{y \in \mathbb{R}^n} \left\{ \mathcal{L}(x, y) := f(x) + \langle Kx, y \rangle - g^*(y) \right\}, \tag{1}$$

✉ Quoc Tran-Dinh
quoctd@email.unc.edu

1   Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill,
    333 Hanes Hall, UNC-CH, Chapel Hill, NC 27599, USA

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, $K : \mathbb{R}^p \to \mathbb{R}^n$ is a linear operator, and $g^*(y) := \sup_x \{\langle y, x \rangle - g(x)\}$ is the Fenchel conjugate of g.

The convex–concave saddle-point problem (1) can be written in the primal and dual forms. The primal problem is defined as

$$F^\star := \min_{x \in \mathbb{R}^p} \left\{ F(x) := f(x) + g(Kx) \right\}. \tag{2}$$

The corresponding dual problem is written as

$$D^\star := \min_{y \in \mathbb{R}^n} \left\{ D(y) := f^*(-K^\top y) + g^*(y) \right\}. \tag{3}$$

Clearly, both the primal problem (2) and its dual form (3) are convex.

Motivation   In [23], two accelerated smoothed gap reduction algorithms are proposed to solve (2) and its special constrained convex problem. Both algorithms achieve optimal sublinear convergence rates (up to a constant factor) in the sense of black-box first-order oracles [15,16], when f and g are convex, and when f is strongly convex and g is just convex, respectively. The first algorithm in [23] is called ASGARD (*Accelerated Smoothed GAp ReDuction*). To the best of our knowledge, except for a special case [24], ASGARD was the first primal–dual first-order algorithm that achieves a non-asymptotic optimal convergence rate on the last primal iterate. ASGARD is also different from the alternating direction method of multipliers (ADMM) and its variants, where it does not require solving complex subproblems but uses the proximal operators of f and g^*. However, ASGARD (i.e. [23, Algorithm 1]) only covers the general convex case, and it needs only one proximal operation of f and of g^* per iteration. To handle the strong convexity of f, a different variant is developed in [23], called ADSGARD, but requires two proximal operations of f per iteration. Therefore, the following natural question is arising:

  Can we develop a unified variant of ASGARD that covers three settings: general convexity, strong convexity, and strong convexity and smoothness?

Contribution   In this paper, we affirmatively answer this question by developing a unified variant of ASGARD that covers the following three settings:

Case 1: Both f and g^* in (1) are only convex, but not necessarily strongly convex.
Case 2: Either f or g^* is strongly convex, but not both f and g^*.
Case 3: Both f and g^* are strongly convex.

The new variant only requires one proximal operation of f and of g^* at each iteration as in the original ASGARD and existing primal–dual methods, e.g., in [4,5,7,8,11, 12,19,28]. Our algorithm reduces to ASGARD in Case 1, but uses a different update rule for η_k (see Step 5 of Algorithm 1) compared to ASGARD. In Cases 2 and 3, our algorithm is completely new by incorporating the strong convexity parameter μ_f of f and/or μ_{g^*} of g^* in the parameter update rules to achieve optimal $\mathcal{O}\left(1/k^2\right)$ sublinear

and $\left(1 - \mathcal{O}\left(1/\sqrt{\kappa_F}\right)\right)$ linear rates, respectively, where k is the iteration counter and $\kappa_F := \|K\|^2/(\mu_f \mu_{g^*})$.

In terms of convergence guarantees, we establish that, in all cases, our algorithm achieves optimal rates on the last primal sequence $\{x^k\}$ and an averaging dual sequence $\{\tilde{y}^k\}$. Moreover, our convergence guarantees are on three different criteria: (i) the gap function for (1), (ii) the primal objective residual $F(x^k) - F^\star$ for (2), and (iii) the dual objective residual $D(\tilde{y}^k) - D^\star$ for (3). Our paper therefore provides a unified and full analysis on convergence rates of ASGARD for solving three problems (1), (2), and (3) simultaneously.

We emphasize that primal–dual first-order methods for solving (1), (2), and (3), and their convergence analysis have been well studied in the literature. To avoid overloading this paper, we refer to our recent works [23,25] for a more thorough discussion and comparison between existing methods. Hitherto, there have been numerous papers studying convergence rates of primal–dual first-order methods, including [4–7,9,14,27]. However, the best known and optimal rates are only achieved via averaging or weighted averaging sequences, which are also known as ergodic rates. The convergence rates on the last iterate sequence are often slower and suboptimal. Recently, the optimal convergence rates of the last iterates have been studied for primal–dual first-order methods, including [23,25,27]. As pointed out in [5,10,21,23,25], the last iterate convergence guarantee is very important in various applications to maintain some desirable structures of the final solutions such as sparsity, low-rankness, or sharp-edgedness in images. This also motivates us to develop ASGARD.

Paper outline   The rest of this paper is organized as follows. Section 2 recalls some basic concepts, states our assumptions, and characterizes the optimality condition of (1). Section 3 presents our main results on the algorithm and its convergence analysis. Section 4 provides a set of experiments to verify our theoretical results and compare our method with Nesterov's smoothing scheme in [17]. Some technical proofs are deferred to the "Appendix".

2 Basic concepts, assumptions, and optimality condition

We are working with Euclidean spaces, $\mathbb{R}^p$ and $\mathbb{R}^n$, equipped with the standard inner product $\langle \cdot, \cdot \rangle$ and the Euclidean norm $\|\cdot\|$. We will use the Euclidean norm for the entire paper. Given a proper, closed, and convex function f, we use $\mathrm{dom}\,(f)$ and $\partial f(x)$ to denote its domain and its subdifferential at x, respectively. We also use $\nabla f(x)$ for a subgradient or the gradient (if f is differentiable) of f at x. We denote by $f^*(y) := \sup\{\langle y, x \rangle - f(x) : x \in \mathrm{dom}\,(f)\}$ the Fenchel conjugate of f. We denote by $\mathrm{ri}(\mathcal{X})$ the relative interior of $\mathcal{X}$.

A function f is called M_f-Lipschitz continuous if $|f(x) - f(\tilde{x})| \leq M_f \|x - \tilde{x}\|$ for all $x, \tilde{x} \in \mathrm{dom}\,(f)$, where $M_f \in [0, +\infty)$ is called a Lipschitz constant. A proper, closed, and convex function f is M_f-Lipschitz continuous if and only if $\partial f(\cdot)$ is uniformly bounded by M_f on $\mathrm{dom}\,(f)$. For a smooth function f, we say that f is L_f-smooth (or Lipschitz gradient) if for any $x, \tilde{x} \in \mathrm{dom}\,(f)$, we have

$\|\nabla f(x) - \nabla f(\tilde{x})\| \leq L_f \|x - \tilde{x}\|$, where $L_f \in [0, +\infty)$ is a Lipschitz constant. A function f is called μ_f-strongly convex with a strong convexity parameter $\mu_f \geq 0$ if $f(\cdot) - \frac{\mu_f}{2} \| \cdot \|^2$ remains convex. For a proper, closed, and convex function f, $\text{prox}_{\gamma f}(x) := \text{argmin} \left\{ f(\tilde{x}) + \frac{1}{2\gamma} \|\tilde{x} - x\|^2 : \tilde{x} \in \text{dom}(f) \right\}$ is called the proximal operator of γf, where $\gamma > 0$.

2.1 Basic assumptions and optimality condition

In order to show the relationship between (1), (2) and its dual form (3), we require the following assumptions.

Assumption 1 The following assumptions hold for (1).

(a) Both functions f and g are proper, closed, and convex on their domain.
(b) There exists a saddle-point $z^\star := (x^\star, y^\star)$ of $\mathcal{L}$ defined in (1), i.e.:

$$\mathcal{L}(x^\star, y) \leq \mathcal{L}(x^\star, y^\star) \leq \mathcal{L}(x, y^\star), \quad \forall (x, y) \in \text{dom}(f) \times \text{dom}(g^*). \quad (4)$$

(c) The Slater condition $0 \in \text{ri}(\text{dom}(g) - K\text{dom}(f))$ holds.

Assumption 1 is standard in convex–concave saddle-point settings. Under Assumption 1, strong duality holds, i.e. $F^\star = \mathcal{L}(x^\star, y^\star) = -D^\star$.

To characterize a saddle-point of (1), we define the following gap function:

$$\mathcal{G}_{\mathcal{X} \times \mathcal{Y}}(x, y) := \sup \left\{ \mathcal{L}(x, \tilde{y}) - \mathcal{L}(\tilde{x}, y) : \tilde{x} \in \mathcal{X}, \ \tilde{y} \in \mathcal{Y} \right\}, \quad (5)$$

where $\mathcal{X} \subseteq \text{dom}(f)$ and $\mathcal{Y} \subseteq \text{dom}(g^*)$ are nonempty, closed, and convex subsets such that $\mathcal{X} \times \mathcal{Y}$ contains a saddle-point $(x^\star, y^\star)$ of (1). Clearly, we have $\mathcal{G}_{\mathcal{X} \times \mathcal{Y}}(x, y) \geq 0$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Moreover, if $(x^\star, y^\star)$ is a saddle-point of (1) in $\mathcal{X} \times \mathcal{Y}$, then $\mathcal{G}_{\mathcal{X} \times \mathcal{Y}}(x^\star, y^\star) = 0$.

2.2 Smoothing technique for *g*

We first smooth g in (2) using Nesterov's smoothing technique [17] as

$$g_\beta(u, \dot{y}) := \max_{y \in \mathbb{R}^n} \left\{ \langle u, y \rangle - g^*(y) - \frac{\beta}{2} \|y - \dot{y}\|^2 \right\}, \quad (6)$$

where g^* is the Fenchel conjugate of g, $\beta > 0$ is a smoothness parameter, and $\dot{y}$ is a given proximal center. We denote by $\nabla_u g_\beta(u) = \text{prox}_{g^*/\beta}(\dot{y} + \frac{1}{\beta}u)$ the gradient of g_β w.r.t. u.

Given g_β defined by (6), we can approximate F in (2) by

$$F_\beta(x, \dot{y}) := f(x) + g_\beta(Kx, \dot{y}). \quad (7)$$

The following lemma provides two key inequalities to link F_β to $\mathcal{L}$ and D.

Lemma 1 *Let F_β be defined by* (7) *and* $(x^\star, y^\star)$ *be a saddle-point of* (1). *Then, for any* $x, \bar{x} \in \text{dom}(f)$ *and* $\dot{y}, \tilde{y}, y \in \text{dom}(g^*)$, *we have*

$$
\begin{aligned}
\mathcal{L}(\bar{x}, y) &\leq F_\beta(\bar{x}, \dot{y}) + \tfrac{\beta}{2}\|y - \dot{y}\|^2, \\
D(\tilde{y}) - D(y^\star) &\leq F_\beta(\bar{x}, \dot{y}) - \mathcal{L}(\tilde{x}, \tilde{y}) + \tfrac{\beta}{2}\|y^\star - \dot{y}\|^2, \quad \forall \tilde{x} \in \partial f^*(-K^\top \tilde{y}).
\end{aligned}
\tag{8}
$$

Proof Using the definition of $\mathcal{L}$ in (1) and of F_β in (7), we have $\mathcal{L}(\bar{x}, y) = f(\bar{x}) + \langle K\bar{x}, y\rangle - g^*(y) \leq f(\bar{x}) + \sup_y\{\langle K\bar{x}, y\rangle - g^*(y) - \tfrac{\beta}{2}\|y - \dot{y}\|^2\} + \tfrac{\beta}{2}\|y - \dot{y}\|^2 = F_\beta(\bar{x}, \dot{y}) + \tfrac{\beta}{2}\|y - \dot{y}\|^2$, which proves the first line of (8).

Next, for any $\tilde{x} \in \partial f^*(-K^\top \tilde{y})$, we have $f^*(-K^\top \tilde{y}) = \langle -K^\top \tilde{y}, \tilde{x}\rangle - f(\tilde{x})$ by Fenchel's equality [1]. Hence, $D(\tilde{y}) = f^*(-K^\top \tilde{y}) + g^*(\tilde{y}) = -\mathcal{L}(\tilde{x}, \tilde{y})$. On the other hand, by (4), we have $\mathcal{L}(\bar{x}, y^\star) \geq \mathcal{L}(x^\star, y^\star) = -D(y^\star)$. Combining these two expressions and the first line of (8), we obtain $D(\tilde{y}) - D(y^\star) \leq \mathcal{L}(\bar{x}, y^\star) - \mathcal{L}(\tilde{x}, \tilde{y}) \leq F_\beta(\bar{x}, \dot{y}) - \mathcal{L}(\tilde{x}, \tilde{y}) + \tfrac{\beta}{2}\|\dot{y} - y^\star\|^2$. □

3 New ASGARD variant and its convergence guarantees

In this section, we derive a new and unified variant of ASGARD in [23] and analyze its convergence rate guarantees for three settings.

3.1 The derivation of algorithm and one-iteration analysis

Given $\dot{y} \in \mathbb{R}^n$ and $\hat{x}^k \in \mathbb{R}^p$, the main step of ASGARD consists of one primal and one dual updates as follows:

$$
\begin{cases}
y^{k+1} := \text{prox}_{g^*/\beta_k}\big(\dot{y} + \tfrac{1}{\beta_k}K\hat{x}^k\big), \\
x^{k+1} := \text{prox}_{f/L_k}\big(\hat{x}^k - \tfrac{1}{L_k}K^\top y^{k+1}\big),
\end{cases}
\tag{9}
$$

where $\beta_k > 0$ is the smoothness parameter of g, and $L_k > 0$ is an estimate of the Lipschitz constant of ∇g_{β_k}. Here, (9) serves as basic steps of various primal–dual first-order methods in the literature, including [5,14,23]. However, instead of updating $\dot{y}$, ASGARD fixes it for all iterations.

The following lemma serves as a key step for our analysis in the sequel. Since its statement and proof are rather different from [23, Lemma 2], we provide its proof in "Appendix B.1".

Lemma 2 [23] *Let* (x^{k+1}, y^{k+1}) *be generated by* (9), F_β *be defined by* (7), *and* $\mathcal{L}$ *be given by* (1). *Then, for any* $x \in \text{dom}(f)$ *and* $\tau_k \in [0, 1]$, *we have*

$$
\begin{aligned}
F_{\beta_k}(x^{k+1}, \dot{y}) \leq {} & (1 - \tau_k) F_{\beta_{k-1}}(x^k, \dot{y}) + \tau_k \mathcal{L}(x, y^{k+1}) \\
& + \tfrac{L_k \tau_k^2}{2} \left\| \tfrac{1}{\tau_k}[\hat{x}^k - (1 - \tau_k)x^k] - x \right\|^2 - \tfrac{\mu_f(1-\tau_k)\tau_k}{2} \|x - x^k\|^2 \\
& - \tfrac{\tau_k^2}{2} \left(L_k + \mu_f \right) \left\| \tfrac{1}{\tau_k}[x^{k+1} - (1 - \tau_k)x^k] - x \right\|^2 \\
& - \tfrac{L_k}{2} \|x^{k+1} - \hat{x}^k\|^2 + \tfrac{1}{2(\mu_{g^*}+\beta_k)} \|K(x^{k+1} - \hat{x}^k)\|^2 \\
& - \tfrac{1-\tau_k}{2} \left[\tau_k \beta_k - (\beta_{k-1} - \beta_k) \right] \|\nabla_u g_{\beta_k}(Kx^k, \dot{y}) - \dot{y}\|^2.
\end{aligned}
\tag{10}
$$

Together with the primal–dual step (9), we also apply Nesterov's accelerated step to $\hat{x}^k$ and an averaging step to $\tilde{y}^k$ as follows:

$$
\begin{cases}
\hat{x}^{k+1} := x^{k+1} + \eta_{k+1}(x^{k+1} - x^k), \\
\tilde{y}^{k+1} := (1 - \tau_k)\tilde{y}^k + \tau_k y^{k+1},
\end{cases}
\tag{11}
$$

where $\tau_k \in (0, 1)$ and $\eta_{k+1} > 0$ will be determined in the sequel.

To analyze the convergence of the new ASGARD variant, we define the following Lyapunov function (also called a potential function):

$$
\begin{aligned}
\mathcal{V}_k(x) := {} & F_{\beta_{k-1}}(x^k, \dot{y}) - \mathcal{L}(x, \tilde{y}^k) \\
& + \tfrac{(L_{k-1}+\mu_f)\tau_{k-1}^2}{2} \| \tfrac{1}{\tau_{k-1}}[x^k - (1 - \tau_{k-1})x^{k-1}] - x\|^2.
\end{aligned}
\tag{12}
$$

The following lemma provides a key recursive estimate to analyze the convergence of (9) and (11), whose proof is given in "Appendix B.2".

Lemma 3 *Let $(x^k, \hat{x}^k, y^k, \tilde{y}^k)$ be updated by (9) and (11). Given $\beta_0 > 0$, τ_k, $\tau_{k+1} \in (0, 1]$, let β_k, L_k, and η_{k+1} be updated by*

$$
\beta_k := \frac{\beta_{k-1}}{1 + \tau_k}, \quad L_k := \frac{\|K\|^2}{\beta_k + \mu_{g^*}}, \quad and \quad \eta_{k+1} := \frac{(1 - \tau_k)\tau_k}{\tau_k^2 + m_{k+1}\tau_{k+1}},
\tag{13}
$$

where $m_{k+1} := \frac{L_{k+1}+\mu_f}{L_k+\mu_f}$. Suppose further that $\tau_k \in (0, 1]$ satisfies

$$
\begin{cases}
(L_{k-1} + \mu_f)(1 - \tau_k)\tau_{k-1}^2 + \mu_f(1 - \tau_k)\tau_k & \geq L_k \tau_k^2, \\
(L_{k-1} + \mu_f)(L_k + \mu_f)\tau_k \tau_{k-1}^2 + (L_k + \mu_f)^2 \tau_k^2 \geq (L_{k-1} + \mu_f)L_k \tau_{k-1}^2.
\end{cases}
\tag{14}
$$

Then, for any $x \in \mathrm{dom}\,(f)$, the Lyapunov function $\mathcal{V}_k$ defined by (12) satisfies

$$
\mathcal{V}_{k+1}(x) \leq (1 - \tau_k)\mathcal{V}_k(x).
\tag{15}
$$

The unified ASGARD algorithm Our next step is to expand (9), (11), and (13) algorithmically to obtain a new ASGARD variant (called ASGARD+) as presented in Algorithm 1.

Algorithm 1 (New Accelerated Smoothed GAp ReDuction (ASGARD+))

1: **Initialization:** Choose $x^0 \in \text{dom}(f)$, $\tilde{y}^0 \in \text{dom}(g^*)$, and $\dot{y} \in \mathbb{R}^n$.

2: Choose $\tau_0 \in (0, 1]$ and $\beta_0 > 0$. Set $L_0 := \frac{\|K\|^2}{\mu_{g^*} + \beta_0}$ and $\hat{x}^0 := x^0$.

3: **For** $k := 0, 1, \cdots, k_{\max}$ **do**

4:   Update τ_{k+1} as in Theorems 1, 2, or 3, and update $\beta_{k+1} := \frac{\beta_k}{1+\tau_{k+1}}$.

5:   Let $L_{k+1} := \frac{\|K\|^2}{\mu_{g^*} + \beta_{k+1}}$, $m_{k+1} := \frac{L_{k+1} + \mu_f}{L_k + \mu_f}$, and $\eta_{k+1} := \frac{(1-\tau_k)\tau_k}{\tau_k^2 + m_{k+1}\tau_{k+1}}$.

6:   Update
$$
\begin{cases}
y^{k+1} := \text{prox}_{g^*/\beta_k}\left(\dot{y} + \frac{1}{\beta_k} K\hat{x}^k\right), \\
x^{k+1} := \text{prox}_{f/L_k}\left(\hat{x}^k - \frac{1}{L_k} K^\top y^{k+1}\right), \\
\hat{x}^{k+1} := x^{k+1} + \eta_{k+1}(x^{k+1} - x^k), \\
\tilde{y}^{k+1} := (1 - \tau_k)\tilde{y}^k + \tau_k y^{k+1}.
\end{cases}
$$

7: **EndFor**

Compared to the original ASGARD in [23], Algorithm 1 requires one additional averaging dual step on $\tilde{y}^k$ at Step 6 to obtain the dual convergence. Note that Algorithm 1 also incorporates the strong convexity parameters μ_f of f and μ_{g^*} of g^* to cover three settings: general convexity ($\mu_f = \mu_{g^*} = 0$), strong convexity ($\mu_f > 0$ and $\mu_{g^*} = 0$), and strong convexity and smoothness ($\mu_f > 0$ and $\mu_{g^*} > 0$). More precisely, L_k and the momentum step-size η_{k+1} are also different from [23] by incorporating μ_{g^*} and μ_f. The per-iteration complexity of ASGARD+ remains the same as ASGARD except for the averaging dual update $\tilde{y}^k$. However, this step is not required if we only solve (2). We highlight that if we apply a new approach from [25] to (1), then we can also update the proximal center $\dot{y}$ at each iteration.

3.2 Case 1: Both f and g^* are just convex ($\mu_f = \mu_{g^*} = 0$)

The following theorem establishes convergence rates of Algorithm 1 for the general convex case where both f and g^* are just convex.

Theorem 1 *Suppose that Assumption 1 holds and both f and g^* are only convex, i.e. $\mu_f = \mu_{g^*} = 0$. Let $\{(x^k, \tilde{y}^k)\}$ be generated by Algorithm 1 for solving (1), where $\tau_0 := 1$ and τ_{k+1} is the unique solution of the cubic equation $\tau^3 + \tau^2 + \tau_k^2\tau - \tau_k^2 = 0$ in τ, which always exists. Then, for all $k \geq 1$, we have:*

(a) *The gap function $\mathcal{G}_{\mathcal{X} \times \mathcal{Y}}$ defined by (5) satisfies*

$$
\mathcal{G}_{\mathcal{X} \times \mathcal{Y}}(x^k, \tilde{y}^k) \leq \frac{\|K\|^2}{2\beta_0 k} \sup_{x \in \mathcal{X}} \|x^0 - x\|^2 + \frac{\beta_0}{k+1} \sup_{y \in \mathcal{Y}} \|y - \dot{y}\|^2. \tag{16}
$$

(b) *If g is M_g-Lipschitz continuous on* dom (g), *then for* (2), *it holds that*

$$F(x^k) - F^\star \leq \frac{\|K\|^2}{2\beta_0 k} \|x^0 - x^\star\|^2 + \frac{\beta_0}{k+1}(\|\dot{y}\| + M_g)^2. \qquad (17)$$

(c) *If f^* is M_{f^*}-Lipschitz continuous on* dom (f^*), *then for* (3), *it holds that*

$$D(\tilde{y}^k) - D^\star \leq \frac{\|K\|^2}{2\beta_0 k} (\|x^0\| + M_{f^*})^2 + \frac{\beta_0}{k+1} \|\dot{y} - y^\star\|^2. \qquad (18)$$

Proof First, since $\mu_f = \mu_{g^*} = 0$, $L_k = \frac{\|K\|^2}{\beta_k}$, and $(1 + \tau_k)\beta_k = \beta_{k-1}$, the two conditions of (14) respectively reduce to

$$(1 - \tau_k)\tau_{k-1}^2 \geq (1 + \tau_k)\tau_k^2 \quad \text{and} \quad (1 + \tau_k)\tau_k^2 \geq \tau_{k-1}^2(1 - \tau_k).$$

These conditions hold if $\tau_k^3 + \tau_k^2 + \tau_{k-1}^2\tau_k - \tau_{k-1}^2 = 0$. We first choose $\tau_0 := 1$, and update τ_k by solving the cubic equation $\tau^3 + \tau + \tau_{k-1}^2\tau - \tau_{k-1}^2 = 0$ for $k \geq 1$. Note that this equation has a unique positive real solution $\tau_k \in (0, 1)$ due to Lemma 5(b). Moreover, we have $\prod_{i=1}^k (1 - \tau_i) \leq \frac{1}{k+1}$ and $\beta_k \leq \frac{2\beta_0}{k+2}$.

Next, by induction, (15) leads to $\mathcal{V}_k(x) \leq \left[\prod_{i=1}^{k-1}(1 - \tau_i)\right] \mathcal{V}_1(x) \leq \frac{1}{k}\mathcal{V}_1(x)$, where we have used $\prod_{i=1}^{k-1}(1 - \tau_i) \leq \frac{1}{k}$ from Lemma 5(b). However, from (45) in the proof of Lemma 3 and $\tau_0 = 1$, we have $\mathcal{V}_1(x) \leq (1 - \tau_0)\mathcal{V}_0(x) + \frac{L_0\tau_0^2}{2}\|\hat{x}^0 - x\|^2 = \frac{\|K\|^2}{2\beta_0}\|x^0 - x\|^2$. Hence, we eventually obtain

$$\mathcal{V}_k(x) \leq \frac{\|K\|^2}{2\beta_0 k}\|x^0 - x\|^2. \qquad (19)$$

Using (8) from Lemma 1 and $\beta_{k-1} \leq \frac{2\beta_0}{k+1}$ from Lemma 5(b), we get

$$\begin{aligned}
\mathcal{L}(x^k, y) - \mathcal{L}(x, \tilde{y}^k) &\overset{(8)}{\leq} F_{\beta_{k-1}}(x^k, \dot{y}) - \mathcal{L}(x, \tilde{y}^k) + \frac{\beta_{k-1}}{2}\|y - \dot{y}\|^2 \\
&\overset{(12)}{\leq} \mathcal{V}_k(x) + \frac{\beta_{k-1}}{2}\|y - \dot{y}\|^2 \\
&\overset{(19)}{\leq} \frac{\|K\|^2}{2\beta_0 k}\|x^0 - x\|^2 + \frac{\beta_0}{k+1}\|\dot{y} - y\|^2.
\end{aligned}$$

Taking the supreme over $\mathcal{X}$ and $\mathcal{Y}$ both sides of the last estimate and using (5), we obtain (16).

Now, since $F_{\beta_{k-1}}(x^k, \dot{y}) - F^\star \overset{(4)}{\leq} F_{\beta_{k-1}}(x^k, \dot{y}) - \mathcal{L}(x^\star, \tilde{y}^k) \overset{(12)}{\leq} \mathcal{V}_k(x^\star)$, combining this inequality and (19), we get

$$F_{\beta_{k-1}}(x^k, \dot{y}) - F^\star \leq \frac{\|K\|^2}{2\beta_0 k}\|x^0 - x^\star\|^2.$$

On the other hand, since g is M_g-Lipschitz continuous, we have

$$\sup\left\{\|y - \dot{y}\| : y \in \partial g(Kx^k)\right\} \le \|\dot{y}\| + \sup\{\|y\| : \|y\| \le M_g\} = \|\dot{y}\| + M_g.$$

Hence, by (31) of Lemma 4, we have $F(x^k) \le F_{\beta_{k-1}}(x^k, \dot{y}) + \frac{\beta_{k-1}}{2}(\|\dot{y}\| + M_g)^2 \le F_{\beta_{k-1}}(x^k, \dot{y}) + \frac{\beta_0}{k+1}(\|\dot{y}\| + M_g)^2$. Combining both estimates, we obtain (17).

Finally, using (8), we have

$$
\begin{aligned}
D(\tilde{y}^k) - D^\star &\le F_{\beta_{k-1}}(x^k, \dot{y}) - \mathcal{L}(\tilde{x}^k, \tilde{y}^k) + \frac{\beta_{k-1}}{2}\|\dot{y} - y^\star\|^2 \\
&\overset{(12)}{\le} \mathcal{V}_k(\tilde{x}^k) + \frac{\beta_0}{k+1}\|\dot{y} - y^\star\|^2 \\
&\overset{(19)}{\le} \frac{\|K\|^2}{2\beta_0 k}(\|x^0\| + M_{f^*})^2 + \frac{\beta_0}{k+1}\|\dot{y} - y^\star\|^2,
\end{aligned}
$$

which proves (18). Here, since $\tilde{x}^k \in \partial f^*(-K^\top \tilde{y}^k)$, we have $\|x^0 - \tilde{x}^k\| \le \|\tilde{x}^k\| + \|x^0\| \le M_{f^*} + \|x^0\|$, which has been used in the last inequality. □

3.3 Case 2: f is strongly convex and g^* is convex ($\mu_f > 0$ and $\mu_{g^*} = 0$)

Next, we consider the case when only f or g^* is strongly convex. Without loss of generality, we assume that f is strongly convex with a strong convexity parameter $\mu_f > 0$, but g^* is only convex with $\mu_{g^*} = 0$. Otherwise, we switch the role of f and g^* in Algorithm 1.

The following theorem establishes an optimal $\mathcal{O}\left(1/k^2\right)$ convergence rate (up to a constant factor) of Algorithm 1 in this case (i.e. $\mu_f > 0$ and $\mu_{g^*} = 0$).

Theorem 2 *Suppose that Assumption 1 holds and that f is strongly convex with a convexity parameter $\mu_f > 0$, but g^* is just convex (i.e. $\mu_{g^*} = 0$). Let us choose $\tau_0 := 1$ and $\beta_0 \ge \frac{0.382\|K\|^2}{\mu_f}$. Let $\{(x^k, \tilde{y}^k)\}$ be generated by Algorithm 1 using the update rule $\tau_{k+1} := \frac{\tau_k}{2}\left(\sqrt{\tau_k^2 + 4} - \tau_k\right)$ for τ_k. Then, we have:*

(a) *The gap function $\mathcal{G}_{\mathcal{X} \times \mathcal{Y}}$ be defined by (5) satisfies*

$$\mathcal{G}_{\mathcal{X} \times \mathcal{Y}}(x^k, \tilde{y}^k) \le \frac{2\|K\|^2}{\beta_0(k+1)^2} \sup_{x \in \mathcal{X}} \|x^0 - x\|^2 + \frac{10\beta_0}{(k+3)^2} \sup_{y \in \mathcal{Y}} \|\dot{y} - y\|^2. \quad (20)$$

(b) *If g is M_g-Lipschitz continuous on $\mathrm{dom}\,(g)$, then for (2), it holds that*

$$F(x^k) - F^\star \le \frac{2\|K\|^2\|x^0 - x^\star\|^2}{\beta_0(k+1)^2} + \frac{10\beta_0(\|\dot{y}\| + M_g)^2}{(k+3)^2}. \quad (21)$$

(c) *If f^* is M_{f^*}-Lipschitz continuous on $\mathrm{dom}\,(f^*)$, then for (3), it holds that*

$$D(\tilde{y}^k) - D^\star \le \frac{2\|K\|^2(\|x^0\| + M_{f^*})^2}{\beta_0(k+1)^2} + \frac{10\beta_0\|\dot{y} - y^\star\|^2}{(k+3)^2}. \quad (22)$$

Proof Since $\mu_{g^*} = 0$ and $\tau_k^2 = \tau_{k-1}^2(1 - \tau_k)$ by the update rule of τ_k, $\beta_k = \frac{\beta_{k-1}}{1+\tau_k}$, and $L_k = \frac{\|K\|^2}{\beta_k}$, the first condition of (14) is equivalent to $\beta_{k-1} \geq \frac{\|K\|^2\tau_k^2}{\mu_f}$. However, since $\beta_{k-1} \geq \frac{\beta_0\tau_k^2}{\tau_1^2}$ due to Lemma 5(a), and $\tau_1 = 0.6180$, $\beta_{k-1} \geq \frac{\|K\|^2\tau_k^2}{\mu_f}$ holds if $\beta_0 \geq \frac{\|K\|^2\tau_1^2}{\mu_f} = \frac{0.382\|K\|^2}{\mu_f}$. Thus we can choose $\beta_0 \geq \frac{0.382\|K\|^2}{\mu_f}$ to guarantee the first condition of (14).

Similarly, using $m_k = \frac{L_k+\mu_f}{L_{k-1}+\mu_f} \geq 1$, the second condition of (14) is equivalent to $m_k^2\tau_k^2 + m_k\tau_k\tau_{k-1}^2 \geq \frac{L_k\tau_{k-1}^2}{L_{k-1}+\mu_f}$. Since $\frac{L_k}{L_{k-1}+\mu_f} \leq m_k$, the last condition holds if $m_k^2\tau_k^2 + m_k\tau_k\tau_{k-1}^2 \geq m_k\tau_{k-1}^2$. Using again $\tau_k^2 = \tau_{k-1}^2(1 - \tau_k)$, this condition becomes $m_k\tau_k^2 \geq \tau_{k-1}^2(1 - \tau_k) = \tau_k^2$. This always holds true since $m_k \geq 1$. Therefore, the second condition of (14) is satisfied.

As a result, we have the recursive estimate (15), i.e.:

$$\mathcal{V}_{k+1}(x) \leq (1 - \tau_k)\mathcal{V}_k(x). \tag{23}$$

From (27), Lemma 5(a), (45), and noting that $\tilde{x}^0 := x^0$ and $\tau_0 := 1$, we have

$$\mathcal{V}_k(x) \leq \left[\prod_{i=1}^{k-1}(1 - \tau_i)\right]\mathcal{V}_1(x) = \Theta_{1,k-1}\mathcal{V}_1(x) \leq \frac{4}{(k+1)^2}\mathcal{R}_0(x).$$

where $\mathcal{R}_0(x) := \frac{\|K\|^2}{2\beta_0}\|x^0 - x\|^2$. Similar to the proof of Theorem 1, using the last inequality and $\beta_{k-1} \leq \frac{4\beta_0\tau_0^2}{\tau_k^2[\tau_0(k+1)+2]^2} \leq \frac{20\beta_0}{(k+3)^2}$ from Lemma 5(a), we obtain the bounds (20), (21), and (22), respectively. □

Remark 1 The variant of Algorithm 1 in Theorem 2 is completely different from [23, Algorithm 2] and [25, Algorithm 2], where it requires only one $\text{prox}_{f/L_k}(\cdot)$ as opposed to two proximal operations of f as in [23,25].

3.4 Case 3: Both f and g^* are strongly convex ($\mu_f > 0$ and $\mu_{g^*} > 0$)

Finally, we assume that both f and g^* are strongly convex with strong convexity parameters $\mu_f > 0$ and $\mu_{g^*} > 0$, respectively. Then, the following theorem proves the optimal linear rate (up to a constant factor) of Algorithm 1.

Theorem 3 *Suppose that Assumption 1 holds and both f and g^* in (1) are strongly convex with $\mu_f > 0$ and $\mu_{g^*} > 0$, respectively. Let $\{(x^k, \tilde{y}^k)\}$ be generated by Algorithm 1 using $\tau_k := \tau = \frac{1}{\sqrt{1+\kappa_F}} \in (0, 1)$ and $\beta_0 > 0$, where $\kappa_F := \frac{\|K\|^2}{\mu_f\mu_{g^*}}$. Then, the following statements hold:*

(a) *The gap function $\mathcal{G}_{\mathcal{X}\times\mathcal{Y}}$ defined by (5) satisfies*

$$\mathcal{G}_{\mathcal{X}\times\mathcal{Y}}(x^k, \tilde{y}^k) \leq (1 - \tau)^k\bar{\mathcal{R}}_p + \frac{1}{2(1 + \tau)^k} \sup_{y\in\mathcal{Y}} \|\dot{y} - y\|^2, \tag{24}$$

$$\text{where } \bar{\mathcal{R}}_p := \sup_{x \in \mathcal{X}} \left\{ (1 - \tau) \left[\mathcal{F}_{\beta_0}(x^0) - \mathcal{L}(x, \tilde{y}^0) \right] + \frac{\|K\|^2 \tau^2}{2(\mu_{g^*} + \beta_0)} \|x^0 - x\|^2 \right\}.$$

(b) *If g is M_g-Lipschitz continuous on* dom (g), *then for* (2), *it holds that*

$$F(x^k) - F^\star \le (1 - \tau)^k \bar{\mathcal{R}}_p^\star + \frac{\beta_0}{2(1 + \tau)^k} \left(\|\dot{y}\| + M_g \right)^2, \tag{25}$$

$$\text{where } \bar{\mathcal{R}}_p^\star := (1 - \tau) \left[\mathcal{F}_{\beta_0}(x^0) - \mathcal{L}(x^\star, \tilde{y}^0) \right] + \frac{\|K\|^2 \tau^2}{2(\mu_{g^*} + \beta_0)} \|x^0 - x^\star\|^2.$$

(d) *If f^* is M_{f^*}-Lipschitz continuous on* dom (f^*), *then for* (3), *it holds that*

$$D(\tilde{y}^k) - D^\star \le (1 - \tau)^k \bar{\mathcal{R}}_d^\star + \frac{\beta_0 \|\dot{y} - y^\star\|^2}{2(1 + \tau)^k}, \tag{26}$$

$$\text{where } \bar{\mathcal{R}}_d^\star := (1 - \tau) \left[\mathcal{F}_{\beta_0}(x^0) - D(\tilde{y}^0) \right] + \frac{\|K\|^2 \tau^2}{2(\mu_{g^*} + \beta_0)} \left(\|x^0\| + M_{f^*} \right)^2.$$

Proof Since $\tau_k = \tau = \frac{1}{\sqrt{1 + \kappa_F}} = \sqrt{\frac{\mu_f \mu_{g^*}}{\|K\|^2 + \mu_f \mu_{g^*}}} \in (0, 1)$ and $\beta_{k-1} = (1 + \tau)\beta_k$, after a few elementary calculations, we can show that the first condition of (14) automatically holds. The second condition of (14) is equivalent to $m_k \tau + m_k^2 \ge \frac{L_k}{L_{k-1} + \mu_f}$. Since $m_k \ge \frac{L_k}{L_{k-1} + \mu_f}$, this condition holds if $m_k \tau + m_k^2 \ge m_k$, which is equivalent to $\tau + m_k \ge 1$. This obviously holds true since $\tau > 0$ and $m_k \ge 1$.

From (15) of Lemma 3, we have $\mathcal{V}_{k+1}(x) \le (1 - \tau)\mathcal{V}_k(x)$. Therefore, by induction and using again (45), we get

$$\mathcal{V}_k(x) \le (1 - \tau)^k \mathcal{V}_1(x) \le (1 - \tau)^k \bar{\mathcal{R}}_0(x). \tag{27}$$

where $\bar{\mathcal{R}}_p(x) := (1 - \tau) \left[\mathcal{F}_{\beta_0}(x^0) - \mathcal{L}(x, \tilde{y}^0) \right] + \frac{\|K\|^2 \tau^2}{2(\mu_{g^*} + \beta_0)} \|x^0 - x\|^2.$

Now, since $\beta_{k-1} = \frac{\beta_0}{(1 + \tau)^k}$ due to the update rule of β_k, by (8), we have

$$\mathcal{L}(x^k, y) - \mathcal{L}(x, \tilde{y}^k) \le \mathcal{V}_k(x) + \frac{\beta_{k-1}}{2} \|\dot{y} - y\|^2 \le (1 - \tau)^k \bar{\mathcal{R}}_p(x) + \frac{\|\dot{y} - y\|^2}{2(1 + \tau)^k}.$$

This implies (24). The estimates (25) and (26) can be proved similarly as in Theorem 1, and we omit the details here. □

Remark 2 Since g^* is μ_{g^*}-strongly convex, it is well-known that $g \circ K$ is $\frac{\|K\|^2}{\mu_{g^*}}$-smooth. Hence, $\kappa_F := \frac{\|K\|^2}{\mu_f \mu_{g^*}}$ is the condition number of F in (2). Theorem shows that Algorithm 1 can achieve a $\left(1 - \frac{1}{(1 + \sqrt{2})\sqrt{\kappa_F}} \right)^k$-linear convergence rate. Consequently, it also achieves $\mathcal{O}\left(\sqrt{\kappa_F} \log \left(\frac{1}{\varepsilon} \right) \right)$ oracle complexity to obtain an ε-primal–dual solution $(x^k, \tilde{y}^k)$. This linear rate and complexity are optimal (up to a constant factor) under the given assumptions in Theorem 3. However, Algorithm 1 is very different from existing accelerated proximal gradient methods, e.g., [2,18,26] for solving (2) since our method uses the proximal operator of g^* (and therefore, the proximal operator of g) instead of the gradient of g as in [2,18,26].

Remark 3 The $\mathcal{O}(1/k)$, $\mathcal{O}(1/k^2)$, and linear convergence rates in Theorems 1, 2, and 3, respectively are already optimal (up to a constant factor) under given assumptions as discussed, e.g., in [20,25]. The primal convergence rate on $\{F(x^k) - F^\star\}$ has been proved in [23, Theorem 4], but only for the case $\mathcal{O}(1/k)$. The convergence rates on $\{\mathcal{G}_{\mathcal{X} \times \mathcal{Y}}(x^k, \tilde{y}^k)\}$ and $\{D(\tilde{y}^k) - D^\star\}$ are new. Moreover, the convergence of the primal sequence is on the last iterate x^k, while the convergence of the dual sequence is on the averaging iterate $\tilde{y}^k$.

4 Numerical experiments

In this section, we provide four numerical experiments to verify the theoretical convergence aspects and the performance of Algorithm 1. Our algorithm is implemented in Matlab R.2019b running on a MacBook Laptop with 2.8GHz Quad-Core Intel Core i7 and 16 GB RAM. We also compare our method with Nesterov's smoothing algorithm in [17] as a baseline. We emphasize that our experiments bellow follow exactly the parameter update rules as stated in Theorems 1 and 2 without any parameter tuning trick. To further improve practical performance of Algorithm 1, one can exploit the restarting strategy in [23], where its theoretical guarantee is established in [22].

The nonsmooth and convex optimization problem we use for our experiments is the following representative model:

$$\min_{x \in \mathbb{R}^p} \left\{ F(x) := \|Kx - b\|_2 + \lambda \|x\|_1 + \frac{\rho}{2} \|x\|_2^2 \right\}, \tag{28}$$

where $K \in \mathbb{R}^{n \times p}$ is a given matrix, $b \in \mathbb{R}^n$ is also given, and $\lambda > 0$ and $\rho \geq 0$ are two given regularization parameters. The norm $\|\cdot\|_2$ is the ℓ_2-norm (or Euclidean norm). If $\rho = 0$, then (28) reduces to the square-root LASSO model proposed in [3]. If $\rho > 0$, then (28) becomes a square-root regression problem with elastic net regularizer similar to [30]. Clearly, if we define $g(y) := \|y - b\|_2$ and $f(x) := \lambda \|x\|_1 + \frac{\rho}{2} \|x\|_2^2$, then (28) can be rewritten into (2).

To generate the input data for our experiments, we first generate K from standard i.i.d. Gaussian distributions with either uncorrelated or 50% correlated columns. Then, we generate an observed vector b as $b := Kx^\natural + \mathcal{N}(0, \sigma)$, where $x^\natural$ is a predefined sparse vector and $\mathcal{N}(0, \sigma)$ stands for standard Gaussian noise with variance $\sigma = 0.05$. The regularization parameter λ to promote sparsity is chosen as suggested in [3], and the parameter ρ is set to $\rho := 0.1$. We first fix the size of problem at $p := 1000$ and $n := 350$ and choose the number of nonzero entries of $x^\natural$ to be $s := 100$. Then, for each experiment, we generate 30 instances of the same size but with different input data (K, b).

For Nesterov's smoothing method, following [17], we smooth g as

$$g_\gamma(y) := \max_{v \in \mathbb{R}^n} \left\{ \langle y - b, v \rangle - \frac{\gamma}{2} \|v\|^2 : \|v\|_2 \leq 1 \right\},$$

where $\gamma > 0$ is a smoothness parameter. In order to correctly choose γ for Nesterov's smoothing method, we first solve (28) with CVX [13] using Mosek with high precision

to get a high accurate solution $x^\star$ of (28). Then, we set $\gamma^* := \frac{\sqrt{2}\|K\|\|x^0-x^\star\|}{k_{\max}\sqrt{D_\mathcal{V}}}$ by minimizing its theoretical bound from [17] w.r.t. $\gamma > 0$, where $D_\mathcal{V} := \frac{1}{2}$ is the prox-diameter of the unit ℓ_2-norm ball, and $k_{\max}$ is the maximum number of iterations. For Algorithm 1, using (17) and $\dot{y} := 0$, we can set $\beta_0 = \beta^* := \frac{\|K\|\|x^0-x^\star\|}{M_g}$ by minimizing the right-hand side of (17) w.r.t. $\beta_0 > 0$, where $M_g := 1$. We choose $k_{\max} := 5000$ for all experiments. To see the effect of the smoothness parameters γ and β_0 on the performance of both algorithms, we also consider two variants by increasing or decreasing these parameters 10 times, respectively. More specifically, we set them as follows.

- For Nesterov's smoothing scheme, we consider two additional variants by setting $\gamma := 10\gamma^*$ and $\gamma = 0.1\gamma^*$, respectively.
- For Algorithm 1, we also consider two other variants with $\beta_0 := 10\beta^*$ and $\beta_0 := 0.1\beta^*$, respectively.

We first conduct two different experiments for the square-root LASSO model (i.e. setting $\rho := 0$ in (28)). In this case, the underlying optimization problem is non-strongly convex and fully nonsmooth.

- *Experiment 1:* We test Algorithm 1 (abbreviated by Alg. 1) and Nesterov's smoothing method (abbreviated by Nes. Alg.) on 30 problem instances with uncorrelated columns of K. Since both algorithms essentially have the same per-iteration complexity, we report the relative primal objective residual $\frac{F(x^k)-F(x^\star)}{\max\{1,|F(x^\star)|\}}$ against the number of iterations.
- *Experiment 2:* We conduct the same test on another set of 30 problem instances, but using 50% correlated columns in the input matrix K.

The results of both experiments are depicted in Fig. 1, where the left plot is for *Experiment 1* and the right plot is for *Experiment 2*. The solid line of each curve shows the mean over 30 problem samples, and the corresponding shaded area represents the sample variance of 30 problem samples (i.e. the area between the lowest and the highest deviation from the mean).

From Fig. 1, we observe that, with the choice $\beta_0 := \beta^*$ and $\gamma := \gamma^*$ as suggested by the theory, both algorithms perform best compared to other smaller or larger values of these parameters. We also see that Algorithm 1 outperforms Nesterov's smoothing scheme in both experiments. If β_0 (respectively, γ) is large, then both algorithms make good progress in early iterations, but become saturated at a given objective value in the last iterations. Alternatively, if β_0 (respectively, γ) is small, then both algorithms perform worse in early iterations, but further decrease the objective value when the number of iterations is increasing. This behavior also confirms the theoretical results stated in Theorem 1 and in [17]. In fact, if β_0 (or γ) is small, then the algorithmic stepsize is small. Hence, the algorithm makes slow progress at early iterations, but it better approximates the nonsmooth function g, leading to more accurate approximation from $F(x^k)$ to $F(x^\star)$. In contrast, if β_0 (or γ) is large, then we have a large stepsize and therefore a faster convergence rate in early iterations. However, the smoothed approximation is less accurate.

In order to test the strongly convex case in Theorem 2, we conduct two additional experiments on (28) with $\rho := 0.1$. In this case, problem (28) is strongly convex with
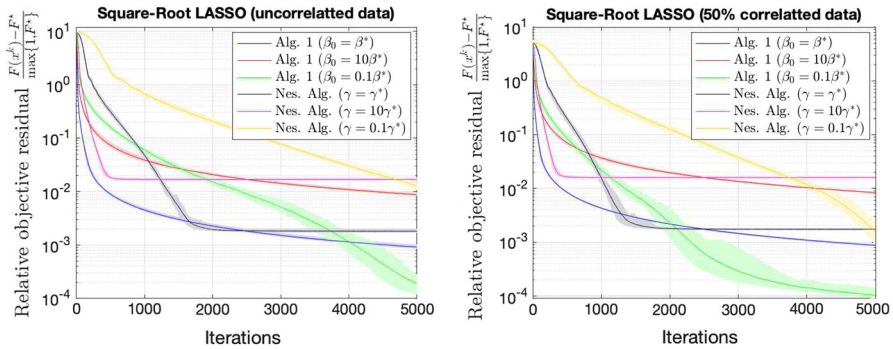
Fig. 1 The convergence behavior of Algorithm 1 and Nesterov's smoothing scheme on 30 problem instances of (28) (the non-strongly convex case). Left plot: uncorrelated columns in K, and Right plot: 50% correlated columns in K
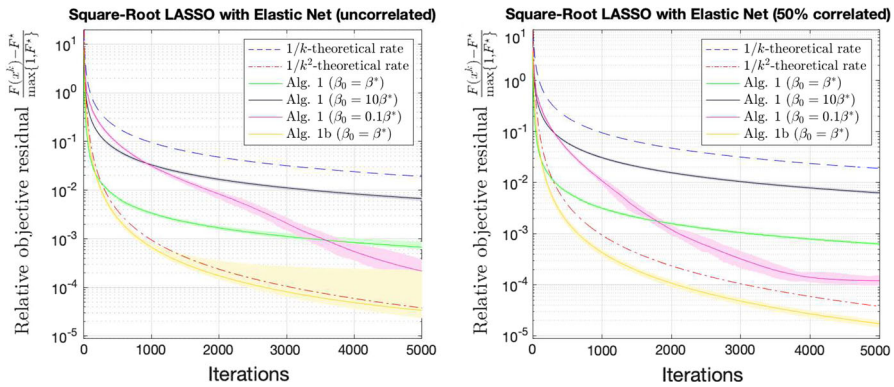
Fig. 2 The convergence behavior of the two variants of Algorithm 1 on a collection of 30 problem instances of (28) (the strongly convex case). Left plot: uncorrelated columns in K, and Right plot: 50% correlated columns in K

$\mu_f := 0.1$. Since [17] does not directly handle the strongly convex case, we only compare two variants of Algorithm 1 stated in Theorem 1 (`Alg. 1`) and Theorem 2 (`Alg. 1b`), respectively. We set $\beta_0 := \frac{0.382\|K\|^2}{\mu_f}$ in `Alg. 1b`) as suggested by Theorem 2. We consider two experiments as follows:

– *Experiment 3:* Test two variants of Algorithm 1 on a collection of 30 problem instances with uncorrelated columns of K.
– *Experiment 4:* Conduct the same test on another set of 30 problem instances, but using 50% correlated columns in K.

The results of both variants of Algorithm 1 are reported in Fig. 2, where the left plot is for *Experiment 3* and the right plot is for *Experiment 4*.

Clearly, as shown in Fig. 2, `Alg. 1b` (i.e. corresponding to Theorem 2) highly outperforms `Alg. 1` (corresponding to Theorem 1). `Alg. 1` matches well the $\mathcal{O}(1/k)$ convergence rate as stated in Theorem 1, while `Alg. 1b` shows its $\mathcal{O}(1/k^2)$ convergence rate as indicated by Theorem 2. Note that since g^* in (28) is non-strongly

convex, we omit testing the result of Theorem 3. This case is rather well studied in the literature, see, e.g., [5].

5 Concluding remarks

We have developed a new variant of ASGARD introduced in [23, Algorithm 1], Algorithm 1, that unifies three different settings: general convexity, strong convexity, and strong convexity and smoothness. We have proved the convergence of Algorithm 1 for three settings on three convergence criteria: gap function, primal objective residual, and dual objective residual. Our convergence rates in all cases are optimal up to a constant factor and the convergence rates of the primal sequence is on the last iterate. Our preliminary numerical experiments have shown that the theoretical convergence rates of Algorithm 1 match well the actual rates observed in practice. The proposed algorithm can be easily extended to solve composite convex problems with three or multi-objective terms. It can also be customized to solve other models, including general linear and nonlinear constrained convex problems as discussed in [21,23,25].

A Appendix 1: Technical lemmas

We need the following technical lemmas for our convergence analysis in the main text.

Lemma 4 ([23, Lemma 10]) *Given $\beta > 0$, $\dot{y} \in \mathbb{R}^n$, and a proper, closed, and convex function $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ with its Fenchel conjugate g^*, we define*

$$g_\beta(u, \dot{y}) := \max_{y \in \mathbb{R}^n} \left\{ \langle u, y \rangle - g^*(y) - \frac{\beta}{2} \| y - \dot{y} \|^2 \right\}. \tag{29}$$

Let $y_\beta^*(u, \dot{y})$ be the unique solution of (29). Then, the following statements hold:

(a) *$g_\beta(\cdot, \dot{y})$ is convex w.r.t. u on $\mathrm{dom}\,(g)$ and $\frac{1}{\beta + \mu_{g^*}}$-smooth w.r.t. u on $\mathrm{dom}\,(g)$, where $\nabla_u g_\beta(u, \dot{y}) = \mathrm{prox}_{g^*/\beta}(\dot{y} + \frac{1}{\beta} u)$. Moreover, for any $u, \hat{u} \in \mathrm{dom}\,(g)$, we have*

$$g_\beta(\hat{u}, \dot{y}) + \langle \nabla g_\beta(\hat{u}, \dot{y}), u - \hat{u} \rangle \leq g_\beta(u, \dot{y}) - \frac{\beta + \mu_{g^*}}{2} \| \nabla_u g_\beta(\hat{u}, \dot{y}) - \nabla_u g_\beta(u, \dot{y}) \|^2. \tag{30}$$

(b) *For any $\beta > 0$, $\dot{y} \in \mathbb{R}^n$, and $u \in \mathrm{dom}\,(g)$, we have*

$$g_\beta(u, \dot{y}) \leq g(u) \leq g_\beta(u, \dot{y}) + \frac{\beta}{2} [D_g(\dot{y})]^2, \text{ where } D_g(\dot{y}) := \sup_{y \in \partial g(u)} \| y - \dot{y} \|. \tag{31}$$

(c) *For $u \in \mathrm{dom}\,(g)$ and $\dot{y} \in \mathbb{R}^n$, $g_\beta(u, \dot{y})$ is convex in β, and for all $\hat{\beta} \geq \beta > 0$, we have*

$$g_\beta(u, \dot{y}) \leq g_{\hat{\beta}}(u, \dot{y}) + \left(\frac{\hat{\beta} - \beta}{2} \right) \| \nabla_u g_\beta(u, \dot{y}) - \dot{y} \|^2. \tag{32}$$

(d) *For any $\beta > 0$, and $u, \hat{u} \in \mathrm{dom}\,(g)$, we have*

$$g_\beta(u, \dot{y}) + \langle \nabla_u g_\beta(u, \dot{y}), \hat{u} - u \rangle \le \ell_\beta(\hat{u}, \dot{y}) - \tfrac{\beta}{2} \| \nabla_u g_\beta(u, \dot{y}) - \dot{y} \|^2, \qquad (33)$$

where $\ell_\beta(\hat{u}, \dot{y}) := \langle \hat{u}, \nabla_u g_\beta(u, \dot{y}) \rangle - g^*(\nabla_u g_\beta(u, \dot{y})) \le g(\hat{u}) - \tfrac{\mu_{g^*}}{2} \| \nabla_u g_\beta(u, \dot{y}) - \nabla g(\hat{u}) \|^2$ for any $\nabla g(\hat{u}) \in \partial g(\hat{u})$.

Lemma 5 *The following statements hold.*

(a) *Let $\{\tau_k\} \subset (0, 1]$ be computed by $\tau_{k+1} := \tfrac{\tau_k}{2}\left[(\tau_k^2 + 4)^{1/2} - \tau_k\right]$ for some $\tau_0 \in (0, 1]$. Then, we have*

$$\tau_k^2 = (1 - \tau_k)\tau_{k-1}^2, \quad \frac{1}{k + 1/\tau_0} \le \tau_k < \frac{2}{k + 2/\tau_0},$$

$$\text{and} \quad \frac{1}{1 + \tau_{k-2}} \le 1 - \tau_k \le \frac{1}{1 + \tau_{k-1}}.$$

Moreover, we also have

$$\Theta_{l,k} := \prod_{i=l}^{k}(1 - \tau_i) = \frac{\tau_k^2}{\tau_{l-1}^2} \quad \text{for } 0 \le l \le k,$$

$$\Theta_{0,k} = \frac{(1 - \tau_0)\tau_k^2}{\tau_0^2} \le \frac{4(1 - \tau_0)}{(\tau_0 k + 2)^2},$$

$$\text{and} \quad \frac{\tau_{l+1}^2}{\tau_{k+2}^2} \le \Gamma_{l,k} := \prod_{i=l}^{k}(1 + \tau_i) \le \frac{\tau_l^2}{\tau_{k+1}^2} \quad \text{for } 0 \le l \le k.$$

If we update $\beta_k := \frac{\beta_{k-1}}{1 + \tau_k}$ for a given $\beta_0 > 0$, then

$$\frac{4\beta_0 \tau_0^2}{\tau_1^2[\tau_0(k + 1) + 2]^2} \le \frac{\beta_0 \tau_{k+1}^2}{\tau_1^2} \le \beta_k = \frac{\beta_0}{\Gamma_{1,k}} \le \frac{\beta_0 \tau_{k+2}^2}{\tau_2^2} \le \frac{4\beta_0 \tau_0^2}{\tau_2^2[\tau_0(k + 2) + 2]^2}.$$

(b) *Let $\{\tau_k\} \subset (0, 1]$ be computed by solving $\tau_k^3 + \tau_k^2 + \tau_{k-1}^2 \tau_k - \tau_{k-1}^2 = 0$ for all $k \ge 1$ and $\tau_0 := 1$. Then, we have $\frac{1}{k+1} \le \tau_k \le \frac{2}{k+2}$ and $\Theta_{1,k} := \prod_{i=1}^{k}(1 - \tau_i) \le \frac{1}{k+1}$. Moreover, if we update $\beta_k := \frac{\beta_{k-1}}{1 + \tau_k}$, then $\beta_k \le \frac{2\beta_0}{k+2}$.*

Proof The first two relations of (a) have been proved, e.g., in [24]. Let us prove the last inequality of (a). Since $\frac{1}{1 + \tau_{k-2}} \le 1 - \tau_k$ is equivalent to $\tau_{k-2}(1 - \tau_k) \ge \tau_k$. Using $1 - \tau_k = \frac{\tau_k^2}{\tau_{k-1}^2}$, we have $\tau_k \tau_{k-2} \ge \tau_{k-1}^2$. Utilizing $\tau_k = \frac{\tau_{k-1}}{2}\left[(\tau_{k-1}^2 + 4)^2 - \tau_{k-1}\right]$, this condition is equivalent to $\tau_{k-2}^2 \ge \tau_{k-1}^2(1 + \tau_{k-2})$. However, since $\tau_{k-1}^2 = (1 - \tau_{k-1})\tau_{k-2}^2$, the last condition becomes $1 \ge (1 - \tau_{k-1})(1 + \tau_{k-2})$, or equivalently, $\tau_{k-1} \le \tau_{k-2}$, which automatically holds.

To prove $1 - \tau_k \leq \frac{1}{1+\tau_{k-1}}$, we write it as $\tau_{k-1}(1 - \tau_k) \leq \tau_k$. Using again $\tau_k^2 = (1-\tau_k)\tau_{k-1}^2$, the last inequality is equivalent to $\tau_k \leq \tau_{k-1}$, which automatically holds. The last statements of (a) is a consequence of $1 - \tau_k = \frac{\tau_k^2}{\tau_{k-1}^2}$ and the previous relations.

(b) We consider the function $\varphi(\tau) := \tau^3 + \tau^2 + \tau_{k-1}^2 \tau - \tau_{k-1}^2$. Clearly, $\varphi(0) = -\tau_{k-1}^2 < 0$ and $\varphi(1) = 2 > 0$. Moreover, $\varphi'(\tau) = 3\tau^2 + 2\tau + \tau_{k-1}^2 > 0$ for $\tau \in [0, 1]$. Hence, the cubic equation $\varphi(\tau) = 0$ has a unique solution $\tau_k \in (0, 1)$. Therefore, $\{\tau_k\}_{k \geq 0}$ is well-defined.

Next, since $\tau_k^3 + \tau_k^2 + \tau_k \tau_{k-1}^2 - \tau_{k-1}^2 = 0$ is equivalent to $\tau_{k-1}^2(1 - \tau_k) = \tau_k^2(1 + \tau_k)$, we have $\tau_{k-1}^2(1 - \tau_k) = \tau_k^2(1 + \tau_k) \leq \frac{\tau_k^2}{1-\tau_k}$. This inequality becomes $\tau_k \geq \frac{\tau_{k-1}}{1+\tau_{k-1}}$. By induction and $\tau_0 = 1$, we can easily show that $\tau_k \geq \frac{1}{k+1}$. On the other hand, $\tau_{k-1}^2(1 - \tau_k) = \tau_k^2(1 + \tau_k) \geq \tau_k^2$. From this inequality, with a similar argument as in the proof of the statement (a), we can also easily show that $\tau_k \leq \frac{2}{k+2}$. Hence, we have $\frac{1}{k+1} \leq \tau_k \leq \frac{2}{k+2}$ for all $k \geq 0$.

Finally, since $\tau_k \geq \frac{1}{k+1}$, we have $\prod_{i=1}^k (1 - \tau_i) \leq \prod_{i=1}^k \left(1 - \frac{1}{i+1}\right) = \frac{1}{k+1}$. Alternatively, $\prod_{i=1}^k (1 + \tau_i) \geq \prod_{i=1}^k \left(1 + \frac{1}{i+1}\right) = \frac{k+2}{2}$. However, since $\beta_k = \frac{\beta_{k-1}}{1+\tau_k}$, we have $\beta_k = \beta_0 \prod_{i=1}^k \frac{1}{1+\tau_i} \leq \frac{2\beta_0}{k+2}$. $\square$

Lemma 6 ([29, Lemma 4] and [23]) *The following statements hold.*

(a) *For any $u, v, w \in \mathbb{R}^p$ and $t_1, t_2 \in \mathbb{R}$ such that $t_1 + t_2 \neq 0$, we have*

$$t_1 \|u - w\|^2 + t_2 \|v - w\|^2 = (t_1 + t_2) \|w - \frac{1}{t_1+t_2}(t_1 u + t_2 v)\|^2 + \frac{t_1 t_2}{t_1+t_2} \|u - v\|^2.$$

(b) *For any $\tau \in (0, 1)$, $\hat{\beta}, \beta > 0$, $w, z \in \mathbb{R}^p$, we have*

$$\beta(1 - \tau) \|w - z\|^2 + \beta\tau \|w\|^2 - (1 - \tau)(\hat{\beta} - \beta) \|z\|^2 = \beta \|w - (1 - \tau)z\|^2$$
$$+ (1 - \tau)[\tau\beta - (\hat{\beta} - \beta)] \|z\|^2.$$

The following lemma is a key step to address the strongly convex case of f in (1).

Lemma 7 *Given $L_k > 0$, $\mu_f > 0$, and $\tau_k \in (0, 1)$, let $m_k := \frac{L_k + \mu_f}{L_{k-1} + \mu_f}$ and $a_k := \frac{L_k}{L_{k-1} + \mu_f}$. Assume that the following two conditions hold:*

$$\begin{cases} (1 - \tau_k)[\tau_{k-1}^2 + m_k \tau_k] \geq a_k \tau_k \\ m_k \tau_k \tau_{k-1}^2 + m_k^2 \tau_k^2 \geq a_k \tau_{k-1}^2. \end{cases} \tag{34}$$

Let $\{x^k\}$ be a given sequence in $\mathbb{R}^p$. We define $\hat{x}^k := x^k + \frac{1}{\omega_k}(x^k - x^{k-1})$, where ω_k is chosen such that

$$\max\left\{\frac{\tau_{k-1} + \sqrt{\tau_{k-1}^2 + 4a_k}}{2(1 - \tau_{k-1})}, \frac{a_k \tau_k}{(1 - \tau_k)(1 - \tau_{k-1})\tau_{k-1}}\right\} \le \omega_k \le \frac{\tau_{k-1}^2 + m_k \tau_k}{\tau_{k-1}(1 - \tau_{k-1})}. \tag{35}$$

Then, ω_k is well-defined, and for any $x \in \mathbb{R}^p$, we have

$$L_k \tau_k^2 \|\frac{1}{\tau_k}[\hat{x}^k - (1 - \tau_k)x^k] - x\|^2 - \mu_f \tau_k(1 - \tau_k)\|x^k - x\|^2$$
$$\le (1 - \tau_k)\left(L_{k-1} + \mu_f\right)\tau_{k-1}^2\|\frac{1}{\tau_{k-1}}[x^k - (1 - \tau_{k-1})x^{k-1}] - x\|^2. \tag{36}$$

Proof Firstly, from the definition $\hat{x}^k := x^k + \frac{1}{\omega_k}(x^k - x^{k-1})$ of $\hat{x}^k$, we have $\omega_k(\hat{x}^k - x^k) = x^k - x^{k-1}$. Hence, we can show that

$$\tau_{k-1}^2\|\frac{1}{\tau_{k-1}}[x^k - (1 - \tau_{k-1})x^{k-1}] - x\|^2 = \|(1 - \tau_{k-1})(x^k - x^{k-1}) + \tau_{k-1}(x^k - x)\|^2$$
$$= \|(1 - \tau_{k-1})\omega_k(\hat{x}^k - x^k) + \tau_{k-1}(x^k - x)\|^2$$
$$= \omega_k^2(1 - \tau_{k-1})^2\|\hat{x}^k - x^k\|^2 + \tau_{k-1}^2\|x^k - x\|^2$$
$$+ 2\omega_k(1 - \tau_{k-1})\tau_{k-1}\langle\hat{x}^k - x^k, x^k - x\rangle.$$

Alternatively, we also have

$$\tau_k^2\|\frac{1}{\tau_k}[\hat{x}^k - (1 - \tau_k)x^k] - x\|^2 = \|\hat{x}^k - x^k\|^2 + \tau_k^2\|x^k - x\|^2 + 2\tau_k\langle\hat{x}^k - x^k, x^k - x\rangle.$$

Utilizing the two last expressions, (36) can be rewritten equivalently to

$$\mathcal{T}_{[1]} := 2\left[\left(L_{k-1} + \mu_f\right)(1 - \tau_k)(1 - \tau_{k-1})\tau_{k-1}\omega_k - L_k\tau_k\right]\langle\hat{x}^k - x^k, x - x^k\rangle$$
$$\le \left[\left(L_{k-1} + \mu_f\right)(1 - \tau_k)(1 - \tau_{k-1})^2\omega_k^2 - L_k\right]\|\hat{x}^k - x^k\|^2$$
$$+ \left[\left(L_{k-1} + \mu_f\right)(1 - \tau_k)\tau_{k-1}^2 - L_k\tau_k^2 + \mu_f\tau_k(1 - \tau_k)\right]\|x^k - x\|^2.$$

Now, let us denote

$$\begin{cases} c_1 := \left(L_{k-1} + \mu_f\right)(1 - \tau_k)(1 - \tau_{k-1})\tau_{k-1}\omega_k - L_k\tau_k \\ c_2 := \left(L_{k-1} + \mu_f\right)(1 - \tau_k)(1 - \tau_{k-1})^2\omega_k^2 - L_k \\ c_3 := \left(L_{k-1} + \mu_f\right)(1 - \tau_k)\tau_{k-1}^2 - L_k\tau_k^2 + \mu_f(1 - \tau_k)\tau_k. \end{cases}$$

Then, (36) is equivalent to

$$2c_1\langle\hat{x}^k - x^k, x - x^k\rangle \le c_2\|\hat{x}^k - x^k\|^2 + c_3\|x - x^k\|^2. \tag{37}$$

Secondly, we need to guarantee that $c_1 \geq 0$. This condition holds if we choose ω_k such that

$$\omega_k \geq \frac{a_k \tau_k}{(1 - \tau_k)(1 - \tau_{k-1})\tau_{k-1}}. \tag{38}$$

Thirdly, we also need to guarantee $c_2 \geq c_1$, which is equivalent to

$$c_2 - c_1 = \left(L_{k-1} + \mu_f \right)(1 - \tau_k)(1 - \tau_{k-1})\left[(1 - \tau_{k-1})\omega_k^2 - \tau_{k-1}\omega_k \right] - L_k(1 - \tau_k) \geq 0.$$

This condition holds if

$$\omega_k \geq \frac{\tau_{k-1} + \sqrt{\tau_{k-1}^2 + 4a_k}}{2(1 - \tau_{k-1})}. \tag{39}$$

Alternatively, we also need to guarantee $c_3 \geq c_1$, which is equivalent to

$$c_3 - c_1 = \left(L_{k-1} + \mu_f \right)(1 - \tau_k)\left[\tau_{k-1}^2 - (1 - \tau_{k-1})\tau_{k-1}\omega_k \right] + (L_k + \mu_f)\tau_k(1 - \tau_k) \geq 0.$$

This condition holds if

$$\omega_k \leq \frac{\tau_{k-1}^2 + m_k \tau_k}{\tau_{k-1}(1 - \tau_{k-1})}. \tag{40}$$

Combining (38), (39), and (40), we obtain

$$\max \left\{ \frac{\tau_{k-1} + \sqrt{\tau_{k-1}^2 + 4a_k}}{2(1 - \tau_{k-1})}, \ \frac{a_k \tau_k}{(1 - \tau_k)(1 - \tau_{k-1})\tau_{k-1}} \right\} \leq \omega_k \leq \frac{\tau_{k-1}^2 + m_k \tau_k}{\tau_{k-1}(1 - \tau_{k-1})},$$

which is exactly (35). Here, under the condition (34), the left-hand side of the last expression is less than or equal to the right-hand side. Therefore, ω_k is well-defined.

Finally, under the choice of ω_k as in (35), we have $c_2 \geq c_1 \geq 0$ and $c_3 \geq c_1 \geq 0$. Hence, (37) holds, which is also equivalent to (36). □

B Appendix 2: Technical proof of Lemmas 2 and 3 in Sect. 3

This section provides the full proof of Lemmas 2 and 3 in the main text.

B.1 The proof of Lemma 2: key estimate of the primal–dual step (9)

Proof From the first line of (9) and Lemma 4(a), we have $\nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}) = K^\top y^{k+1}$. Now, from the second line of (9), we also have

$$0 \in \partial f(x^{k+1}) + L_k(x^{k+1} - \hat{x}^k) + K^\top \nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}).$$

Combining this inclusion and the μ_f-convexity of f, for any $x \in \text{dom}(f)$, we get

$$f(x^{k+1}) \le f(x) + \langle \nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}), K(x - x^{k+1}) \rangle + L_k \langle x^{k+1} - \hat{x}^k, x - x^{k+1} \rangle$$
$$- \frac{\mu_f}{2} \|x^{k+1} - x\|^2.$$

Since $g_\beta(\cdot, \dot{y})$ is $\frac{1}{\beta + \mu_{g*}}$-smooth by Lemma 4(a), for any $x \in \text{dom}(f)$, we have

$$g_{\beta_k}(Kx^{k+1}, \dot{y}) \le g_{\beta_k}(K\hat{x}^k, \dot{y}) + \langle \nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}), K(x^{k+1} - \hat{x}^k) \rangle$$
$$+ \frac{1}{2(\beta_k + \mu_{g*})} \|K(x^{k+1} - \hat{x}^k)\|^2$$
$$= g_{\beta_k}(K\hat{x}^k, \dot{y}) + \langle \nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}), K(x - \hat{x}^k) \rangle$$
$$- \langle \nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}), K(x - x^{k+1}) \rangle$$
$$+ \frac{1}{2(\mu_{g*} + \beta_k)} \|K(x^{k+1} - \hat{x}^k)\|^2.$$

Now, combining the last two estimates, we get

$$f(x^{k+1}) + g_{\beta_k}(Kx^{k+1}, \dot{y}) \le f(x) + g_{\beta_k}(K\hat{x}^k, \dot{y}) + \langle \nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}), K(x - \hat{x}^k) \rangle$$
$$+ L_k \langle x^{k+1} - \hat{x}^k, x - \hat{x}^k \rangle - L_k \|x^{k+1} - \hat{x}^k\|^2$$
$$+ \frac{1}{2(\mu_{g*} + \beta_k)} \|K(x^{k+1} - \hat{x}^k)\|^2 - \frac{\mu_f}{2} \|x - x^{k+1}\|^2. \tag{41}$$

Using Lemma 4(a) again, we have

$$\ell_{\beta_k}(x^k, \dot{y}) := g_{\beta_k}(K\hat{x}^k, \dot{y}) + \langle \nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}), K(x^k - \hat{x}^k) \rangle$$
$$\le g_{\beta_k}(Kx^k, \dot{y}) - \frac{\beta_k + \mu_{g*}}{2} \|\nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}) - \nabla_u g_{\beta_k}(Kx^k, \dot{y})\|^2. \tag{42}$$

Substituting $x := x^k$ into (41), and multiplying the result by $1 - \tau_k$ and adding the result to (41) after multiplying it by τ_k, then using (42), we can derive

$$F_{\beta_k}(x^{k+1}, \dot{y}) := f(x^{k+1}) + g_{\beta_k}(Kx^{k+1}, \dot{y})$$
$$\le (1 - \tau_k)[f(x^k) + g_{\beta_k}(Kx^k, \dot{y})] + \tau_k \left[f(x) + \ell_{\beta_k}(x, \dot{y}) \right]$$
$$- L_k \|x^{k+1} - \hat{x}^k\|^2 + \frac{1}{2(\mu_{g*} + \beta_k)} \|K(x^{k+1} - \hat{x}^k)\|^2$$
$$+ L_k \langle x^{k+1} - \hat{x}^k, \tau_k x - \hat{x}^k + (1 - \tau_k) x^k \rangle$$
$$- \frac{\mu_f}{2} \left[(1 - \tau_k) \|x^{k+1} - x^k\|^2 + \tau_k \|x - x^{k+1}\|^2 \right]$$
$$- \frac{(1 - \tau_k)(\beta_k + \mu_{g*})}{2} \|\nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}) - \nabla_u g_{\beta_k}(Kx^k, \dot{y})\|^2. \tag{43}$$

From Lemma 6(a), we can easily show that

$$(1 - \tau_k) \|x^{k+1} - x^k\|^2 + \tau_k \|x^{k+1} - x\|^2 = \tau_k^2 \|\frac{1}{\tau_k}[x^{k+1} - (1 - \tau_k)x^k] - x\|^2$$
$$+ \tau_k (1 - \tau_k) \|x - x^k\|^2.$$

We also have the following elementary relation

$$\langle x^{k+1} - \hat{x}^k, \tau_k x - [\hat{x}^k - (1 - \tau_k)x^k] \rangle = \frac{\tau_k^2}{2} \| \frac{1}{\tau_k}[\hat{x}^k - (1 - \tau_k)x^k] - x \|^2 + \frac{1}{2}\|x^{k+1} - \hat{x}^k\|^2$$
$$- \frac{\tau_k^2}{2} \| \frac{1}{\tau_k}[x^{k+1} - (1 - \tau_k)x^k] - x \|^2.$$

Substituting the two last expressions into (43), we obtain

$$
\begin{aligned}
F_{\beta_k}(x^{k+1}, \dot{y}) \leq {} & (1 - \tau_k)F_{\beta_k}(x^k, \dot{y}) + \tau_k \left[f(x) + \ell_{\beta_k}(x, \dot{y}) \right] \\
& + \frac{L_k \tau_k^2}{2} \| \frac{1}{\tau_k}[\hat{x}^k - (1 - \tau_k)x^k] - x \|^2 \\
& - \frac{\tau_k^2}{2}\left(L_k + \mu_f\right) \| \frac{1}{\tau_k}[x^{k+1} - (1 - \tau_k)x^k] - x \|^2 \\
& - \frac{(1-\tau_k)(\mu_{g^*}+\beta_k)}{2} \| \nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}) - \nabla_u g_{\beta_k}(Kx^k, \dot{y}) \|^2 \\
& - \frac{L_k}{2}\|x^{k+1} - \hat{x}^k\|^2 + \frac{1}{2(\mu_{g^*}+\beta_k)} \| K(x^{k+1} - \hat{x}^k) \|^2 \\
& - \frac{\mu_f(1-\tau_k)\tau_k}{2}\|x - x^k\|^2.
\end{aligned}
\tag{44}
$$

One the one hand, by (32) of Lemma 4, we have

$$F_{\beta_k}(x^k, \dot{y}) \leq F_{\beta_{k-1}}(x^k, \dot{y}) + \frac{(\beta_{k-1} - \beta_k)}{2}\|\nabla_u g_{\beta_k}(Kx^k, \dot{y}) - \dot{y}\|^2.$$

On the other hand, by (33) of Lemma 4, we get

$$f(x) + \ell_{\beta_k}(x, \dot{y}) \leq \mathcal{L}(x, y^{k+1}) - \frac{\beta_k}{2}\|\nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}) - \dot{y}\|^2,$$

where $\mathcal{L}(x, y^{k+1}) := f(x) + \langle Kx, y^{k+1} \rangle - g^*(y^{k+1})$ is the Lagrange function in (1).

Now, substituting the last two inequalities into (44), and using Lemma 6(b) with $w := \nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}) - \dot{y}$ and $z := \nabla_u g_{\beta_k}(Kx^k, \dot{y}) - \dot{y}$, we arrive at

$$
\begin{aligned}
F_{\beta_k}(x^{k+1}, \dot{y}) \leq {} & (1 - \tau_k)F_{\beta_{k-1}}(x^k, \dot{y}) + \tau_k \mathcal{L}(x, y^{k+1}) + \frac{L_k \tau_k^2}{2} \| \frac{1}{\tau_k}[\hat{x}^k - (1 - \tau_k)x^k] - x \|^2 \\
& - \frac{\tau_k^2}{2}\left(L_k + \mu_f\right) \| \frac{1}{\tau_k}[x^{k+1} - (1 - \tau_k)x^k] - x \|^2 - \frac{\mu_f(1-\tau_k)\tau_k}{2}\|x - x^k\|^2 \\
& - \frac{L_k}{2}\|x^{k+1} - \hat{x}^k\|^2 + \frac{1}{2(\mu_{g^*}+\beta_k)} \| K(x^{k+1} - \hat{x}^k) \|^2 \\
& - \frac{(1-\tau_k)}{2}\left[\tau_k \beta_k - (\beta_{k-1} - \beta_k)\right]\|\nabla_u g_{\beta_k}(Kx^k, \dot{y}) - \dot{y}\|^2 \\
& - \frac{\beta_k}{2}\|\nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}) - \dot{y} - (1 - \tau_k)\left[\nabla_u g_{\beta_k}(Kx^k, \dot{y}) - \dot{y}\right]\|^2 \\
& - \frac{(1-\tau_k)\mu_{g^*}}{2}\|\nabla_u g_{\beta_k}(K\hat{x}^k, \dot{y}) - \nabla_u g_{\beta_k}(Kx^k, \dot{y})\|^2.
\end{aligned}
$$

By dropping the last two nonpositive terms in the last inequality, we obtain (10). □

B.2 The proof of Lemma 3: recursive estimate of the Lyapunov function

Proof First, from the last line $\tilde{y}^{k+1} = (1 - \tau_k)\tilde{y}^k + \tau_k y^{k+1}$ of (11), and the μ_{g^*}-convexity of g^*, we have

$$\mathcal{L}(x, \tilde{y}^{k+1}) := f(x) + \langle Kx, \tilde{y}^{k+1}\rangle - g^*(\tilde{y}^{k+1})$$
$$\geq (1 - \tau_k)\mathcal{L}(x, \tilde{y}^k) + \tau_k \mathcal{L}(x, y^{k+1}) + \frac{\mu_{g^*}\tau_k(1-\tau_k)}{2}\|y^{k+1} - \tilde{y}^k\|^2.$$

Hence, $\tau_k \mathcal{L}(x, y^{k+1}) \leq \mathcal{L}(x, \tilde{y}^{k+1}) - (1 - \tau_k)\mathcal{L}(x, \tilde{y}^k) - \frac{\mu_{g^*}\tau_k(1-\tau_k)}{2}\|y^{k+1} - \tilde{y}^k\|^2$.
Substituting this estimate into (10) and dropping the term $-\frac{\mu_{g^*}\tau_k(1-\tau_k)}{2}\|y^{k+1} - \tilde{y}^k\|^2$, we can derive

$$F_{\beta_k}(x^{k+1}, \dot{y}) \leq (1 - \tau_k)F_{\beta_{k-1}}(x^k, \dot{y}) + \mathcal{L}(x, \tilde{y}^{k+1}) - (1 - \tau_k)\mathcal{L}(x, \tilde{y}^k)$$
$$+ \frac{L_k \tau_k^2}{2}\left\|\frac{1}{\tau_k}[\hat{x}^k - (1 - \tau_k)x^k] - x\right\|^2$$
$$- \frac{\tau_k^2}{2}\left(L_k + \mu_f\right)\left\|\frac{1}{\tau_k}[x^{k+1} - (1 - \tau_k)x^k] - x\right\|^2 \qquad (45)$$
$$- \frac{L_k}{2}\|x^{k+1} - \hat{x}^k\|^2 + \frac{1}{2(\mu_{g^*}+\beta_k)}\|K(x^{k+1} - \hat{x}^k)\|^2$$
$$- \frac{\mu_f \tau_k(1-\tau_k)}{2}\|x^k - x\|^2.$$

Now, it is obvious to show that the condition (14) is equivalent to the condition (34) of Lemma 7. In addition, we choose $\eta_k = \frac{1}{\omega_k}$ in our update (13), where $\omega_k := \frac{\tau_{k-1}^2 + m_k \tau_k}{\tau_{k-1}(1-\tau_{k-1})}$, which is the upper bound of (35). Hence, (35) automatically holds. Using (36), we have

$$\mathcal{T}_{[2]} := \frac{L_k \tau_k^2}{2}\left\|\frac{1}{\tau_k}[\hat{x}^k - (1 - \tau_k)x^k] - x\right\|^2 - \frac{\mu_f \tau_k(1-\tau_k)}{2}\|x^k - x\|^2$$
$$\leq \frac{\tau_{k-1}^2}{2}(1 - \tau_k)\left(L_{k-1} + \mu_f\right)\left\|\frac{1}{\tau_{k-1}}[x^k - (1 - \tau_{k-1})x^{k-1}] - x\right\|^2.$$

Moreover, $\frac{1}{2(\mu_{g^*}+\beta_k)}\|K(x^{k+1} - \hat{x}^k)\|^2 \leq \frac{\|K\|^2}{2(\mu_{g^*}+\beta_k)}\|x^{k+1} - \hat{x}^k\|^2 = \frac{L_k}{2}\|x^{k+1} - \hat{x}^k\|^2$ due to the definition of L_k in (13). Substituting these two estimates into (45), and utilizing the definition (12) of $\mathcal{V}_k$, we obtain (15). $\qquad\square$

References

1. Bauschke, H.H., Combettes, P.: Convex Analysis and Monotone Operators Theory in Hilbert Spaces, 2nd edn. Springer, Berlin (2017)
2. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
3. Belloni, A., Chernozhukov, V., Wang, L.: Square-root LASSO: Pivotal recovery of sparse signals via conic programming. Biometrika **94**(4), 791–806 (2011)
4. Boţ, R.I., Böhm, A.: Variable smoothing for convex optimization problems using stochastic gradients. J. Sci. Comput. **85**(2), 1–29 (2020)

5. Chambolle, A., Pock, T.: A first-order primal–dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
6. Chambolle, A., Pock, T.: On the ergodic convergence rates of a first-order primal–dual algorithm. Math. Program. **159**(1–2), 253–287 (2016)
7. Chen, Y., Lan, G., Ouyang, Y.: Optimal primal–dual methods for a class of saddle-point problems. SIAM J. Optim. **24**(4), 1779–1814 (2014)
8. Condat, L.: A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**, 460–479 (2013)
9. Davis, D.: Convergence rate analysis of primal–dual splitting schemes. SIAM J. Optim. **25**(3), 1912–1943 (2015)
10. Davis, D., Yin, W.: A three-operator splitting scheme and its optimization applications. Set-Valued Var. Anal. **25**(4), 829–858 (2017)
11. Esser, E., Zhang, X., Chan, T.: A general framework for a class of first order primal–dual algorithms for TV-minimization. SIAM J. Imaging Sci. **3**(4), 1015–1046 (2010)
12. Goldstein, T., Esser, E., Baraniuk, R.: Adaptive primal-dual hybrid gradient methods for saddle point problems. Technical Report, pp. 1–26 (2013). arxiv: 1305.0546v1.pdf
13. Grant, M.: Disciplined Convex Programming. Ph.D. thesis, Stanford University (2004)
14. He, B.S., Yuan, X.M.: On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method. SIAM J. Numer. Anal. **50**, 700–709 (2012)
15. Nemirovskii, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley Interscience, London (1983)
16. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, Volume 87 of Applied Optimization. Kluwer Academic Publishers, London (2004)
17. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
18. Nesterov, Y.: Gradient methods for minimizing composite objective function. Math. Program. **140**(1), 125–161 (2013)
19. O'Connor, D., Vandenberghe, L.: Primal-dual decomposition by operator splitting and applications to image deblurring. SIAM J. Imaging Sci. **7**(3), 1724–1754 (2014)
20. Ouyang, Y., Xu, Y.: Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. Math. Program. **185**, 1–35 (2019)
21. Sabach, S., Teboulle, M.: Faster Lagrangian-based methods in convex optimization (2020). arXiv preprint arXiv:2010.14314
22. Tran-Dinh, Q., Alacaoglu, A., Fercoq, O., Cevher, V.: An adaptive primal–dual framework for nonsmooth convex minimization. Math. Program. Compt. **12**, 451–491 (2020)
23. Tran-Dinh, Q., Fercoq, O., Cevher, V.: A smooth primal–dual optimization framework for nonsmooth composite convex minimization. SIAM J. Optim. **28**(1), 96–134 (2018)
24. Tran-Dinh, Q., Savorgnan, C., Diehl, M.: Combining Lagrangian decomposition and excessive gap smoothing technique for solving large-scale separable convex optimization problems. Comput. Optim. Appl. **55**(1), 75–111 (2013)
25. Tran-Dinh, Q., Zhu, Y.: Non-stationary first-order primal–dual algorithms with faster convergence rates. SIAM J. Optim. **30**(4), 2866–2896 (2020)
26. Tseng, P.: On accelerated proximal gradient methods for convex–concave optimization. SIAM J. Optim. (2008)
27. Valkonen, T.: Inertial, corrected, primal–dual proximal splitting. SIAM J. Optim. **30**(2), 1391–1420 (2020)
28. Vu, B.C.: A variable metric extension of the forward–backward–forward algorithm for monotone operators. Numer. Funct. Anal. Optim. **34**(9), 1050–1065 (2013)
29. Zhu, Y., Liu, D., Tran-Dinh, Q.: Primal–dual algorithms for a class of nonlinear compositional convex optimization problems, pp. 1–26 (2020). arXiv preprint arXiv:2006.09263
30. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B **67**(2), 301–320 (2005)