NON-STATIONARY FIRST-ORDER PRIMAL-DUAL ALGORITHMS WITH FASTER CONVERGENCE RATES

QUOC TRAN-DINH* AND YUZIXUAN ZHU*

Abstract. In this paper, we propose two novel non-stationary first-order primal-dual algorithms to solve nonsmooth composite convex optimization problems. Unlike existing primal-dual schemes where the parameters are often fixed, our methods use pre-defined and dynamic sequences for parameters. We prove that our first algorithm can achieve $\mathcal{O}(1/k)$ convergence rate on the primal-dual gap, and primal and dual objective residuals, where k is the iteration counter. Our rate is on the non-ergodic (i.e., the last iterate) sequence of the primal problem and on the ergodic (i.e., the averaging) sequence of the dual problem, which we call semi-ergodic rate. By modifying the step-size update rule, this rate can be boosted even faster on the primal objective residual. When the problem is strongly convex, we develop a second primal-dual algorithm that exhibits $\mathcal{O}\left(1/k^2\right)$ convergence rate on the same three types of guarantees. Again by modifying the step-size update rule, this rate becomes faster on the primal objective residual. Our primal-dual algorithms are the first ones to achieve such fast convergence rate guarantees under mild assumptions compared to existing works, to the best of our knowledge. As byproducts, we apply our algorithms to solve constrained convex optimization problems and prove the same convergence rates on both the objective residuals and the feasibility violation. We still obtain at least $\mathcal{O}\left(1/k^2\right)$ rates even when the problem is "semi-strongly" convex. We verify our theoretical results via two well-known numerical examples.

Keywords: Non-stationary primal-dual method; non-ergodic convergence rate; fast convergence rates; composite convex minimization; constrained convex optimization.

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

1. Introduction.

Problem statement. In this paper, we develop new first-order primal-dual algorithms to solve the following classical composite convex minimization problem:

$$F^\star := \min_{x\in\mathbb{R}^p}\Big\{F(x) := f(x) + g(Kx)\Big\}, \tag{1}$$

where $f : \mathbb{R}^p \to \mathbb{R}\cup\{+\infty\}$ and $g : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ are two proper, closed, and convex functions, and $K : \mathbb{R}^p \to \mathbb{R}^n$ is a given general linear operator. Associated with the primal problem (1), we also consider its dual form as

$$G^\star := \min_{y\in\mathbb{R}^n}\Big\{G(y) := f^*(-K^\top y) + g^*(y)\Big\}, \tag{2}$$

where f^* and g^* are the Fenchel conjugates of f and g, respectively. We can combine the primal and dual problems (1) and (2) into the following min-max setting:

$$\min_{x\in\mathbb{R}^p}\max_{y\in\mathbb{R}^n}\Big\{\widetilde{\mathcal{L}}(x,y) := f(x) + \langle Kx, y\rangle - g^*(y)\Big\}, \tag{3}$$

where $\widetilde{\mathcal{L}}(x,y)$ can be referred to as the Lagrange function of (1) and (2), see [2].

A brief overview of primal-dual methods. The study of first-order primal-dual methods for solving (1) and (2) has become extremely active in recent years, ranging from algorithmic development and convergence theory to applications, see, e.g., [2, 11, 27, 29]. This type of methods has close connection to other fields such as monotone inclusions, variational inequalities, and game theory [2, 28]. They also

*Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill (UNC), 318-Hanes Hall, Chapel Hill, NC, 27599-3260, USA.
Corresponding author: quoctd@email.unc.edu.

have various direct applications in image and signal processing, machine learning, statistics, economics, and engineering, see, e.g., [10, 17, 26].

In our view, the study of first-order primal-dual methods for convex optimization can be divided into three main streams. The first one is algorithmic development with numerous variants using different frameworks such as fixed-point theory, projective methods, monotone operator splitting, Fenchel duality and augmented Lagrangian frameworks, and variational inequality, see, e.g., [8, 13, 14, 17, 26, 31, 32, 37, 39, 40, 52, 63, 64, 67, 68]. Among different primal-dual variants for convex optimization, the general primal-dual hybrid gradient (PDHG) method proposed in [10, 26, 51, 68] appears to be the most general scheme that covers many existing variants, as investigated in [11, 30, 49]. Using an appropriate reformulation of (1), [49] showed that the general PDHG scheme is in fact equivalent to Douglas-Rachford's splitting method [2, 25, 36], and, therefore, to ADMM in the dual setting. Extensions to three operators and three objective functions have also been studied in several works, including [5, 18, 23, 61]. Other extensions to non-bilnear terms, Bregman distances, multi-objective terms, and stochastic variants have been also intensively investigated, see, e.g., in [2, 9, 44, 53, 58, 65, 66].

The second stream is convergence analysis. Existing works often use a gap function to measure the optimality of given approximate solutions [10, 44]. This approach usually combines both primal and dual variables in one and uses, e.g., variational inequality frameworks to prove convergence, see, e.g., [25, 32, 39, 40]. An algorithmic-independent framework to characterize primal-dual gap certificates can be found in [24]. Together with asymptotic convergence and linear convergence rates, many researchers have recently focused on sublinear convergence rates under weaker assumptions than strong convexity and smoothness or strong monotonicity-type and Lipschitz continuity conditions, see [4, 5, 12, 14, 20, 22, 32, 33, 39, 40, 56, 63] for more details. We emphasize that in general convex settings, such convergence rates are often achieved via averaging sequences on both primal and dual variables, which are much faster and easier to derive than on the sequence of last iterates.

The third stream is applications, especially in image and signal processing, see, e.g., [10, 11, 15, 16, 27, 48]. Recently, many primal-dual methods have also been applied to solving problems from machine learning, statistics, and engineering, see, e.g., [11, 29]. While theoretical results have shown that primal-dual methods may suffer from slow sublinear convergence rates under mild assumptions, their empirical convergence rates are much better on concrete applications [10, 26].

Motivation. In many applications, the desired solutions often have special structures such as sharp-edgedness in images, sparsity in signal processing and model selection, and low-rankness in matrix approximation. Such structures can be modeled using regularizers, constraints, or penalty functions, but unfortunately can be destroyed by algorithms that use ***ergodic*** (i.e., averaging or weighted averaging) sequences as outputs, which is one of the reasons why many algorithms eventually take the ***non-ergodic*** (i.e., last iterate) sequence as output while ignoring the fact that their convergence rate guarantee is proved based on an ergodic sequence. In addition, as observed in [23], the last-iterate sequence often has fast empirical convergence rate (e.g., up to linear rate). This mismatch between theory and practice motivates us to develop new primal-dual algorithms that return the last iterates as outputs with rigorous convergence rate guarantees. While non-ergodic convergence guarantees have recently been discussed in [19, 22] for several methods, it did not achieve the *optimal* rate. This paper develops two new first-order primal-dual schemes to fulfill this gap by using dynamic step-sizes, which leads to non-stationary methods, where the term

"non-stationary" is adopted from [35] for Douglas-Rachford methods.

Whereas $\mathcal{O}(1/k)$ convergence rate appears to be optimal under only convexity and strong duality assumptions when $k \leq \mathcal{O}(p)$, faster convergence rate for $k > \mathcal{O}(p)$ in primal-dual methods seems to not be known yet, especially in non-ergodic sense. Recently, [1] showed that Nesterov's accelerated method can exhibit up to $o\left(1/k^2\right)$ convergence rate when k is sufficiently large compared to the problem dimension p. This rate can only be achieved if g has Lipschitz continuous gradient. This motivates us to consider such an acceleration in first-order primal-dual methods by adopting the approach in [1]. We show $\underline{o}\big(1/(k\sqrt{\log k})\big)$ non-ergodic convergence rate on the objective residual sequence in the sense that $\liminf_{k\to\infty}(k\sqrt{\log k})[F(x^k) - F^\star] = 0$ (*cf.* (5)) without any smoothness or strong convexity-type assumption. A similar type of rate is also proved in [20, 22] with $o\big(1/\sqrt{k}\big)$ rate under the same assumption as ours, and $o(1/k)$ rate under additional assumption of strong convexity or smoothness (our non-ergodic rates are both $\mathcal{O}\left(1/k^2\right)$ and $\underline{o}\big(1/(k^2\sqrt{\log k})\big)$ in this case).

Our contributions. To this end, our contributions are summarized as follows:

(a) We develop a new first-order primal-dual scheme, Algorithm 1, to solve primal and dual problems (1) and (2). We prove the $\mathcal{O}(1/k)$ optimal convergence rate on three criteria: primal-dual gap, primal objective residual, and dual objective residual under only convexity and strong duality assumptions. Our guarantee is achieved in semi-ergodic sense, i.e., non-ergodic in primal variable and ergodic in dual variable. For sufficiently large k (i.e., $k > \mathcal{O}(p)$), by modifying the parameter update rules, we can show that our algorithm can be boosted up to $\min\{\mathcal{O}(1/k)\,,\ \underline{o}(1/(k\sqrt{\log k}))\}$ non-ergodic convergence rate on the primal objective residual. This rate is not slower than $\mathcal{O}(1/k)$ and empirically significantly faster than its counterpart with only $\mathcal{O}(1/k)$ rate.

(b) If we apply Algorithm 1 to solve nonsmooth constrained convex optimization problems, then we can prove the same $\mathcal{O}(1/k)$ and $\underline{o}\big(1/(k\sqrt{\log k})\big)$ convergence rates on the primal objective residual and the feasibility violation.

(c) If f of (1) is strongly convex (or equivalently, its Fenchel conjugate f^* is L-smooth), then we propose another first-order primal-dual algorithm, Algorithm 2, which achieves the $\mathcal{O}\left(1/k^2\right)$ optimal convergence rate on the same three criteria as of Algorithm 1. When k is sufficiently large (i.e., $k > \mathcal{O}(p)$), by modifying the parameter update rules of Algorithm 2, we obtain up to $\min\{\mathcal{O}\left(1/k^2\right)\,,\ \underline{o}(1/(k^2\sqrt{\log k}))\}$ non-ergodic convergence rates on (1).

(d) If we modify Algorithm 2 to solve the constrained convex problem (38), where the objective is semi-strongly convex (i.e., one objective term is strongly convex while the other term is non-strongly convex), then we prove the same $\mathcal{O}\left(1/k^2\right)$ and $\underline{o}\big(1/(k^2\sqrt{\log k})\big)$ rates for both the primal objective residual and the feasibility violation.

Comparison. We highlight some key differences between our algorithms and existing methods in terms of approach, algorithmic appearance, and theoretical guarantees. First, unlike existing augmented Lagrangian-based methods, we view the augmented term as a smoothed term for the indicator of linear constraints in the constrained reformulation (7) of (1). Next, we apply Nesterov's accelerated scheme to minimize this smoothed Lagrange function and simultaneously update the smoothness parameter (i.e., the penalty parameter) at each iteration in a homotopy fashion.

Second, Algorithm 1 has similar structure as Chambolle-Pock's method [10, 12, 49], a special case of PDHG, but it possesses a three-point momentum step depending on the iterates at the iterations k, $k-1$, and $k-2$, and makes use of dynamic parameters and step-sizes. Algorithm 2 uses two proximal operators of the primal objective

to obtain a non-ergodic convergence rate (but not required, see Subsection 4.2).

Third, unlike existing works where the best-known convergence rates are often obtained via ergodic sequences, see, e.g., [12, 19, 22, 32, 33, 39, 40], our methods achieve the optimal convergence rates in non-ergodic sense. The $\mathcal{O}(1/k)$ ergodic optimal rate of primal-dual methods for solving (1) is not new and has been proved in many papers. Their non-ergodic rates have just recently been proved, e.g., in [54, 55, 56, 60]. More precisely, [55, 56] utilize the Nesterov's smoothing technique in [47] and only derive primal convergence rates. [54] only handles constrained problems by applying the quadratic penalty function approach. [60] relies on the well-known Chambolle-Pock scheme in [10] by adding inertial correction terms and adapting the parameters to achieve non-ergodic rates. Nevertheless, our algorithm in this paper uses a completely different approach and achieves the $\mathcal{O}(1/k)$ rate on three criteria.

Finally, in addition to the $\mathcal{O}(1/k)$ non-ergodic rate on the primal objective residual, we also prove its $\underline{o}\big(1/(k\sqrt{\log k})\big)$ non-ergodic rate. In comparison, [19] provides an intensive analysis of convergence rates for several methods to solve a more general problem than (1). However, [19] does not provide new algorithms, and its convergence rate if applied to (1) would become $o\big(1/\sqrt{k}\big)$. Other related works include [20, 21, 22, 23]. Table 1 non-exhaustively summarizes the best-known convergence rates of first-order primal-dual methods for solving (1), where we highlight that this paper contributes the fastest rates under corresponding assumptions.

TABLE 1

State-of-the-art results and our contributions to convergence rates of the primal objective residual sequence $\{F(x) - F^\star\}$ of first-order primal-dual algorithms for solving (1). Here, f and g are convex and possibly nonsmooth, and g is Lipschitz continuous. In addition, we consider the assumption where f is strongly convex.

Assumption	Convergence type	Convergence rate	References
convex f and g	ergodic	$\mathcal{O}(1/k)$	[6, 10, 19, 20, 21, 31, 38, 39, 40, 50], etc.
	non-ergodic	$\mathcal{O}(1/k)$	[54, 55, 56, 57, 60] and **this work**
	non-ergodic	$o\big(1/\sqrt{k}\big)$	[19, 20, 21]
	non-ergodic	$\min\big\{\mathcal{O}(1/k), \underline{o}\big(1/(k\sqrt{\log k})\big)\big\}$	**this work**
strongly convex f or g^*	ergodic	$\mathcal{O}\big(1/k^2\big)$	[10, 31, 38, 39, 50], etc.
	non-ergodic	$\mathcal{O}\big(1/k^2\big)$	[54, 55, 56, 57, 60] and **this work**
	best-iterate	$o(1/k)$	[20, 22]
	non-ergodic	$\min\big\{\mathcal{O}\big(1/k^2\big), \underline{o}\big(1/(k^2\sqrt{\log k})\big)\big\}$	**this work**

Paper organization. The rest of this paper is organized as follows. Section 2 reviews some preliminary tools used in the sequel. Section 3 develops a new algorithm for the general convex case, investigates its convergence rate guarantees, and applies it to solve constrained convex problems. Section 4 studies the strongly convex case with a new algorithm and its convergence guarantees. It also presents an application to constrained convex problems under semi-strongly convex assumption. Section 5 provides two illustrative numerical examples. For the sake of presentation, all technical proofs of the results in the main text are deferred to the appendices.

2. Basic Assumption and Optimality Conditions.

Basic notation and concepts. We work with Euclidean spaces $\mathbb{R}^p$ and $\mathbb{R}^n$ equipped with standard inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. For any nonempty, closed, and convex set $\mathcal{X}$ in $\mathbb{R}^p$, $\mathrm{ri}(\mathcal{X})$ denotes the relative interior of $\mathcal{X}$ and $\delta_{\mathcal{X}}(\cdot)$ denotes the indicator of $\mathcal{X}$. For any proper, closed, and convex function $f : \mathbb{R}^p \to \mathbb{R}\cup\{+\infty\}$, $\mathrm{dom}(f) := \{x \in \mathbb{R}^p \mid f(x) < +\infty\}$ is its (effective) domain, $f^*(y) := \sup_x\{\langle x, y\rangle - f(x)\}$ denotes

the Fenchel conjugate of f, $\partial f(x) := \{w \in \mathbb{R}^p \mid f(y)-f(x) \geq \langle w, y-x\rangle,\ \forall y \in \mathrm{dom}(f)\}$ stands for the subdifferential of f at x, and ∇f is the gradient or subgradient of f.

A function f is called M_f-Lipschitz continuous on $\mathrm{dom}(f)$ with a Lipschitz constant $M_f \in [0,+\infty)$ if $|f(x) - f(y)| \leq M_f\|x - y\|$ for all $x, y \in \mathrm{dom}(f)$. If f is differentiable on $\mathrm{dom}(f)$ and ∇f is Lipschitz continuous with a Lipschitz constant $L_f \in [0,+\infty)$, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L_f\|x - y\|$ for $x, y \in \mathrm{dom}(f)$, then we say that f is L_f-smooth. If $f(\cdot) - \frac{\mu_f}{2}\|\cdot\|^2$ is still convex for some $\mu_f > 0$, then we say that f is μ_f-strongly convex with a strong convexity parameter μ_f. We also denote $\mathrm{prox}_f(x) := \arg\min_y\{f(y) + \frac{1}{2}\|y - x\|^2\}$ the proximal operator of f. For any $\rho > 0$, we have the following Moreau's identity [2]:

$$\mathrm{prox}_{f/\rho}(x) + \rho^{-1}\mathrm{prox}_{\rho f^*}(\rho x) = x. \tag{4}$$

We use $\mathcal{O}(\cdot)$, $o(\cdot)$, and $\Omega(\cdot)$ to denote the order of complexity bounds as usual. With convergence rates, for two scalar sequences $\{u_k\} \subseteq \mathbb{R}_+$ and $\{v_k\} \subseteq \mathbb{R}_{++}$, we say that

$$u_k = \mathcal{O}(v_k) \quad \text{if} \quad \limsup_{k\to\infty}\left(\frac{u_k}{v_k}\right) < +\infty \qquad \text{and} \qquad u_k = o(v_k) \quad \text{if} \quad \lim_{k\to\infty}\left(\frac{u_k}{v_k}\right) = 0.$$

In this paper, we further define a new $\underline{o}(\cdot)$ notation for convergence rates as follows:

$$u_k = \underline{o}(v_k) \quad \text{if} \quad \liminf_{k\to\infty}\left(\frac{u_k}{v_k}\right) = 0. \tag{5}$$

That is, there is a subsequence $\{k_j\} \subseteq \mathbb{N}$ such that $u_{k_j} = o\big(v_{k_j}\big)$.

Our algorithms rely on the following assumption imposed on the problem (1):

Assumption 2.1. The functions f and g in (1) are proper, closed, and convex. The solution set $\mathcal{X}^\star$ of (1) is nonempty, and $0 \in \mathrm{ri}\,(\mathrm{dom}(g) - K\mathrm{dom}(f))$.

Assumption 2.1 is fundamental and required in any primal-dual method for theoretical convergence guarantees. Since $\mathcal{X}^\star$ is nonempty, under Assumption 2.1, the strong duality holds, thus we have $F^\star + G^\star = 0$, and the solution set $\mathcal{Y}^\star$ of the dual problem (2) is also nonempty.

Optimality conditions. The optimality conditions of (1) and (2) are

$$\text{primal: } 0 \in \partial f(x^\star) + K^\top\partial g(Kx^\star) \quad \text{ or dual: } 0 \in -K\partial f^*(-K^\top y^\star) + \partial g^*(y^\star).$$

These two conditions can be written into the following primal-dual optimality condition, which can also be derived from the min-max form (3):

$$\text{primal-dual: } 0 \in K^\top y^\star + \partial f(x^\star) \quad \text{and} \quad 0 \in -Kx^\star + \partial g^*(y^\star).$$

Gap function. Let $\widetilde{\mathcal{L}}(x, y) := f(x) + \langle Kx, y\rangle - g^*(y)$ be defined by (3) and $\mathcal{X}$ and $\mathcal{Y}$ be given nonempty, closed, and convex sets such that $\mathcal{X}^\star \cap \mathcal{X} \neq \emptyset$ and $\mathcal{Y}^\star \cap \mathcal{Y} \neq \emptyset$. We define a gap function $\mathcal{G}_{\mathcal{X}\times\mathcal{Y}}(\cdot)$ as follows:

$$\mathcal{G}_{\mathcal{X}\times\mathcal{Y}}(x, y) := \sup_{(\hat{x},\hat{y})\in\mathcal{X}\times\mathcal{Y}}\left\{\widetilde{\mathcal{L}}(x, \hat{y}) - \widetilde{\mathcal{L}}(\hat{x}, y)\right\} = \sup_{\hat{y}\in\mathcal{Y}}\widetilde{\mathcal{L}}(x, \hat{y}) - \inf_{\hat{x}\in\mathcal{X}}\widetilde{\mathcal{L}}(\hat{x}, y). \tag{6}$$

Then, we immediately have

$$\mathcal{G}_{\mathcal{X}\times\mathcal{Y}}(x, y) \geq \widetilde{\mathcal{L}}(x, y^\star) - \widetilde{\mathcal{L}}(x^\star, y) \geq \widetilde{\mathcal{L}}(x^\star, y^\star) - \widetilde{\mathcal{L}}(x^\star, y^\star) = 0,$$

where $(x^\star, y^\star) \in \mathcal{X}^\star \times \mathcal{Y}^\star$ is a primal-dual solution of (1) and (2), i.e., a saddle-point of $\widetilde{\mathcal{L}}$. Moreover, $\mathcal{G}_{\mathcal{X}\times\mathcal{Y}}$ vanishes at any saddle point $(x^\star, y^\star)$. Thus this gap function can be considered as a measure of optimality for both (1) and (2).

Constrained reformulation and merit function. The primal problem (1) can be reformulated into the following equivalent constrained setting:

$$F^\star := \min_{x\in\mathbb{R}^p,\ r\in\mathbb{R}^n}\Big\{F(x,r) := f(x) + g(r) \ \text{ s.t. } \ Kx - r = 0\Big\}. \tag{7}$$

Let $\mathcal{L}(x,r,y) := f(x) + g(r) + \langle Kx - r, y\rangle$ be the Lagrange function associated with (7), where $y \in \mathbb{R}^n$ is the corresponding Lagrange multiplier, and $\widetilde{\mathcal{L}}(x,y) := f(x) + \langle Kx, y\rangle - g^*(y)$ be defined by (3). Since $g^*(y) := \sup_{r\in\mathbb{R}^n}\{\langle y, r\rangle - g(r)\}$, we can show that, for any $r \in \mathbb{R}^n$, one has

$$\widetilde{\mathcal{L}}(x,y) \le f(x) + g(r) + \langle Kx - r, y\rangle = \mathcal{L}(x,r,y). \tag{8}$$

Moreover, $\widetilde{\mathcal{L}}(x,y) = \mathcal{L}(x,r,y)$ if and only if $y \in \partial g(r)$ or equivalently $r \in \partial g^*(y)$.

Together with $\mathcal{L}$, we define an augmented Lagrangian $\mathcal{L}_\rho$ as

$$\mathcal{L}_\rho(x,r,y) := \mathcal{L}(x,r,y) + \frac{\rho}{2}\|Kx - r\|^2 = f(x) + g(r) + \phi_\rho(x,r,y), \tag{9}$$

where $\phi_\rho(x,r,y) := \langle Kx - r, y\rangle + \frac{\rho}{2}\|Kx - r\|^2$ and $\rho > 0$ is a penalty parameter. Note that the term $\frac{\rho}{2}\|Kx - r\|^2$ can be viewed as a smoothed approximation of $\delta_{\{(x,r)\mid Kx-r=0\}}(x,r)$, the indicator of $\{(x,r) \mid Kx - r = 0\}$. The function $\mathcal{L}_\rho$ will serve as a ***merit function*** to develop our algorithms in the sequel.

3. A New Primal-Dual Algorithm for General Convex Case. In this section, we develop a novel primal-dual algorithm to solve (1) and its dual form (2) with fast convergence rate guarantees, where f and g are both ***merely convex***.

3.1. Algorithm derivation and one-iteration analysis. Our main idea is to combine four techniques in one: alternating minimization, linearization, acceleration, and homotopy. While each individual technique is classical, their entire combination is ***new***. At the iteration $k \ge 0$, given x^k, $\tilde{x}^k$, r^k, and $\tilde{y}^k$, we update

$$\left\{\begin{array}{lcl}
\hat{x}^k &:=& (1-\tau_k)x^k + \tau_k\tilde{x}^k,\\
r^{k+1} &:=& \mathrm{prox}_{g/\rho_k}\big(\tilde{y}^k/\rho_k + K\hat{x}^k\big),\\
x^{k+1} &:=& \mathrm{prox}_{\beta_k f}\big(\hat{x}^k - \beta_k\nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k)\big),\\
\tilde{x}^{k+1} &:=& \tilde{x}^k + \frac{1}{\tau_k}(x^{k+1} - \hat{x}^k),\\
\tilde{y}^{k+1} &:=& \tilde{y}^k + \eta_k\left[Kx^{k+1} - r^{k+1} - (1-\tau_k)(Kx^k - r^k)\right].
\end{array}\right. \tag{10}$$

We now explain each step of the scheme (10) as follows:

- Line 2 and line 3 of (10) alternatively minimize the merit function $\mathcal{L}_\rho$ w.r.t. r and x to obtain r^{k+1} and x^{k+1}, respectively. However, since the subproblem in x^{k+1} is difficult to solve, we linearize the coupling term $\frac{\rho}{2}\|Kx - r\|^2$ as

$$\begin{array}{lcl}
\frac{\rho_k}{2}\|Kx - r^{k+1}\|^2 &\approx& \frac{\rho_k}{2}\|K\hat{x}^k - r^{k+1}\|^2\\
&&+\ \frac{\rho_k}{2}\langle\nabla_x\|K\hat{x}^k - r^{k+1}\|^2, x - \hat{x}^k\rangle + \frac{1}{2\beta_k}\|x - \hat{x}^k\|^2,
\end{array}$$

  so that we can simply use the proximal operator of f as in line 3.
- Line 1 and line 4 update $\hat{x}^k$ and $\tilde{x}^{k+1}$, respectively to accelerate the primal iterates using Nesterov's acceleration strategy [45].
- Line 5 updates the dual variable $\tilde{y}^{k+1}$ as in augmented Lagrangian methods.

All the parameters $\tau_k \in (0,1]$, $\rho_k > 0$, $\beta_k > 0$, and $\eta_k > 0$ will be updated in a homotopy fashion. We will explicitly provide update rules for these parameters in Algorithm 1 based on our convergence analysis.

The following lemma provides a key estimate on the difference $\mathcal{L}_{\rho_{k-1}}(x^k, r^k, y) - \mathcal{L}(x, r, \bar{y}^k)$ to prove Theorems 2 and 5. The proof is deferred to Appendix B.1.

LEMMA 1. *Let $(x^k, \hat{x}^k, \tilde{x}^k, r^k, \tilde{y}^k)$ be updated by (10) with $\rho_k > \eta_k$ and $\bar{y}^{k+1} := (1-\tau_k)\bar{y}^k + \tau_k\left[\tilde{y}^k + \rho_k(K\hat{x}^k - r^{k+1})\right]$. Then, for any $(x, r, y) \in \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^n$, the following inequality holds:*

$$\begin{aligned}\left[\mathcal{L}_{\rho_k}(x^{k+1}, r^{k+1}, y) - \mathcal{L}(x, r, \bar{y}^{k+1})\right] &\le (1-\tau_k)\left[\mathcal{L}_{\rho_{k-1}}(x^k, r^k, y) - \mathcal{L}(x, r, \bar{y}^k)\right] \\ &+ \tfrac{\tau_k^2}{2\beta_k}\left[\|\tilde{x}^k - x\|^2 - \|\tilde{x}^{k+1} - x\|^2\right] + \tfrac{1}{2\eta_k}\left[\|\tilde{y}^k - y\|^2 - \|\tilde{y}^{k+1} - y\|^2\right] \\ &- \tfrac{1}{2}\left(\tfrac{1}{\beta_k} - \tfrac{\rho_k^2\|K\|^2}{\rho_k - \eta_k}\right)\|x^{k+1} - \hat{x}^k\|^2 - \tfrac{(1-\tau_k)}{2}\left[\rho_{k-1} - (1-\tau_k)\rho_k\right]\|Kx^k - r^k\|^2.\end{aligned} \tag{11}$$

3.2. The complete algorithm. To transform our scheme (10) into a primal-dual format, we first eliminate r^k and r^{k+1}. By Moreau's identity (4), we have

$$r^{k+1} = \tfrac{1}{\rho_k}\big(\tilde{y}^k + \rho_k K\hat{x}^k - y^{k+1}\big), \quad \text{where} \quad y^{k+1} := \operatorname{prox}_{\rho_k g^*}\big(\tilde{y}^k + \rho_k K\hat{x}^k\big). \tag{12}$$

Now, from the definition of ϕ_ρ in (9), we can write

$$\nabla_x \phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k) = K^\top\big(\tilde{y}^k + \rho_k K\hat{x}^k - \rho_k r^{k+1}\big) = K^\top y^{k+1}.$$

Substituting this expression into line 3 of (10), we can eliminate r^{k+1}. Next, we combine line 1 and line 4 of (10) to obtain $\hat{x}^{k+1} = x^{k+1} + \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}(x^{k+1} - x^k)$. Finally, substituting r^k using (12) into line 5 of (10), we can express $\tilde{y}^{k+1}$ as

$$\begin{aligned}\tilde{y}^{k+1} &= \tilde{y}^k + \eta_k K\big(x^{k+1} - \hat{x}^k - (1-\tau_k)(x^k - \hat{x}^{k-1})\big) \\ &\quad - \tfrac{\eta_k}{\rho_k}(\tilde{y}^k - y^{k+1}) + \tfrac{\eta_k(1-\tau_k)}{\rho_{k-1}}(\tilde{y}^{k-1} - y^k).\end{aligned} \tag{13}$$

In addition to (10), we also update y using the following weighted averaging step:

$$\bar{y}^{k+1} := (1-\tau_k)\bar{y}^k + \tau_k y^{k+1}, \tag{14}$$

where y^{k+1} is defined in (12). This is consistent with the condition in Lemma 1.

For the parameters, as guided by Lemma 1, we propose the following update:

$$\tau_k := \frac{c}{k+c}, \qquad \rho_k := \frac{\rho_0}{\tau_k}, \qquad \beta_k := \frac{\gamma}{\|K\|^2\rho_k} \quad \text{and} \quad \eta_k := (1-\gamma)\rho_k, \tag{15}$$

where $c \ge 1$, $\gamma \in (0,1)$, and $\rho_0 > 0$ are given.

In summary, we describe the complete primal-dual algorithm as in Algorithm 1. Let us highlight the following features of Algorithm 1.

- Algorithm 1 updates its parameters at Step 4 dynamically. The update of τ_k is often seen in Nesterov's accelerated-based schemes. The penalty parameter ρ_k is not fixed, but is updated in a homotopy fashion and also different from the dual step-size η_k. The dual update at Step 6 is completely new and depends on three consecutive iterations. All these properties are fundamentally different from existing primal-dual and augmented Lagrangian-based methods.
- We use two parameters $\gamma \in (0,1)$ and $\rho_0 > 0$ to balance the primal term $\|x^0 - x^\star\|^2$ and dual term $\|y^0 - y^\star\|^2$ in the bound (16) of Theorem 2 below. Note that our update leads to $\rho_k\beta_k\|K\|^2 = \gamma < 1$, which is the same as the parameter condition in the Chambolle-Pock primal-dual method [10].
- The per-iteration complexity of Algorithm 1 is essentially the same as in existing primal-dual methods. It requires one $\operatorname{prox}_{\rho_k g^*}$, one $\operatorname{prox}_{\beta_k f}$, one Kx, and one $K^\top y$. The matrix-vector multiplication at Step 6 can be eliminated if we store Kx^k and Kx^{k+1}, and use the last line of Step 5 to compute $K\hat{x}^k$.

Algorithm 1 (New Primal-Dual Algorithm for (1) and (2): General Convex Case)

1: **Initialization:** Choose $x^0 \in \mathbb{R}^p$, $y^0 \in \mathbb{R}^n$, $\rho_0 > 0$, $c \geq 1$, and $\gamma \in (0,1)$.
2: Set $\tau_0 := 1$, $x^{-1} := \hat{x}^0 := x^0$, and $\tilde{y}^{-1} := \tilde{y}^0 := \bar{y}^0 := y^0$.
3: **For** $k := 0, 1, \cdots, k_{\max}$ **do**
4: Update $\rho_k := \frac{\rho_0}{\tau_k}$, $\beta_k := \frac{\gamma}{\|K\|^2\rho_k}$, $\eta_k := (1-\gamma)\rho_k$, and $\tau_{k+1} := \frac{c}{k+c+1}$.
5: Update the primal-dual step:

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

6: Update the intermediate dual step:

$$\begin{aligned} \tilde{y}^{k+1} &:= \tilde{y}^k + \eta_k K\big[x^{k+1} - \hat{x}^k - (1-\tau_k)(x^k - \hat{x}^{k-1})\big] \\ &\quad + (1-\gamma)\left[y^{k+1} - \tilde{y}^k - \tfrac{\tau_{k-1}(1-\tau_k)}{\tau_k}(y^k - \tilde{y}^{k-1})\right]. \end{aligned}$$

7: Update the dual averaging step: $\bar{y}^{k+1} := (1-\tau_k)\bar{y}^k + \tau_k y^{k+1}$.
8: **EndFor**

3.3. Convergence analysis. The following theorem states convergence guarantees of Algorithm 1 under Assumption 2.1 with $c = 1$ without any smoothness or strong convexity assumption. Its proof is given in Appendix B.2.

THEOREM 2 ($\mathcal{O}(1/k)$ convergence rates when $c = 1$). *Let $\{(x^k, \bar{y}^k)\}$ be generated by Algorithm 1 with $c := 1$, and $\widetilde{\mathcal{L}}$ be defined by (3). Then, under Assumption 2.1, the following bound is valid for any given $x, x^0 \in \mathbb{R}^p$ and $y, y^0 \in \mathbb{R}^n$:*

$$\widetilde{\mathcal{L}}(x^k, y) - \widetilde{\mathcal{L}}(x, \bar{y}^k) \leq \frac{1}{2k}\left[\frac{\rho_0\|K\|^2\|x^0 - x\|^2}{\gamma} + \frac{\|y^0 - y\|^2}{(1-\gamma)\rho_0}\right]. \tag{16}$$

Furthermore, the following statements hold:

(a) (Semi-ergodic convergence on the primal-dual gap). *The gap function $\mathcal{G}_{\mathcal{X}\times\mathcal{Y}}$ defined by (6) satisfies the following bound for all $k \geq 1$:*

$$0 \leq \mathcal{G}_{\mathcal{X}\times\mathcal{Y}}(x^k, \bar{y}^k) \leq \frac{1}{2k}\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\left[\frac{\rho_0\|K\|^2\|x^0 - x\|^2}{\gamma} + \frac{\|y^0 - y\|^2}{(1-\gamma)\rho_0}\right]. \tag{17}$$

Hence, the primal-dual gap sequence $\big\{\mathcal{G}_{\mathcal{X}\times\mathcal{Y}}(x^k, \bar{y}^k)\big\}$ converges to zero at $\mathcal{O}(1/k)$ rate in semi-ergodic sense, i.e., non-ergodic in x^k and ergodic in $\bar{y}^k$.

(b) (Non-ergodic convergence on the primal objective). *If g is M_g-Lipschitz continuous on $\mathrm{dom}(g)$ and $x^\star$ is an optimal solution of (1), then, for $k \geq 1$, the primal objective residual based on the last-iterate sequence $\{x^k\}$ satisfies:*

$$0 \leq F(x^k) - F^\star \leq \frac{1}{2k}\left[\frac{\rho_0\|K\|^2\|x^0 - x^\star\|^2}{\gamma} + \frac{D_g^2}{(1-\gamma)\rho_0}\right], \tag{18}$$

where $D_g := \sup\big\{\|y^0 - y\| \mid \|y\| \leq M_g\big\}$. Hence, $\{F(x^k)\}$ converges to the primal optimal value $F^\star$ of (1) at $\mathcal{O}(1/k)$ rate in non-ergodic sense.

(c) (Ergodic convergence on the dual objective). *If f^* is M_{f^*}-Lipschitz continuous on $\mathrm{dom}(f^*)$ and $y^\star$ is an optimal solution of (2), then, for all $k \geq 1$, the dual objective residual based on the averaging sequence $\{\bar{y}^k\}$ satisfies:*

$$0 \le G(\bar{y}^k) - G^\star \le \frac{1}{2k}\left[\frac{\rho_0\|K\|^2 D_{f^*}^2}{\gamma} + \frac{\|y^0 - y^\star\|^2}{(1-\gamma)\rho_0}\right], \tag{19}$$

where $D_{f^*} := \sup\{\|x^0 - x\| \mid \|x\| \le M_{f^*}\}$. *Hence,* $\{G(\bar{y}^k)\}$ *converges to the dual optimal value* $G^\star$ *of* (2) *at* $\mathcal{O}(1/k)$ *rate in ergodic sense.*

Remark 3 (*Optimal rate*). It was shown in [34, 62] that, under Assumption 2.1, the rate $\mathcal{O}(1/k)$ is **optimal**, in the sense that for any algorithm $\mathcal{A}$ for solving (1), in order to achieve the bound $F(x^k) - F^\star \le \varepsilon$, there exists an instance of f and g with their arguments' dimensions p and n dependent on ε, such that $\mathcal{A}$ makes $\Omega(1/\varepsilon)$ queries to the first-order oracle of f and g (e.g., $f(x)$, $\nabla f(x)$, or $\mathrm{prox}_{\rho f}(x)$). In other words, the convergence rate of $\mathcal{A}$ cannot exceed $\mathcal{O}(1/k)$ rate under Assumption 2.1 when the problem dimension p is much larger than the number of iterations k, i.e., $k \le \mathcal{O}(p)$. Consequently, Algorithm 1 indeed achieves **optimal** convergence rate.

Remark 4 (*Symmetry*). Since the primal-dual problems (1) and (2) are symmetric, to obtain a non-ergodic convergence rate on the dual problem (2), we could apply Algorithm 1 to the dual-primal pair instead of the primal-dual pair.

If we choose $c > 1$, then Algorithm 1 still converges. In fact, it achieves the same $\mathcal{O}(1/k)$ and a potentially faster[1] $\underline{o}\big(1/(k\sqrt{\log k})\big)$ convergence rate on the primal objective residual, as shown in Theorem 5, whose proof is given in Appendix B.3.

THEOREM 5 ($\mathcal{O}(1/k)$ and $\underline{o}\big(1/(k\sqrt{\log k})\big)$ convergence rates when $c > 1$). *Let* $\{x^k\}$ *be generated by Algorithm 1 with* $c > 1$. *Let* $\widetilde{\mathcal{L}}$ *be defined by* (3) *and* $y^\star$ *be an optimal solution of* (2). *Then, under Assumption 2.1, for any* $k \ge 0$, *we have*

$$0 \le \widetilde{\mathcal{L}}(x^k, y^\star) - F^\star \le \frac{R_0^2}{k+c-1} \quad \text{and} \quad \liminf_{k\to\infty} k\log(k)\big[\widetilde{\mathcal{L}}(x^k, y^\star) - F^\star\big] = 0, \tag{20}$$

where $R_0^2 := (c-1)\big[F(x^0) - F^\star\big] + \frac{c}{2}\left[\frac{\rho_0\|K\|^2}{\gamma}\|x^0 - x^\star\|^2 + \frac{1}{(1-\gamma)\rho_0}\|y^0 - y^\star\|^2\right]$.

Moreover, if g *is* M_g*-Lipschitz continuous on* $\mathrm{dom}(g)$, *then the primal last-iterate sequence* $\{x^k\}$ *satisfies the following statements for all* $k \ge 0$*:*

$$0 \le F(x^k) - F^\star \le \frac{R_1^2}{k+c-1} \quad \text{and} \quad \liminf_{k\to\infty} k\sqrt{\log k}\big[F(x^k) - F^\star\big] = 0, \tag{21}$$

where $R_1^2 := R_0^2 + \sqrt{2c/\rho_0}(\|y^\star\| + M_g)R_0$. *Hence,* $\{F(x^k)\}$ *converges to the primal optimal value* $F^\star$ *of* (1) *at both* $\mathcal{O}(1/k)$ *and* $\underline{o}\big(1/(k\sqrt{\log k})\big)$ *convergence rates in non-ergodic sense, where* $\underline{o}(\cdot)$ *is defined in* (5).

Remark 6. The $\underline{o}\big(1/(k\sqrt{\log k})\big)$ rate does not contradict our discussion in Remark 3, since our problem dimensions p and n are fixed, while k can be sufficiently large.

3.4. Application to constrained problems. Let us apply Algorithm 1 to solve the following nonsmooth constrained convex optimization problem:

$$F^\star := \min_{x\in\mathbb{R}^p}\Big\{F(x) := f(x) + \psi(x) \quad \text{s.t.} \quad Kx = b\Big\}, \tag{22}$$

where f and K are defined as in (1), ψ is proper, closed, and convex, and $b \in \mathbb{R}^n$. This problem is a special case of (1) where f is replaced by $f + \psi$, and $g(u) := \delta_{\{b\}}(u)$, the indicator of $\{b\}$. In this case, $g^*(y) = \langle b, y\rangle$, and the last condition of Assumption 2.1 reduces to the Slater condition: $\mathrm{ri}\,(\mathrm{dom}(f) \cap \mathrm{dom}(\psi)) \cap \{x \mid Kx = b\} \ne \emptyset$. In addition, we require the following assumption on the new objective term ψ:

[1] In fact, our numerical experiments in Section 5 show significantly faster empirical convergence rates of Algorithm 1 when we use the parameter update rules (15) with $c > 1$.

Assumption 3.1. The function ψ in (22) is convex and L_ψ-smooth.

We specify Algorithm 1 to solve (22) and its dual problem as follows:

$$\left\{\begin{array}{lcl} y^{k+1} &:=& \tilde{y}^k + \rho_k(K\hat{x}^k - b),\\ x^{k+1} &:=& \mathrm{prox}_{\beta_k f}\left(\hat{x}^k - \beta_k\left[K^\top y^{k+1} + \nabla\psi(\hat{x}^k)\right]\right),\\ \hat{x}^{k+1} &:=& x^{k+1} + \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}(x^{k+1} - x^k),\\ \tilde{y}^{k+1} &:=& \tilde{y}^k + \eta_k\left[K\left(x^{k+1} - (1-\tau_k)x^k\right) - \tau_k b\right],\\ \bar{y}^{k+1} &:=& (1-\tau_k)\bar{y}^k + \tau_k y^{k+1}, \end{array}\right. \tag{23}$$

where all the parameters are updated as in Algorithm 1 with a small modification $\beta_k := \gamma/(\|K\|^2\rho_k + \gamma L_\psi)$. Its convergence guarantee is summarized in the following corollary, whose proof is given in Appendix B.4.

COROLLARY 7. *Let $\left\{(x^k, \bar{y}^k)\right\}$ be generated by scheme (23) to solve (22) and its dual problem under Assumptions 2.1 and 3.1. Let $(x^\star, y^\star)$ be a pair of primal-dual optimal solution of (22). If we choose $c := 1$, then, for all $k \geq 1$, we have the following primal convergence rate guarantee:*

$$|F(x^k) - F^\star| \leq \frac{R_0^2}{2k} \quad \text{and} \quad \|Kx^k - b\| \leq \frac{R_0^2}{2k}, \tag{24}$$

where $R_0^2 := \frac{\rho_0\|K\|^2 + \gamma L_\psi}{\gamma}\|x^0 - x^\star\|^2 + \frac{1}{(1-\gamma)\rho_0}(2\|y^\star\| + \|y^0\| + 1)^2$. Hence, the objective residual and the feasibility violation both converge to zero at $\mathcal{O}(1/k)$ non-ergodic rate.

If, in addition, $\mathrm{dom}(F)$ is bounded, then we have the dual convergence guarantee:

$$G(\bar{y}^k) - G^\star \leq \frac{1}{2k}\left[\frac{(\rho_0\|K\|^2 + \gamma L_\psi)\mathcal{D}_F^2}{\gamma} + \frac{\|y^0 - y^\star\|^2}{(1-\gamma)\rho_0}\right], \tag{25}$$

where $\mathcal{D}_F := \sup\left\{\|x - x^0\| \mid x \in \mathrm{dom}(F)\right\} < +\infty$.

If we choose $c > 1$ in the variant of Algorithm 1 for solving (22), then the $\mathcal{O}(1/k)$ non-ergodic rate bounds on $|F(x^k) - F^\star|$ and $\|Kx^k - b\|$ still hold, and

$$\liminf_{k\to\infty} k\sqrt{\log k}\left[|F(x^k) - F^\star| + \|Kx^k - b\|\right] = 0. \tag{26}$$

Hence, the objective residual and feasibility violation sequences both converge to zero at both $\mathcal{O}(1/k)$ and $o\big(1/(k\sqrt{\log k})\big)$ non-ergodic convergence rates.

Remark 8. The $\mathcal{O}(1/k)$ rate results of Corollary 7 are similar to [54, 57]. However, [54] studied only primal methods for constrained convex problems using quadratic penalty framework and alternating minimization techniques without updating dual variables, and thus does not have convergence guarantee on the dual problem. The other work [57] relies on a different approach called smoothing techniques and excessive gap framework introduced in [46].

Remark 9. We can extend the scheme (23) to solve (22) with general linear constraint $Kx - b \in \mathcal{S}$ instead of $Kx - b = 0$, where $\mathcal{S}$ is a nonempty, closed, and convex set in $\mathbb{R}^n$. This constraint covers linear inequality constraints as special cases. One simple trick is to introduce a slack variable s and reformulating this constraint into $Kx - s = b$ and $s \in \mathcal{S}$. Next, we replace the objective function $F(x) = f(x) + \psi(x)$ in (22) by $F(x, s) := f(x) + \psi(x) + \delta_{\mathcal{S}}(s)$, where $\delta_{\mathcal{S}}(s)$ is the indicator of $\mathcal{S}$. Then, we can apply (23) to solve the resulting problem in x and s. In this case, our new

scheme requires projection onto $\mathcal{S}$. As another option, we can adopt the approach in [54] to handle $Kx - b \in \mathcal{S}$ directly without reformulation. We omit this extension to avoid overloading the paper.

4. A New Primal-Dual Method for Strongly Convex Case. In this section, we consider a special case of problem (1), where f is strongly convex. More precisely, we impose the following assumption.

Assumption 4.1. The function f in (1) is strongly convex with a strong convexity parameter $\mu_f > 0$, but not necessarily smooth.

4.1. Algorithm derivation and one-iteration analysis. We follow the same diagram as in Section 3, but replacing Nesterov's accelerated step [45] by Tseng's scheme [59], which allows us to achieve a $\mathcal{O}\left(1/k^2\right)$ non-ergodic convergence rate. With this modification, we now describe our primal-dual scheme for solving (1)-(2):

$$\left\{\begin{array}{lcl} \hat{x}^k &:=& (1-\tau_k)x^k + \tau_k\tilde{x}^k,\\ r^{k+1} &:=& \mathrm{prox}_{g/\rho_k}(\tilde{y}^k/\rho_k + K\hat{x}^k),\\ \tilde{x}^{k+1} &:=& \mathrm{prox}_{(\beta_k/\tau_k)f}\big(\tilde{x}^k - \frac{\beta_k}{\tau_k}\nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k)\big),\\ x^{k+1} &:=& \mathrm{prox}_{f/(\rho_k\|K\|^2)}\big(\hat{x}^k - \frac{1}{\rho_k\|K\|^2}\nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k)\big),\\ \tilde{y}^{k+1} &:=& \tilde{y}^k + \eta_k\big[Kx^{k+1} - r^{k+1} - (1-\tau_k)(Kx^k - r^k)\big]. \end{array}\right. \tag{27}$$

The parameters τ_k, ρ_k, β_k, and η_k will be specified later based on our analysis.

We first analyze one iteration of the primal-dual scheme (27) in the following lemma to obtain a recursive estimate. Its proof can be found in Appendix C.1.

LEMMA 10. *Let $(x^k, \hat{x}^k, \tilde{x}^k, r^k, \tilde{y}^k)$ be generated by (27), and $\bar{y}^{k+1} := (1-\tau_k)\bar{y}^k + \tau_k\left[\tilde{y}^k + \rho_k(K\hat{x}^k - r^{k+1})\right]$. Assume that $\rho_k > \eta_k$ and $\rho_k\beta_k\|K\|^2 < 1$. Then, for any $(x, r, y) \in \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^n$, it holds that*

$$\begin{aligned} \big[\mathcal{L}_{\rho_k}(x^{k+1}, r^{k+1}, y) - \mathcal{L}(x, r, \bar{y}^{k+1})\big] &\le (1-\tau_k)\big[\mathcal{L}_{\rho_{k-1}}(x^k, r^k, y) - \mathcal{L}(x, r, \bar{y}^k)\big]\\ &+ \tfrac{\tau_k^2}{2\beta_k}\|\tilde{x}^k - x\|^2 - \tfrac{\tau_k(\tau_k+\beta_k\mu_f)}{2\beta_k}\|\tilde{x}^{k+1} - x\|^2 + \tfrac{1}{2\eta_k}\left[\|\tilde{y}^k - y\|^2 - \|\tilde{y}^{k+1} - y\|^2\right]\\ &- \tfrac{\rho_k}{2}\Big(1 - \rho_k\beta_k\|K\|^2 - \tfrac{\eta_k}{\rho_k - \eta_k}\Big)\|K(x^{k+1} - \hat{x}^k)\|^2\\ &- \tfrac{(1-\tau_k)}{2}\left[\rho_{k-1} - (1-\tau_k)\rho_k\right]\|Kx^k - r^k\|^2. \end{aligned} \tag{28}$$

4.2. Parameter update and complete algorithm. As before, we can eliminate r^k and r^{k+1} in (27) following the same lines as in Subsection 3.2, in order to transform (27) into a primal-dual form. We also add an averaging sequence $\{\bar{y}^k\}$ as in Lemma 10 to obtain a dual convergence rate. Furthermore, as guided by Lemma 10, we propose the following parameter update rules:

$$\rho_k := \frac{\rho_0}{\tau_k^2}, \qquad \beta_k := \frac{\Gamma}{\rho_k\|K\|^2}, \quad \text{and} \quad \eta_k := (1-\gamma)\rho_k, \tag{29}$$

where $\gamma \in (\frac{1}{2}, 1)$ and $\Gamma := 2 - \frac{1}{\gamma} \in (0, 1)$ are given. For the choice of ρ_0 and the update of τ_k, we provide two cases:

$$\textbf{Case 1:} \quad \rho_0 \in \left(0, \ \frac{\Gamma\mu_f}{2\|K\|^2}\right] \text{ and } \tau_{k+1} := \frac{\tau_k}{2}\left(\sqrt{\tau_k^2 + 4} - \tau_k\right), \text{ where } \tau_0 := 1, \tag{30}$$

or

$$\text{(31)}\quad \textbf{Case 2:}\ \ \rho_0 \in \left(0,\ \frac{c(c-1)\Gamma\mu_f}{(2c-1)\|K\|^2}\right] \quad\text{and}\quad \tau_k := \frac{c}{k+c}, \quad \text{where } c > 2 \text{ is given.}$$

Now, we can describe our second first-order primal-dual algorithm as in Algorithm 2, and highlight the following features.

Algorithm 2 (New Primal-Dual Algorithm for (1) and (2): Strongly Convex Case)

1: **Initialization:** Choose $y^0 \in \mathbb{R}^n$, $x^0 \in \mathbb{R}^p$, and $\gamma \in (\frac{1}{2}, 1)$. Set $\Gamma := 2 - \frac{1}{\gamma}$.
2: Set ρ_0 and τ_0 according to (30) or (31).
3: Set $x^{-1} := \hat{x}^0 := x^0$ and $\tilde{y}^{-1} := \tilde{y}^0 := \bar{y}^0 := y^0$.
4: **For** $k := 0, 1, \cdots, k_{\max}$ **do**
5: Update $\rho_k := \frac{\rho_0}{\tau_k^2}$, $\beta_k := \frac{\Gamma}{\rho_k\|K\|^2}$, and $\eta_k := (1-\gamma)\rho_k$.
6: Update τ_{k+1} according to (30) or (31), consistent with the update in Step 2.
7: Update the primal-dual step:

$$\begin{cases} y^{k+1} &:= \ \mathrm{prox}_{\rho_k g^*}(\tilde{y}^k + \rho_k K\hat{x}^k), \\ \tilde{x}^{k+1} &:= \ \mathrm{prox}_{(\beta_k/\tau_k)f}(\tilde{x}^k - \frac{\beta_k}{\tau_k}K^\top y^{k+1}), \\ x^{k+1} &:= \ \mathrm{prox}_{f/(\rho_k\|K\|^2)}\left(\hat{x}^k - \frac{1}{\rho_k\|K\|^2}K^\top y^{k+1}\right), \\ \hat{x}^{k+1} &:= \ (1-\tau_{k+1})x^{k+1} + \tau_{k+1}\tilde{x}^{k+1}. \end{cases}$$

8: Update the intermediate dual step:

$$\begin{aligned} \tilde{y}^{k+1} \ :=\ & \tilde{y}^k + \eta_k K\big[x^{k+1} - \hat{x}^k - (1-\tau_k)(x^k - \hat{x}^{k-1})\big] \\ & + (1-\gamma)\Big[y^{k+1} - \tilde{y}^k - \tfrac{\tau_{k-1}(1-\tau_k)}{\tau_k}(y^k - \tilde{y}^{k-1})\Big]. \end{aligned}$$

9: Update the dual averaging step: $\bar{y}^{k+1} := (1-\tau_k)\bar{y}^k + \tau_k y^{k+1}$.
10: **EndFor**

- Since we aim at obtaining non-ergodic convergence rate, Algorithm 2 requires one additional $\mathrm{prox}_{f/(\rho_k\|K\|^2)}(\cdot)$ compared to Algorithm 1. We could replace this proximal step by an averaging step: $x^{k+1} := (1-\tau_k)x^k + \tau_k\tilde{x}^{k+1}$, but the convergence rate would no longer be non-ergodic in $\{x^k\}$.
- At each iteration, Algorithm 2 requires one $\mathrm{prox}_{\rho_k g^*}$, one $\mathrm{prox}_{\beta_k f}$, one $\mathrm{prox}_{f/(\rho_k\|K\|^2)}$, one Kx, and one $K^\top y$, which incur one more proximal operation of f than in existing primal-dual methods. Again, the matrix-vector multiplication at Step 8 can be eliminated by storing vectors Kx^k and Kx^{k+1}.
- Due to the symmetry between (1) and (2), if g^* is μ_{g^*}-strongly convex with $\mu_{g^*} > 0$ (or equivalently, g is L_g-smooth with $L_g := 1/\mu_{g^*}$), then we can apply Algorithm 2 to the dual-primal pair instead of the primal-dual pair.

4.3. Convergence analysis. We state the convergence of Algorithm 2 under Case 1, i.e., (30), in the following theorem, whose proof is in Appendix C.2.

THEOREM 11 ($\mathcal{O}\left(1/k^2\right)$ convergence rates under Case 1). *Let $\{(x^k, \bar{y}^k)\}$ be generated by Algorithm 2 using the update (29) and (30), and $\widetilde{\mathcal{L}}$ be defined by (3). Then, under Assumptions 2.1 and 4.1, for any $x,\ x^0 \in \mathbb{R}^p$ and $y,\ y^0 \in \mathbb{R}^n$, we have*

$$\text{(32)}\qquad \widetilde{\mathcal{L}}(x^k, y) - \widetilde{\mathcal{L}}(x, \bar{y}^k) \le \frac{2}{(k+1)^2}\left[\frac{\rho_0\|K\|^2\|x^0 - x\|^2}{\Gamma} + \frac{\|y^0 - y\|^2}{(1-\gamma)\rho_0}\right].$$

Moreover, the following statements hold:

(a) (Semi-ergodic convergence on the primal-dual gap). *The gap function* $\mathcal{G}_{\mathcal{X}\times\mathcal{Y}}$ *defined by* (6) *satisfies the following bound for all* $k \geq 1$*:*

$$\mathcal{G}_{\mathcal{X}\times\mathcal{Y}}(x^k, \bar{y}^k) \leq \frac{2}{(k+1)^2} \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left[\frac{\rho_0\|K\|^2\|x^0 - x\|^2}{\Gamma} + \frac{\|y^0 - y\|^2}{(1-\gamma)\rho_0}\right]. \tag{33}$$

Hence, the primal-dual gap sequence $\{\mathcal{G}_{\mathcal{X}\times\mathcal{Y}}(x^k, \bar{y}^k)\}$ *converges to zero at* $\mathcal{O}\left(1/k^2\right)$ *semi-ergodic rate, i.e., non-ergodic in* x^k *and ergodic in* $\bar{y}^k$*.*

(b) (Non-ergodic convergence on the primal objective). *If* g *is* M_g*-Lipschitz continuous on* $\mathrm{dom}(g)$ *and* $x^\star$ *is an optimal solution of* (1)*, then the primal objective residual based on the last-iterate sequence* $\{x^k\}$ *satisfies:*

$$0 \leq F(x^k) - F^\star \leq \frac{2}{(k+1)^2}\left[\frac{\rho_0\|K\|^2\|x^0 - x^\star\|^2}{\Gamma} + \frac{D_g^2}{(1-\gamma)\rho_0}\right], \tag{34}$$

where $D_g := \sup\{\|y^0 - y\| \mid \|y\| \leq M_g\}$*. Hence,* $\{F(x^k)\}$ *converges to the primal optimal value* $F^\star$ *of* (1) *at* $\mathcal{O}\left(1/k^2\right)$ *rate in non-ergodic sense.*

(c) (Ergodic convergence on the dual objective). *If* f^* *is* M_{f^*}*-Lipschitz continuous on* $\mathrm{dom}(f^*)$ *and* $y^\star$ *is an optimal solution of* (2)*, then the dual objective residual based on the averaging sequence* $\{\bar{y}^k\}$ *satisfies:*

$$0 \leq G(\bar{y}^k) - G^\star \leq \frac{2}{(k+1)^2}\left[\frac{\rho_0\|K\|^2 D_{f^*}^2}{\Gamma} + \frac{\|y^0 - y^\star\|^2}{(1-\gamma)\rho_0}\right], \tag{35}$$

where $D_{f^*} := \sup\{\|x^0 - x\| \mid \|x\| \leq M_{f^*}\}$*. Hence,* $\{G(\bar{y}^k)\}$ *converges to the dual optimal value* $G^\star$ *of* (2) *at* $\mathcal{O}\left(1/k^2\right)$ *rate in ergodic sense.*

Remark 12 (*Optimal rate*). As shown in [62, Theorem 2], the $\mathcal{O}\left(1/k^2\right)$ convergence rate of Algorithm 2 is optimal in the sense of Remark 3. Moreover, by Assumption 4.1, we can show that $\{\|x^k - x^\star\|^2\}$ converges to zero at $\mathcal{O}\left(1/k^2\right)$ rate.

If we update the parameters using Case 2, i.e., (31), then Algorithm 2 achieves the same $\mathcal{O}\left(1/k^2\right)$ and an empirically faster $\underline{o}\big(1/(k^2\sqrt{\log k})\big)$ rate on the primal objective residual, as shown in the following theorem, whose proof is given in Appendix C.3.

THEOREM 13 ($\mathcal{O}\left(1/k^2\right)$ and $\underline{o}\big(1/(k^2\sqrt{\log k})\big)$ convergence rates under Case 2). *Let* $\{x^k\}$ *be generated by Algorithm 2 using the update* (29) *and* (31)*. Let* $\widetilde{\mathcal{L}}$ *be defined by* (3) *and* $y^\star$ *be an optimal solution of* (2)*. Then, under Assumptions 2.1 and 4.1, the following statements holds:*

$$0 \leq \widetilde{\mathcal{L}}(x^k, y^\star) - F^\star \leq \frac{R_0^2}{(k+c-1)^2} \quad \text{and} \quad \liminf_{k\to\infty} k^2\log(k)\big[\widetilde{\mathcal{L}}(x^k, y^\star) - F^\star\big] = 0, \tag{36}$$

where $R_0^2 := (c\text{–}1)\left[F(x^0) - F^\star\right] + \frac{c-1}{2}\left[\frac{(c-1)\rho_0\|K\|^2}{\Gamma} + c\mu_f\right]\|x^0 - x^\star\|^2 + \frac{c^2}{2(1-\gamma)\rho_0}\|y^0 - y^\star\|^2$.

Moreover, if g *is* M_g*-Lipschitz continuous on* $\mathrm{dom}(g)$*, then the primal last-iterate sequence* $\{x^k\}$ *satisfies*

$$0 \leq F(x^k) - F^\star \leq \frac{R_1^2}{(k+c-1)^2} \quad \text{and} \quad \liminf_{k\to\infty} k^2\sqrt{\log k}\big[F(x^k) - F^\star\big] = 0, \tag{37}$$

where $R_1^2 := R_0^2 + \sqrt{2c^2/\rho_0}(\|y^\star\| + M_g)R_0$*. Hence,* $\{F(x^k)\}$ *converges to the primal optimal value* $F^\star$ *of* (1) *at both* $\mathcal{O}\left(1/k^2\right)$ *and* $\underline{o}\big(1/(k^2\sqrt{\log k})\big)$ *convergence rates in non-ergodic sense.*

Remark 14. The $o\big(1/(k^2\sqrt{\log k})\big)$ convergence rate stated in Theorem 13 is attained for sufficiently large k. This does not conflict with the optimal upper bound stated in Remark 12, where the number of iterations $k \leq \mathcal{O}(p)$ with p being the dimension of the problem.

4.4. Application to constrained problems with semi-strongly convex objective. Consider the following constrained convex optimization problem:

$$F^\star := \min_{x\in\mathbb{R}^p,\ w\in\mathbb{R}^q}\Big\{F(x,w) := f(x) + \psi(w) \quad \text{s.t.} \quad Kx + Bw = b\Big\}, \tag{38}$$

where $f : \mathbb{R}^p \to \mathbb{R}\cup\{+\infty\}$ and $\psi : \mathbb{R}^q \to \mathbb{R}\cup\{+\infty\}$ are proper, closed, and convex, $K \in \mathbb{R}^{n\times p}$, $B \in \mathbb{R}^{n\times q}$, and $b \in \mathbb{R}^n$. Different from (22), we assume that:

Assumption 4.2. The first objective term f is strongly convex with a modulus $\mu_f > 0$, but the second one ψ is not necessarily strongly convex.

Note that problem (38) is not necessarily strongly convex due to the separability of variables x and w in f and ψ, respectively.

We modify Algorithm 2 as follows to solve (38):

$$\left\{\begin{array}{ll}
w^{k+1} &:= \mathrm{argmin}_w\Big\{\psi(w) + \langle B^\top\tilde{y}^k, w\rangle + \frac{\rho_k}{2}\|K\hat{x}^k + Bw - b\|^2 + \frac{\nu_0}{2}\|w - \hat{w}^k\|^2\Big\},\\
y^{k+1} &:= \tilde{y}^k + \rho_k(K\hat{x}^k + Bw^{k+1} - b),\\
\tilde{x}^{k+1} &:= \mathrm{prox}_{(\beta_k/\tau_k)f}\left(\tilde{x}^k - \frac{\beta_k}{\tau_k}K^\top y^{k+1}\right),\\
x^{k+1} &:= \mathrm{prox}_{f/(\rho_k\|K\|^2)}\left(\hat{x}^k - \frac{1}{\rho_k\|K\|^2}K^\top y^{k+1}\right),\\
\hat{x}^{k+1} &:= (1-\tau_{k+1})x^{k+1} + \tau_{k+1}\tilde{x}^{k+1},\\
\hat{w}^{k+1} &:= w^{k+1} + \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}(w^{k+1} - w^k),\\
\tilde{y}^{k+1} &:= \tilde{y}^k + \eta_k\left[(Kx^{k+1} + Bw^{k+1} - b) - (1-\tau_k)(Kx^k + Bw^k - b)\right].
\end{array}\right. \tag{39}$$

Here, the parameters τ_k, ρ_k, β_k, and η_k are updated as in Algorithm 2.

In (39), we combine Algorithm 2 and an alternating strategy between x and w, but we do not linearize the w^{k+1}-subproblem to avoid imposing strong convexity on ψ. When necessary, we add a proximal term $\frac{\nu_0}{2}\|w - \hat{w}^k\|^2$ to guarantee that the w^{k+1}-subproblem always has optimal solution. Note that if B is invertible, then (38) reduces to (1) with $g(\hat{K}x) := \psi(-B^{-1}(Kx - b))$, where $\hat{K} := -B^{-1}K$. In this case, we could apply accelerated proximal gradient methods in [1, 3] to the dual problem (2) and using the strategy in [42, 43] to recover a primal approximate solution, but the optimal rate would no longer be non-ergodic. Our method is accelerated on the primal problem instead of the dual one as in [42, 43].

Finally, we state the convergence of our new scheme (39) to solve (38) in the following corollary, whose proof is given in Appendix C.4.

Corollary 15. *Let $\{(x^k, w^k, \tilde{y}^k)\}$ be generated by (39) to solve (38) and its dual problem under Assumptions 2.1 and 4.2. Let $(x^\star, w^\star, y^\star)$ be a triple of primal-dual optimal solution. If we update the parameters as in (29) and (30), then*

$$|F(x^k, w^k) - F^\star| \leq \frac{2R_0^2}{(k+1)^2} \quad \text{and} \quad \|Kx^k + Bw^k - b\| \leq \frac{2R_0^2}{(k+1)^2}, \tag{40}$$

where $R_0^2 := \frac{\rho_0\|K\|^2}{\Gamma}\|x^0 - x^\star\|^2 + \nu_0\|w^0 - w^\star\|^2 + \frac{1}{\rho_0(1-\gamma)}(2\|y^\star\| + \|y^0\| + 1)^2$. Hence, the objective residual and the feasibility violation both converge to zero at $\mathcal{O}\left(1/k^2\right)$ rate in non-ergodic sense.

If we update the parameters as in (29) and (31), then the $\mathcal{O}\left(1/k^2\right)$ non-ergodic convergence rate bounds on $|F(x^k, w^k) - F^\star|$ and $\|Kx^k + Bw^k - b\|$ still hold, and

$$\liminf_{k\to\infty} k^2\sqrt{\log k}\left[|F(x^k, w^k) - F^\star| + \|Kx^k + Bw^k - b\|\right] = 0. \tag{41}$$

Hence, the objective residual and feasibility violation sequences both converge to zero at $\mathcal{O}\left(1/k^2\right)$ and $\underline{o}\big(1/(k^2\sqrt{\log k})\big)$ non-ergodic convergence rate.

Remark 16. To the best of our knowledge, Corollary 15 presents the first fast $\min\{\mathcal{O}\left(1/k^2\right), \underline{o}\big(1/(k^2\sqrt{\log k})\big)\}$ convergence result for general constrained convex problem (38) under the semi-strong convexity assumption, i.e., f is strongly convex, but ψ is non-strongly convex.

5. Numerical illustrations. We verify the theoretical statements in this paper through two well-known examples. Our code is implemented in MATLAB (R2014b) running on a MacBook Pro with 2.7 GHz Intel Core i5 and 16GB memory, and available at https://github.com/quoctd/PrimalDualCvxOpt.

5.1. Ergodic vs. non-ergodic convergence rates. We consider the following nonsmooth composite convex minimization problem:

$$F^\star := \min_{x\in\mathbb{R}^p}\Big\{F(x) := f(x) + \|Kx - b\|_1\Big\}, \tag{42}$$

where $K \in \mathbb{R}^{n\times p}$, $b \in \mathbb{R}^n$, and $f(x)$ is a regularizer. This problem fits template (1) with $g(r) := \|r - b\|_1$. We compare Algorithm 1 with two well-established methods: Chambolle-Pock's method (CP) [10] and ADMM [7]. When f is strongly convex, we compare Algorithm 2 with the strongly convex variant of CP (CP-scvx) [10, 12].

Case 1 (*General convex case*). We choose $f(x) := \lambda\|x\|_1$ in (42) with a regularization parameter $\lambda = 0.05$. We generate the entries of K from $\mathcal{N}(0, 1)$, and set $b := Kx^\natural + e$, where $x^\natural$ is an s-sparse vector, and e is a sparse Gaussian noise with variance $\sigma^2 := 0.01$ and 10% nonzero entries. The problem size is $(n, p) = (2000, 640)$.

We run two variants of Algorithm 1 with $c = 1$ and $c = 2$. Since CP and ADMM both have $\mathcal{O}\left(1/k\right)$ convergence rate on only the *ergodic* sequence of the relative objective residual $\frac{F(\bar{x}^k)-F^\star}{\max\{1,\ |F^\star|\}}$, we compare this sequence with the *non-ergodic* sequence of Algorithm 1, so that all algorithms under comparison have some theoretical guarantee. Here, $F^\star$ is computed by Mosek [41] with the highest precision.

For Algorithm 1, we use $\rho_0 := 5\cdot\left(\frac{\gamma}{1-\gamma}\right)^{1/2}\cdot\frac{\|y^0-y^\star\|}{\|K\|\|x^0-x^\star\|}$ with $\gamma := 0.999$ as guided by Theorems 2 and 5. For CP method, we choose its step-sizes $\rho := \rho_0$ and $\beta := \frac{\gamma}{\|K\|^2\rho}$. To be fair in our comparison, we also try step-sizes $0.1\rho_0$ and $10\rho_0$. For ADMM, we reformulate (42) into the constrained problem (7) by introducing $r := Kx - b$. Similar to the CP method, we tune the penalty parameter for ADMM and find that three different values $\rho := 0.5\rho_0$, $10\rho_0$, and $30\rho_0$ represent the best range for ρ.

The relative objective residuals are plotted in Figure 1 (left) for the non-ergodic (last-iterate) sequence of Algorithm 1 and for the ergodic (averaging) sequence of CP and ADMM. All algorithms achieve $\mathcal{O}\left(1/k\right)$ rate. The ergodic sequences of CP and ADMM, while having theoretical convergence guarantees, are slower than ours.

Case 2 (Strongly convex case). We choose $f(x) := \lambda\|x\|_1 + \frac{\mu_f}{2}\|x\|^2$ in (42) with $\lambda := 0.05$ and $\mu_f := 0.1$, and generate problem instances the same way as in *Case 1* but with 50% correlated columns in K.

Since f is μ_f-strongly convex, we test Algorithm 2 on (42). If we use (30) to update parameters, then we choose $\gamma = 0.999$ and $\rho_0^1 := \frac{\Gamma\mu_f}{2\|K\|^2}$. We also run a variant

with $\rho_0^{1+} := \frac{5\Gamma\mu_f}{2\|K\|^2}$ since it leads to empirically better performance, suggesting that our analysis in Theorem 11 may not be tight. If we use (31) to update parameters, we choose $c := 4$, $\gamma = 0.75$, and $\rho_0^2 := \frac{c(c-1)\Gamma\mu_f}{(2c-1)\|K\|^2}$. For comparison, we implement CP-scvx in [10] with initial penalty parameter $\rho^{\text{CP}} := \frac{1}{\|K\|}$ as suggested in convergence analysis in [12]. We also test its variants with $0.01\rho^{\text{CP}}$, $0.75\rho^{\text{CP}}$ and $5\rho^{\text{CP}}$.

The convergence behavior of this test is plotted in Figure 1 (right). Both Algorithm 2 and CP-scvx show $\mathcal{O}\left(1/k^2\right)$ convergence rate as predicted by the theory. Algorithm 2 using the update (31) with $c = 4$ is the fastest. CP-scvx, on the other hand, is sensitive to the parameter choice, and even its best variant underperforms variants of Algorithm 2 with parameters ρ_0^{1+} and ρ_0^2.

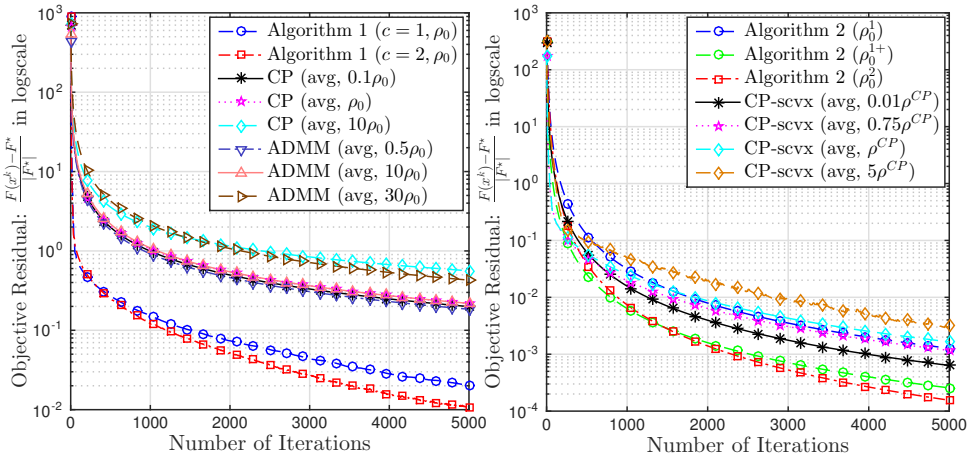

FIG. 1. *Convergence behavior of algorithmic variants on* (42) *with* K *of size* $(n, p) = (2000, 640)$. *Left: Case 1 (general convex) with* 8 *variants; Right: Case 2 (strongly convex) with* 7 *variants.*

5.2. Primal-dual methods vs. smoothing techniques. Consider the following matrix min-max game problem studied, e.g., in [47]:

$$F^\star := \min_{x\in\Delta_p}\Big\{F(x) := \max_{y\in\Delta_n}\langle Kx, y\rangle\Big\}, \tag{43}$$

where $K \in \mathbb{R}^{n\times p}$, $\Delta_p := \{x \in \mathbb{R}^p_+ \mid \sum_{j=1}^p x_j = 1\}$ and $\Delta_n := \{y \in \mathbb{R}^n_+ \mid \sum_{i=1}^n y_i = 1\}$ are two standard simplexes in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively. This problem can be cast into our template (3) with $f(x) := \delta_{\Delta_p}(x)$ and $g^*(y) := \delta_{\Delta_n}(y)$, where Δ is the indicator function. In our experiment, the problem size is $(n, p) = (1000, 2000)$, and K is 10%-sparse with nonzero entries generated from Uniform$(-1, 1)$ distribution, then K is normalized such that $\|K\| = 1$.

We compare Algorithm 1 and Nesterov's smoothing technique in [47]. They both achieve the same theoretical $\mathcal{O}\left(1/k\right)$ convergence rate, but the performance of smoothing techniques depends on the choice of accuracy, as illustrated [47].

For Algorithm 1, we choose $\gamma := 0.5$ and $\rho_0 := \frac{1}{\|K\|} = 1$ to balance the upper bound in Theorem 2. We also update τ_k with $c := 1$ and $c := 2$ to obtain two variants.

For Nesterov's smoothing technique, since $\|K\| = 1$, we use Euclidean distance to smooth $F(x) := \max_{y\in\Delta_n}\langle Kx, y\rangle$ as $F_\mu(x) := \max_{y\in\Delta_n}\{\langle Kx, y\rangle - \frac{\mu}{2}\|y - y_c\|^2\}$, which gives a better complexity bound than entropy proximity functions [47, (4.11)]. Here

$\mu > 0$ is the smoothness parameter and $y_c := (1/n, \cdots, 1/n)^\top$ is the center of Δ_n. As suggested in [47, (4.8)], once the accuracy $\varepsilon > 0$ is fixed, we accordingly set the number of iterations $k_{\max} := \frac{4\|K\|}{\varepsilon}\left[(1-\frac{1}{n})(1-\frac{1}{p})\right]^{1/2}$ and the smoothness parameter $\mu := \frac{\varepsilon}{2(1-1/n)}$. We also run this algorithm with 5μ and $\mu/5$ to observe its sensitivity to the choice of μ.

We run Algorithm 1 and Nesterov's smoothing method using the above configurations. If we test two cases with $\varepsilon_1 = 10^{-3}$ and $\varepsilon_2 = 10^{-4}$, then the corresponding numbers of iterations are $k_{\max,\,1} := 3,997$ and $k_{\max,\,2} := 39,970$, respectively. The duality gap $F(x^k) + G(\bar{y}^k)$ of this test is plotted in Figure 2, where we observe that all algorithms indeed follow the $\mathcal{O}(1/k)$ convergence rate. However, the performance of Nesterov's smoothing method crucially depends on the choice of smoothness parameter μ. Algorithm 1 with $c = 2$ outperforms all other methods in both cases.

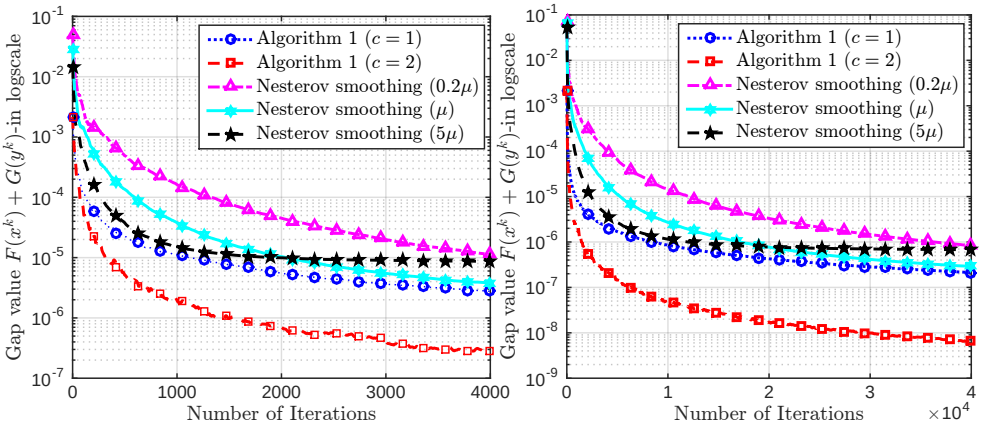

FIG. 2. *The convergence behavior of* 5 *algorithmic variants on* (43) *with a* 10%*-sparse matrix* K *of size* $(n, p) := (1000, 2000)$. *Left:* $\varepsilon_1 = 10^{-3}$; *Right:* $\varepsilon_2 = 10^{-4}$.

Acknowledgments. This paper is based upon work partially supported by the National Science Foundation (NSF), grant no. DMS-1619884, and the Office of Naval Research (ONR), grant No. N00014-20-1-2088.

Appendix A. Some elementary results. We recall some useful facts that will be used in the proof of our main results.

LEMMA 17. *The following statements hold:*

(a) *Let* $\phi_\rho(x, r, y) := \frac{\rho}{2}\|Kx - r\|^2 + \langle y, Kx - r\rangle$ *be defined in* (9) *for* $\rho > 0$. *Then,* $\nabla_x\phi_\rho(x, r, y) = K^\top(y + \rho(Kx - r))$ *and* $\nabla_r\phi_\rho(x, r, y) = \rho(r - Kx) - y$. *Furthermore, for any* $x, x' \in \mathbb{R}^p$ *and* $r, r', y \in \mathbb{R}^n$, *we have*

$$\begin{aligned}\phi_\rho(x', r', y) \;=\; &\phi_\rho(x, r, y) + \langle \nabla_x\phi_\rho(x, r, y), x' - x\rangle \\ &+ \langle \nabla_r\phi_\rho(x, r, y), r' - r\rangle + \tfrac{\rho}{2}\|K(x' - x) - (r' - r)\|^2.\end{aligned} \tag{44}$$

(b) *For any* $u, v, w \in \mathbb{R}^p$ *and* $\alpha_1, \alpha_2 \in \mathbb{R}$, $\alpha_1 + \alpha_2 \neq 0$, *it holds that*

$$\alpha_1\|u-w\|^2+\alpha_2\|v-w\|^2 = (\alpha_1+\alpha_2)\|w-\tfrac{1}{\alpha_1+\alpha_2}(\alpha_1u+\alpha_2v)\|^2+\tfrac{\alpha_1\alpha_2}{\alpha_1+\alpha_2}\|u-v\|^2.$$

(c) *If a nonnegative sequence* $\{u_k\} \subseteq [0, +\infty)$ *satisfies* $\sum_{i=0}^{\infty} u_k < +\infty$, *then* $\liminf_{k\to\infty} k\log(k)u_k = 0$.

(d) *Let* $\{u_k\}$ *and* $\{v_k\}$ *be two nonnegative sequences in* $\mathbb{R}$ *and* $\alpha_1, \alpha_2 \in \mathbb{R}_{++}$ *be two positive constants. Then, the following statements hold:*
  (i) *If* $\liminf_{k\to\infty}\left[k\log(k)(u_k+\alpha_1 k v_k^2)\right] = 0$, *then* $\liminf_{k\to\infty}\left[k\sqrt{\log k}(u_k+\alpha_2 v_k)\right] = 0$.
  (ii) *If* $\liminf_{k\to\infty}\left[k^2\log(k)(u_k + \alpha_1 k^2 v_k^2)\right] = 0$, *then* $\liminf_{k\to\infty}\left[k^2\sqrt{\log k}(u_k + \alpha_2 v_k)\right] = 0$.

Proof. The statements (a) and (b) are trivial. We only prove parts (c) and (d).

(c) Since $u_k \geq 0$ and the lim inf of a lower bounded sequence always exists, we set $\bar{u} := \liminf_{k\to\infty} k\log(k)u_k \geq 0$. Suppose that $\bar{u} > 0$. Then, by definition, for any $\varepsilon > 0$ such that $\bar{u} - \varepsilon > 0$, there exists an integer $k_\varepsilon > 0$ such that for any $k \geq k_\varepsilon$, we have $k\log(k)u_k \geq \bar{u} - \varepsilon$. This leads to

$$+\infty > \sum_{k=0}^{\infty} u_k \geq \sum_{k=k_\varepsilon}^{\infty} u_k \geq \sum_{k=k_\varepsilon}^{\infty} \frac{\bar{u}-\varepsilon}{k\log k} = (\bar{u}-\varepsilon)\sum_{k=k_\varepsilon}^{\infty}\frac{1}{k\log k} = +\infty,$$

which is a contradiction. Hence, we must have $\bar{u} = 0$.

(d) Since $\liminf_{k\to\infty} k\log(k)(u_k + \alpha_1 k v_k^2) = 0$, there exists a convergent subsequence $\{k_j\log(k_j)(u_{k_j} + \alpha_1 k_j v_{k_j}^2)\}_{j\geq 0}$ converging to 0, i.e., for any $\varepsilon > 0$, there exists $j_0 \geq 0$ such that for all $j \geq j_0$, we have $k_j\log(k_j)(u_{k_j} + \alpha_1 k_j v_{k_j}^2) < \min\left\{\frac{\varepsilon}{2}, \frac{\alpha_1\varepsilon^2}{4\alpha_2^2}\right\}$. This inequality implies that

$$\begin{aligned} & k_j\sqrt{\log(k_j)}u_{k_j} < \tfrac{\varepsilon}{2} \\ \text{and} \quad & \alpha_2 k_j\sqrt{\log(k_j)}v_{k_j} = \tfrac{\alpha_2}{\sqrt{\alpha_1}}\sqrt{\alpha_1 k_j^2\log(k_j)v_{k_j}^2} < \tfrac{\alpha_2}{\sqrt{\alpha_1}}\sqrt{\tfrac{\alpha_1\varepsilon^2}{4\alpha_2^2}} = \tfrac{\varepsilon}{2}. \end{aligned}$$

Combining both inequalities, we can show that $k_j\sqrt{\log(k_j)}(u_{k_j}+\alpha_1 k_j v_{k_j}) < \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon$, which proves part (i) of (d). Part (ii) of (d) can be proved analogously. ∎

Appendix B. Technical proofs in Section 3: General convex case. This appendix provides the full proof of technical results in Section 3.

B.1. The proof of Lemma 1: One-iteration analysis. First, we write down the optimality conditions of x^{k+1} and r^{k+1} in (10) as follows:

$$\left\{\begin{array}{ll} 0 &\in \partial g(r^{k+1}) + \rho_k(r^{k+1} - K\hat{x}^k) - \tilde{y}^k \equiv \partial g(r^{k+1}) + \nabla_r\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k),\\ 0 &\in \partial f(x^{k+1}) + \nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k) + \frac{1}{\beta_k}(x^{k+1} - \hat{x}^k). \end{array}\right. \tag{45}$$

By convexity of f and g, and the above optimality conditions, we can derive

$$\left\{\begin{array}{ll} g(r^{k+1}) &\leq g(r) + \langle\nabla g(r^{k+1}), r^{k+1} - r\rangle \overset{(45)}{=} g(r) + \langle\nabla_r\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k), r - r^{k+1}\rangle,\\ f(x^{k+1}) &\leq f(x) + \langle\nabla f(x^{k+1}), x^{k+1} - x\rangle\\ &\overset{(45)}{=} f(x) + \langle\nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k), x - x^{k+1}\rangle + \frac{1}{\beta_k}\langle x^{k+1} - \hat{x}^k, x - x^{k+1}\rangle, \end{array}\right. \tag{46}$$

where $\nabla f(x^{k+1}) \in \partial f(x^{k+1})$ and $\nabla g(r^{k+1}) \in \partial g(r^{k+1})$.
Next, using Lemma 17(a) twice, we can derive

$$\left\{\begin{array}{l} \phi_{\rho_k}(x^{k+1}, r^{k+1}, \tilde{y}^k) = \phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k) + \langle\nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k), x^{k+1} - \hat{x}^k\rangle\\ \qquad\qquad + \frac{\rho_k}{2}\|K(x^{k+1} - \hat{x}^k)\|^2,\\ \phi_{\rho_k}(x, r, \tilde{y}^k) = \phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k) + \langle\nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k), x - \hat{x}^k\rangle\\ \qquad\qquad + \langle\nabla_r\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k), r - r^{k+1}\rangle + \frac{\rho_k}{2}\|K(x - \hat{x}^k) - (r - r^{k+1})\|^2. \end{array}\right. \tag{47}$$

Combining the two expressions in (47), we get

$$\begin{aligned}\phi_{\rho_k}(x^{k+1},r^{k+1},\tilde{y}^k) &= \phi_{\rho_k}(x,r,\tilde{y}^k) + \langle\nabla_x\phi_{\rho_k}(\hat{x}^k,r^{k+1},\tilde{y}^k), x^{k+1}-x\rangle\\ &\quad + \langle\nabla_r\phi_{\rho_k}(\hat{x}^k,r^{k+1},\tilde{y}^k), r^{k+1}-r\rangle + \tfrac{\rho_k}{2}\|K(x^{k+1}-\hat{x}^k)\|^2\\ &\quad - \tfrac{\rho_k}{2}\|K(x-\hat{x}^k)-(r-r^{k+1})\|^2.\end{aligned} \tag{48}$$

Summing up (46) and (48) and using (9), we arrive at

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},r^{k+1},\tilde{y}^k) &\le \mathcal{L}_{\rho_k}(x,r,\tilde{y}^k) + \tfrac{1}{\beta_k}\langle x^{k+1}-\hat{x}^k, x-\hat{x}^k\rangle - \tfrac{1}{\beta_k}\|x^{k+1}-\hat{x}^k\|^2\\ &\quad + \tfrac{\rho_k}{2}\|K(x^{k+1}-\hat{x}^k)\|^2 - \tfrac{\rho_k}{2}\|K(x-\hat{x}^k)-(r-r^{k+1})\|^2.\end{aligned} \tag{49}$$

Since (49) holds for any $x\in\mathbb{R}^p$ and $r\in\mathbb{R}^n$, we can substitute $(x,r) := (x^k,r^k)$ to obtain

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},r^{k+1},\tilde{y}^k) &\le \mathcal{L}_{\rho_k}(x^k,r^k,\tilde{y}^k) + \tfrac{1}{\beta_k}\langle x^{k+1}-\hat{x}^k, x^k-\hat{x}^k\rangle\\ &\quad - \tfrac{1}{\beta_k}\|x^{k+1}-\hat{x}^k\|^2 + \tfrac{\rho_k}{2}\|K(x^{k+1}-\hat{x}^k)\|^2 - \tfrac{\rho_k}{2}\|K(x^k-\hat{x}^k)-(r^k-r^{k+1})\|^2.\end{aligned} \tag{50}$$

Multiplying (49) by τ_k and (50) by $1-\tau_k$, summing up the results, then utilizing $\hat{x}^k = (1-\tau_k)x^k + \tau_k\tilde{x}^k$ from (10), we get

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},r^{k+1},\tilde{y}^k) &\le (1-\tau_k)\mathcal{L}_{\rho_k}(x^k,r^k,\tilde{y}^k) + \tau_k\mathcal{L}_{\rho_k}(x,r,\tilde{y}^k)\\ &\quad + \tfrac{\tau_k}{\beta_k}\langle x^{k+1}-\hat{x}^k, x-\tilde{x}^k\rangle - \tfrac{1}{\beta_k}\|x^{k+1}-\hat{x}^k\|^2 + \tfrac{\rho_k}{2}\|K(x^{k+1}-\hat{x}^k)\|^2\\ &\quad - \tfrac{\tau_k\rho_k}{2}\|K(\hat{x}^k-x)-(r^{k+1}-r)\|^2 - \tfrac{(1-\tau_k)\rho_k}{2}\|K(x^k-\hat{x}^k)-(r^k-r^{k+1})\|^2.\end{aligned} \tag{51}$$

By the definition of $\tilde{y}^{k+1}$ from (10), we have

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},r^{k+1},y) - (1-\tau_k)\mathcal{L}_{\rho_k}(x^k,r^k,y) &= \mathcal{L}_{\rho_k}(x^{k+1},r^{k+1},\tilde{y}^k)\\ &\quad - (1-\tau_k)\mathcal{L}_{\rho_k}(x^k,r^k,\tilde{y}^k) + \tfrac{1}{2\eta_k}\left[\|\tilde{y}^k-y\|^2 - \|\tilde{y}^{k+1}-y\|^2 + \|\tilde{y}^{k+1}-\tilde{y}^k\|^2\right].\end{aligned} \tag{52}$$

Substituting the expression (52) into (51), we get

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},r^{k+1},y) &\le (1-\tau_k)\mathcal{L}_{\rho_k}(x^k,r^k,y) && (=:\mathcal{T}_1)\\ &\quad + \tau_k\mathcal{L}_{\rho_k}(x,r,\tilde{y}^k) - \tfrac{\tau_k\rho_k}{2}\|K\hat{x}^k - r^{k+1} - (Kx-r)\|^2 && (=:\mathcal{T}_2)\\ &\quad + \tfrac{\tau_k}{\beta_k}\langle x^{k+1}-\hat{x}^k, x-\tilde{x}^k\rangle - \tfrac{1}{2\beta_k}\|x^{k+1}-\hat{x}^k\|^2 && (=:\mathcal{T}_3)\\ &\quad - \tfrac{1}{2\beta_k}\|x^{k+1}-\hat{x}^k\|^2 + \tfrac{1}{2\eta_k}\left[\|\tilde{y}^k-y\|^2 - \|\tilde{y}^{k+1}-y\|^2 + \|\tilde{y}^{k+1}-\tilde{y}^k\|^2\right]\\ &\quad + \tfrac{\rho_k}{2}\|K(x^{k+1}-\hat{x}^k)\|^2 - \tfrac{(1-\tau_k)\rho_k}{2}\|K(x^k-\hat{x}^k)-(r^k-r^{k+1})\|^2.\end{aligned} \tag{53}$$

Now, we estimate three terms $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ in (53) as follows. First, we have

$$\mathcal{T}_1 = (1-\tau_k)\left[\mathcal{L}_{\rho_{k-1}}(x^k,r^k,y) + \frac{(\rho_k-\rho_{k-1})}{2}\|Kx^k-r^k\|^2\right]. \tag{54}$$

By the definitions of $\bar{y}^{k+1}$ in Lemma 1 and of $\mathcal{L}$, we can show that

$$\mathcal{T}_2 = \mathcal{L}(x,r,\bar{y}^{k+1}) - (1-\tau_k)\mathcal{L}(x,r,\bar{y}^k) - \frac{\tau_k\rho_k}{2}\|K\hat{x}^k - r^{k+1}\|^2. \tag{55}$$

Moreover, by the update $\tilde{x}^{k+1} := \tilde{x}^k + \frac{1}{\tau_k}(x^{k+1}-\hat{x}^k)$ in (10), we can further derive

$$\mathcal{T}_3 = \frac{\tau_k^2}{2\beta_k}\left[\|\tilde{x}^k-x\|^2 - \|\tilde{x}^{k+1}-x\|^2\right].$$

Substituting these three terms $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ back into (53), we obtain

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},r^{k+1},y) \;\le\; & (1-\tau_k)\mathcal{L}_{\rho_{k-1}}(x^k,r^k,y) + \tfrac{(1-\tau_k)(\rho_k-\rho_{k-1})}{2}\|Kx^k-r^k\|^2\\ & + \mathcal{L}(x,r,\bar{y}^{k+1}) - (1-\tau_k)\mathcal{L}(x,r,\bar{y}^k) - \tfrac{\tau_k\rho_k}{2}\|K\hat{x}^k-r^{k+1}\|^2\\ & + \tfrac{\tau_k^2}{2\beta_k}\left[\|\tilde{x}^k-x\|^2-\|\tilde{x}^{k+1}-x\|^2\right] - \tfrac{1}{2\beta_k}\|x^{k+1}-\hat{x}^k\|^2\\ & + \tfrac{1}{2\eta_k}\left[\|\tilde{y}^k-y\|^2-\|\tilde{y}^{k+1}-y\|^2\right] + \tfrac{1}{2\eta_k}\|\tilde{y}^{k+1}-\tilde{y}^k\|^2\\ & + \tfrac{\rho_k}{2}\|K(x^{k+1}-\hat{x}^k)\|^2 - \tfrac{(1-\tau_k)\rho_k}{2}\|K(x^k-\hat{x}^k)-(r^k-r^{k+1})\|^2,\end{aligned}$$

which, after rearrangement, becomes

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},r^{k+1},y) - \mathcal{L}(x,r,\bar{y}^{k+1}) \le\; & (1-\tau_k)\left[\mathcal{L}_{\rho_{k-1}}(x^k,r^k,y)-\mathcal{L}(x,r,\bar{y}^k)\right]\\ & + \tfrac{\tau_k^2}{2\beta_k}\left[\|\tilde{x}^k-x\|^2-\|\tilde{x}^{k+1}-x\|^2\right] - \tfrac{1}{2\beta_k}\|x^{k+1}-\hat{x}^k\|^2 + \tfrac{\rho_k}{2}\|K(x^{k+1}-\hat{x}^k)\|^2\\ & + \tfrac{1}{2\eta_k}\left[\|\tilde{y}^k-y\|^2-\|\tilde{y}^{k+1}-y\|^2\right] + \tfrac{1}{2\eta_k}\|\tilde{y}^{k+1}-\tilde{y}^k\|^2 + \mathcal{T}_4,\end{aligned} \tag{56}$$

where

$$\begin{aligned}\mathcal{T}_4 \;:=\; & \tfrac{(1-\tau_k)}{2}(\rho_k-\rho_{k-1})\|Kx^k-r^k\|^2 - \tfrac{\tau_k\rho_k}{2}\|K\hat{x}^k-r^{k+1}\|^2\\ & - \tfrac{(1-\tau_k)\rho_k}{2}\|K(x^k-\hat{x}^k)-(r^k-r^{k+1})\|^2\\ =\; & -\tfrac{\rho_k}{2}\|(K\hat{x}^k-r^{k+1})-(1-\tau_k)(Kx^k-r^k)\|^2\\ & - \tfrac{(1-\tau_k)}{2}\left[\rho_{k-1}-(1-\tau_k)\rho_k\right]\|Kx^k-r^k\|^2.\end{aligned} \tag{57}$$

Using Lemma 17(b) with $u := Kx^{k+1}-r^{k+1}$, $v := K\hat{x}^k-r^{k+1}$, $w := (1-\tau_k)(Kx^k-r^k)$, $\alpha_1 := \eta_k/2$ and $\alpha_2 := -\rho_k/2$, and $\tilde{y}^{k+1}$ from (10) with $\rho_k > \eta_k$, we can show that

$$\tfrac{1}{2\eta_k}\|\tilde{y}^{k+1}-\tilde{y}^k\|^2 - \tfrac{\rho_k}{2}\|(K\hat{x}^k-r^{k+1})-(1-\tau_k)(Kx^k-r^k)\|^2 \le \tfrac{\rho_k\eta_k}{2(\rho_k-\eta_k)}\|K(x^{k+1}-\hat{x}^k)\|^2. \tag{58}$$

Substituting this estimate and (57) into (56), we finally arrive at (11). □

B.2. The proof of Theorem 2: $\mathcal{O}(1/k)$ convergence rates when $c = 1$. Using the parameter update rule (15) with $c := 1$, we can easily verify that

$$\frac{\tau_k^2}{\beta_k} = \frac{(1-\tau_k)\tau_{k-1}^2}{\beta_{k-1}}, \quad \frac{1}{\eta_k} = \frac{1-\tau_k}{\eta_{k-1}}, \quad \frac{1}{\beta_k} - \frac{\rho_k^2\|K\|^2}{\rho_k-\eta_k} = 0, \quad \text{and} \quad \rho_{k-1}-(1-\tau_k)\rho_k = 0.$$

Applying these conditions to (11) of Lemma 1, we can simplify it as

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},r^{k+1},y) - \mathcal{L}(x,r,\bar{y}^{k+1}) + \tfrac{\tau_k^2}{2\beta_k}\|\tilde{x}^{k+1}-x\|^2 + \tfrac{1}{2\eta_k}\|\tilde{y}^{k+1}-y\|^2\\ \le (1-\tau_k)\left[\mathcal{L}_{\rho_{k-1}}(x^k,r^k,y)-\mathcal{L}(x,r,\bar{y}^k) + \tfrac{\tau_{k-1}^2}{2\beta_{k-1}}\|\tilde{x}^k-x\|^2 + \tfrac{1}{2\eta_{k-1}}\|\tilde{y}^k-y\|^2\right].\end{aligned}$$

By induction, this inequality implies

$$\begin{aligned}\mathcal{L}_{\rho_{k-1}}(x^k,r^k,y) - \mathcal{L}(x,r,\bar{y}^k) \le\; & \left[\textstyle\prod_{i=1}^{k-1}(1-\tau_i)\right]\times\\ & \left[(1-\tau_0)\left(\mathcal{L}_{\rho_{-1}}(x^0,r^0,y)-\mathcal{L}(x,r,\bar{y}^0)\right) + \tfrac{\tau_0^2}{2\beta_0}\|\tilde{x}^0-x\|^2 + \tfrac{1}{2\eta_0}\|\tilde{y}^0-y\|^2\right].\end{aligned}$$

If $c = 1$, then $\tau_0 = 1$, $\beta_0 = \gamma/\left(\|K\|^2\rho_0\right)$, and $\eta_0 = (1-\gamma)\rho_0$. We also have $\Pi_{i=1}^{k-1}(1-\tau_i) = \frac{1}{k}$, $\tilde{x}^0 = x^0$, and $\tilde{y}^0 = y^0$. Thus the last estimate can be simplified as

$$\mathcal{L}_{\rho_{k-1}}(x^k,r^k,y) - \mathcal{L}(x,r,\bar{y}^k) \le \frac{1}{2k}\left[\frac{\rho_0\|K\|^2\|x^0-x\|^2}{\gamma} + \frac{\|y^0-y\|^2}{(1-\gamma)\rho_0}\right]. \tag{59}$$

Now, let pick any $\bar{r}^k \in \partial g^*(\bar{y}^k)$. Then, by (8), we easily get

$$\widetilde{\mathcal{L}}(x^k, y) - \widetilde{\mathcal{L}}(x, \bar{y}^k) \le \mathcal{L}(x^k, r^k, y) - \mathcal{L}(x, \bar{r}^k, \bar{y}^k). \tag{60}$$

Combining (59), (60), and $\mathcal{L}(x^k, r^k, y) \le \mathcal{L}_{\rho_{k-1}}(x^k, r^k, y)$, we finally get (16).

(a) The estimate (17) directly follows from (16) and the definition of $\mathcal{G}_{\mathcal{X}\times\mathcal{Y}}$ in (6).

(b) By M_g-Lipschitz continuity of g, we have

$$\begin{aligned} F(x^k) - F^\star &\le f(x^k) + g(r^k) + M_g\|Kx^k - r^k\| - F^\star \\ &= f(x^k) + g(r^k) + \langle \breve{y}^k, Kx^k - r^k\rangle - F^\star \quad \text{with } \breve{y}^k := \tfrac{M_g(Kx^k - r^k)}{\|Kx^k - r^k\|} \\ &\le \mathcal{L}_{\rho_{k-1}}(x^k, r^k, \breve{y}^k) - \mathcal{L}(x^\star, r^\star, \bar{y}^k), \end{aligned} \tag{61}$$

where we have used $\mathcal{L}(x^k, r^k, \breve{y}^k) = f(x^k) + g(r^k) + \langle Kx^k - r^k, \breve{y}^k\rangle \le \mathcal{L}_{\rho_{k-1}}(x^k, r^k, \breve{y}^k)$ and $F^\star = \mathcal{L}(x^\star, r^\star, \bar{y}^k)$ in the last inequality. Substituting $(x, r, y) := (x^\star, r^\star, \breve{y}^k)$ into (59), we have

$$0 \le F(x^k) - F^\star \le \frac{1}{2k}\left[\frac{\rho_0\|K\|^2\|x^0 - x^\star\|^2}{\gamma} + \frac{\|y^0 - \breve{y}^k\|^2}{(1-\gamma)\rho_0}\right].$$

By the definition of $\breve{y}^k$ in (61), we have $\|y^0 - \breve{y}^k\| \le \sup_y\{\|y^0 - y\| \mid \|y\| \le M_g\} =: D_g$. Using this estimate into the last inequality, we obtain (18).

(c) For any $x \in \mathbb{R}^p$ and $r \in \mathbb{R}^n$, we have

$$\begin{aligned} \mathcal{L}(x, r, \bar{y}^k) &= f(x) - \langle -K^\top\bar{y}^k, x\rangle + g(r) - \langle \bar{y}^k, r\rangle \\ &\ge -\sup_x\{\langle -K^\top\bar{y}^k, x\rangle - f(x)\} - \sup_r\{\langle \bar{y}^k, r\rangle - g(r)\} \\ &= -f^*(-K^\top\bar{y}^k) - g^*(\bar{y}^k) = -G(\bar{y}^k). \end{aligned}$$

Let $\breve{x}^k \in \partial f^*(-K^\top\bar{y}^k)$ and $\breve{r}^k \in \partial g^*(\bar{y}^k)$. Then, it is clear that the above inequality holds as equality with $(x, r) := (\breve{x}^k, \breve{r}^k)$. We further have

$$G(\bar{y}^k) - G^\star = F^\star - \mathcal{L}(\breve{x}^k, \breve{r}^k, \bar{y}^k) \le \mathcal{L}(x^k, r^k, y^\star) - \mathcal{L}(\breve{x}^k, \breve{r}^k, \bar{y}^k),$$

since $F^\star + G^\star = 0$ and $F^\star \le \mathcal{L}(x^k, r^k, y^\star)$. Combining this inequality and (59) yields

$$0 \le G(\bar{y}^k) - G^\star \le \frac{1}{2k}\left[\frac{\rho_0\|K\|^2\|x^0 - \breve{x}^k\|^2}{\gamma} + \frac{\|y^0 - y^\star\|^2}{(1-\gamma)\rho_0}\right].$$

If f^* is M_{f^*}-Lipschitz continuous, then $\|\breve{x}^k - x^0\| \le \sup_x\{\|x^0 - x\| \mid \|x\| \le M_{f^*}\} =: D_{f^*}$. Substituting this estimate into the last inequality, we prove (19). ☐

B.3. The proof of Theorem 5: $\mathcal{O}(1/k)$ and $o\big(1/(k\sqrt{\log k})\big)$ convergence rates when $c > 1$. Let us first abbreviate $a_k^2 := \frac{\rho_0}{2}\|Kx^k - r^k\|^2$, $b_k^2 := \frac{\rho_0\|K\|^2}{2\gamma}\|\tilde{x}^k - x^\star\|^2 + \frac{1}{2(1-\gamma)\rho_0}\|\tilde{y}^k - y^\star\|^2$, and $\widetilde{\mathcal{G}}_k := \mathcal{L}(x^k, r^k, y^\star) - \mathcal{L}(x^\star, r^\star, \bar{y}^k)$. Since $(x^\star, r^\star, y^\star)$ is a saddle-point of $\mathcal{L}$, we have $\widetilde{\mathcal{G}}_k \equiv \mathcal{L}(x^k, r^k, y^\star) - F^\star \ge 0$. Using the update rules of parameters at Step 4 of Algorithm 1, we can derive from (11) of Lemma 1 that

$$\widetilde{\mathcal{G}}_{k+1} + \frac{k+c}{c}a_{k+1}^2 \le \frac{k}{k+c}\Big(\widetilde{\mathcal{G}}_k + \frac{k+c-1}{c}a_k^2\Big) + \frac{c}{k+c}(b_k^2 - b_{k+1}^2) - \frac{(c-1)k}{c(k+c)}a_k^2.$$

Rearranging this estimate, we obtain

$$\begin{aligned} (c-1)\left(\widetilde{\mathcal{G}}_k + \tfrac{k+c-1}{c}a_k^2\right) &\le (c-1)\Big(\widetilde{\mathcal{G}}_k + \tfrac{2k+c-1}{c}a_k^2\Big) \\ &\le \left[(k+c-1)\widetilde{\mathcal{G}}_k + \tfrac{(k+c-1)^2}{c}a_k^2 + cb_k^2\right] - \left[(k+c)\widetilde{\mathcal{G}}_{k+1} + \tfrac{(k+c)^2}{c}a_{k+1}^2 + cb_{k+1}^2\right]. \end{aligned} \tag{62}$$

Clearly, the estimate (62) implies

$$(k+c)\widetilde{\mathcal{G}}_{k+1} + \frac{(k+c)^2}{c}a_{k+1}^2 + cb_{k+1}^2 \le (k+c-1)\widetilde{\mathcal{G}}_k + \frac{(k+c-1)^2}{c}a_k^2 + cb_k^2.$$

By induction and the definition of $\widetilde{\mathcal{G}}_k$, we can easily show from the last estimate that

$$\begin{aligned}\widetilde{\mathcal{L}}(x^k, y^\star) - F^\star &\le \mathcal{L}(x^k, r^k, y^\star) - F^\star = \widetilde{\mathcal{G}}_k \\ &\le \tfrac{1}{k+c-1}\left[(c-1)\widetilde{\mathcal{G}}_0 + \tfrac{(c-1)^2}{c}a_0^2 + cb_0^2\right] = \tfrac{R_0^2}{k+c-1}.\end{aligned} \tag{63}$$

By the definition of R_0^2 in the statement of Theorem 5, and the fact that $Kx^0 - r^0 = 0$ from the initialization step of Algorithm 1, we have proved the first assertion of (20).

Summing up (62) from $i := 0$ to $i := k$, we get

$$\begin{aligned}(c-1)\textstyle\sum_{i=0}^k\left[\widetilde{\mathcal{G}}_i + \tfrac{(i+c-1)}{c}a_i^2\right] &\le \left[(c-1)\widetilde{\mathcal{G}}_0 + \tfrac{(c-1)^2}{2}a_0^2 + cb_0^2\right] \\ &\quad - \left[(k+c)\widetilde{\mathcal{G}}_{k+1} + \tfrac{(k+c)^2}{c}a_{k+1}^2 + cb_{k+1}^2\right] \\ &\le (c-1)\widetilde{\mathcal{G}}_0 + \tfrac{(c-1)^2}{c}a_0^2 + cb_0^2 < +\infty.\end{aligned} \tag{64}$$

Since $c - 1 > 0$, $i \le i + c - 1$, and $\widetilde{\mathcal{G}}_i \ge 0$, applying Lemma 17(c) to (64), we get

$$\liminf_{k\to\infty} k\log(k)\left[\widetilde{\mathcal{G}}_k + \frac{ka_k^2}{c}\right] = 0. \tag{65}$$

In particular, since $0 \le \widetilde{\mathcal{L}}(x^k, y^\star) - F^\star \le \mathcal{L}(x^k, r^k, y^\star) - F^\star = \widetilde{\mathcal{G}}_k$, we have proved $\liminf_{k\to\infty} k\log(k)\left[\widetilde{\mathcal{L}}(x^k, y^\star) - F^\star\right] = 0$, which is the second assertion of (20).

Analogous to (63), we can show that

$$\|Kx^k - r^k\| \le \frac{\sqrt{2c/\rho_0}R_0}{k+c-1}. \tag{66}$$

By the M_g-Lipschitz continuity of g, similar to (61), we can show that

$$0 \le F(x^k) - F^\star \le \mathcal{L}(x^k, r^k, y^\star) - F^\star + (\|y^\star\| + M_g)\|Kx^k - r^k\|. \tag{67}$$

Combining (63), (66), and (67), we get the first assertion of (21). Moreover, applying Lemma 17(d, part (i)) with $u_k := \widetilde{\mathcal{G}}_k \ge 0$, $v_k := \|Kx^k - r^k\|$, $\alpha_1 := \frac{\rho_0}{2c}$, and $\alpha_2 := \|y^\star\| + M_g$ to (65), we can show that

$$\liminf_{k\to\infty} k\sqrt{\log k}\left[\left(\mathcal{L}(x^k, r^k, y^\star) - F^\star\right) + (\|y^\star\| + M_g)\|Kx^k - r^k\|\right] = 0. \tag{68}$$

Furthermore, using the limit (68) in (67), we obtain the second assertion of (21). □

B.4. The proof of Corollary 7: Constrained problems. From (23), we can write down the optimality condition of x^{k+1} as

$$0 \in \partial f(x^{k+1}) + K^\top y^{k+1} + \nabla\psi(\hat{x}^k) + \tfrac{1}{\beta_k}(x^{k+1} - \hat{x}^k). \tag{69}$$

By convexity of f and L_ψ-smoothness of ψ, for any $x \in \mathbb{R}^p$, we have

$$\psi(x^{k+1}) \le \psi(x) + \langle\nabla\psi(\hat{x}^k), x^{k+1} - x\rangle + \frac{L_\psi}{2}\|x^{k+1} - \hat{x}^k\|^2. \tag{70}$$

Combining (46), (48), (69), and (70) with $r = r^{k+1} = b$, for any $x \in \mathbb{R}^p$, we can derive

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},\tilde{y}^k) &= f(x^{k+1})+\psi(x^{k+1})+\langle\tilde{y}^k,Kx^{k+1}-b\rangle+\tfrac{\rho_k}{2}\|Kx^{k+1}-b\|^2\\ &\le \mathcal{L}_{\rho_k}(x,\tilde{y}^k)+\tfrac{1}{\beta_k}\langle x^{k+1}-\hat{x}^k,x-\hat{x}^k\rangle-\tfrac{1}{\beta_k}\|x^{k+1}-\hat{x}^k\|^2\\ &\quad+\tfrac{\rho_k}{2}\|K(x^{k+1}-\hat{x}^k)\|^2-\tfrac{\rho_k}{2}\|K(x-\hat{x}^k)\|^2+\tfrac{L_\psi}{2}\|x^{k+1}-\hat{x}^k\|^2.\end{aligned}$$

Analogous to the proof for Lemma 1 but using the last estimate, we can show that

$$\begin{aligned}&\mathcal{L}_{\rho_k}(x^{k+1},y)-\mathcal{L}(x,\bar{y}^{k+1})\le(1-\tau_k)\left[\mathcal{L}_{\rho_{k-1}}(x^k,y)-\mathcal{L}(x,\bar{y}^k)\right]\\ &+\tfrac{\tau_k^2}{2\beta_k}\left[\|\tilde{x}^k-x\|^2-\|\tilde{x}^{k+1}-x\|^2\right]+\tfrac{1}{2\eta_k}\left[\|\tilde{y}^k-y\|^2-\|\tilde{y}^{k+1}-y\|^2\right]\\ &-\tfrac{1}{2}\left(\tfrac{1}{\beta_k}-L_\psi-\tfrac{\rho_k^2\|K\|^2}{\rho_k-\eta_k}\right)\|x^{k+1}-\hat{x}^k\|^2-\tfrac{(1-\tau_k)}{2}\left[\rho_{k-1}-(1-\tau_k)\rho_k\right]\|Kx^k-b\|^2.\end{aligned}\tag{71}$$

Using the update (15) with $c:=1$ and $\beta_k:=\gamma/(\|K\|^2\rho_k+\gamma L_\psi)$, for any $x\in\mathbb{R}^p$ and $y\in\mathbb{R}^n$, we follow the same lines as in the proof of Theorem 2 to derive

$$\mathcal{L}_{\rho_{k-1}}(x^k,y)-\mathcal{L}(x,\bar{y}^k)\le\frac{1}{2k}\left[\frac{(\rho_0\|K\|^2+\gamma L_\psi)}{\gamma}\|x^0-x\|^2+\frac{\|y^0-y\|^2}{\rho_0(1-\gamma)}\right],\tag{72}$$

which implies

$$F(x^k)+\langle y,Kx^k-b\rangle+\frac{\rho_{k-1}}{2}\|Kx^k-b\|^2-F^\star\le\frac{R_0^2(y)}{2k},\quad\forall y\in\mathbb{R}^n,$$

where $R_0^2(y):=\frac{(\rho_0\|K\|^2+\gamma L_\psi)}{\gamma}\|x^0-x^\star\|^2+\frac{1}{\rho_0(1-\gamma)}\|y^0-y\|^2$. For any $\lambda>0$, the last inequality leads to

$$F(x^k)-F^\star+\lambda\|Kx^k-b\|+\frac{\rho_{k-1}}{2}\|Kx^k-b\|^2\le\frac{1}{2k}\sup\left\{R_0^2(y)\mid\|y\|\le\lambda\right\}=:\frac{R_0^2}{2k}.$$

On the other hand, we have $F(x^k)-F^\star\ge-\langle y^\star,Kx^k-b\rangle\ge-\|y^\star\|\,\|Kx^k-b\|$. Combining these expressions, we obtain

$$\begin{cases}\left(\lambda-\|y^\star\|\right)\|Kx^k-b\|+\frac{\rho_{k-1}}{2}\|Kx^k-b\|^2\le\frac{R_0^2}{2k},\\ -\|y^\star\|\,\|Kx^k-b\|\le F(x^k)-F^\star\le\frac{R_0^2}{2k}.\end{cases}$$

Choosing $\lambda:=2\|y^\star\|+1$, and noting that $\sup\left\{\|y^0-y\|^2\mid\|y\|\le\lambda\right\}=\left(\lambda+\|y^0\|\right)^2=\left(2\|y^\star\|+\|y^0\|+1\right)^2$, we obtain (24) from the last expression.

Next, let $\breve{x}^k\in\partial F^*(-K^\top\bar{y}^k)$, we have

$$\begin{aligned}G(\bar{y}^k)-G^\star &= \sup_x\left\{\langle-K^\top\bar{y}^k,x\rangle-f(x)-\psi(x)\right\}+\langle b,\bar{y}^k\rangle+F^\star\\ &\le\mathcal{L}(x^k,y^\star)-f(\breve{x}^k)-\psi(\breve{x}^k)-\langle K\breve{x}^k-b,\bar{y}^k\rangle=\mathcal{L}(x^k,y^\star)-\mathcal{L}(\breve{x}^k,\bar{y}^k).\end{aligned}$$

Since $\mathrm{dom}(F)$ is bounded, we have $\|\breve{x}^k-x^0\|\le\sup\left\{\|x-x^0\|\mid x\in\mathrm{dom}(F)\right\}=:\mathcal{D}_F$. Plugging these two last inequalities in (72), we finally obtain (25).

For $c>1$, by the same proof as of (68) but with $\alpha_2:=\|y^\star\|+1$, we get

$$\liminf_{k\to\infty}k\sqrt{\log k}\left[\widetilde{\mathcal{G}}_k+(\|y^\star\|+1)\|Kx^k-b\|\right]=0.$$

In addition, from the proof of Theorem 5, we have $F(x^k)-F^\star+\langle y^\star,Kx^k-b\rangle=\mathcal{L}(x^k,y^\star)-F^\star=:\widetilde{\mathcal{G}}_k\ge0$. Moreover, $F(x^k)-F^\star\ge-\|y^\star\|\|Kx^k-b\|$. Combining these inequalities, we can show that $|F(x^k)-F^\star|\le\widetilde{\mathcal{G}}_k+\|y^\star\|\|Kx^k-b\|$. Consequently, we obtain (26) by combining this inequality and the last limit. □

Appendix C. Technical proofs in Section 4: Strongly convex case. This appendix provides the full proof of technical results in Section 4.

C.1. The proof of Lemma 10: One-iteration analysis. First, we write down the optimality conditions of the updates of r^{k+1} and $\tilde{x}^{k+1}$ in (27) as

$$\left\{\begin{array}{lcl} 0 &\in& \partial g(r^{k+1}) + \nabla_r\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k),\\ 0 &\in& \partial f(\tilde{x}^{k+1}) + \nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k) + \frac{\tau_k}{\beta_k}(\tilde{x}^{k+1} - \tilde{x}^k). \end{array}\right. \tag{73}$$

Let us denote $\breve{x}^{k+1} := (1-\tau_k)x^k + \tau_k\tilde{x}^{k+1}$. Then, by convexity of g and strong convexity of f with a strong convexity parameter $\mu_f > 0$, we can derive

$$\left\{\begin{array}{lcl} g(r^{k+1}) &\le& (1-\tau_k)g(r^k) + \tau_k g(r) + \langle \nabla g(r^{k+1}), r^{k+1} - (1-\tau_k)r^k - \tau_k r\rangle,\\ f(\breve{x}^{k+1}) &\le& (1-\tau_k)f(x^k) + \tau_k f(x) + \tau_k\langle \nabla f(\tilde{x}^{k+1}), \tilde{x}^{k+1} - x\rangle\\ && - \frac{\tau_k\mu_f}{2}\|\tilde{x}^{k+1} - x\|^2 - \frac{\tau_k(1-\tau_k)\mu_f}{2}\|\tilde{x}^{k+1} - x^k\|^2, \end{array}\right. \tag{74}$$

where $\nabla g(r^{k+1}) \in \partial g(r^{k+1})$ and $\nabla f(x^{k+1}) \in \partial f(x^{k+1})$ are subgradients.
Next, using Lemma 17(a) three times by setting (x', r') as (x^k, r^k), (x^{k+1}, r^{k+1}), and (x, r), and (x, r) as $(\hat{x}^k, r^{k+1})$, respectively, similar to (47), we can eventually derive

$$\begin{array}{ll} \phi_{\rho_k}(x^{k+1}, r^{k+1}, \tilde{y}^k) = (1-\tau_k)\phi_{\rho_k}(x^k, r^k, \tilde{y}^k) + \tau_k\phi_{\rho_k}(x, r, \tilde{y}^k)\\ \quad + \langle \nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k),\ x^{k+1} - (1-\tau_k)x^k - \tau_k x\rangle\\ \quad + \langle \nabla_r\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k),\ r^{k+1} - (1-\tau_k)r^k - \tau_k r\rangle + \frac{\rho_k}{2}\|K(x^{k+1} - \hat{x}^k)\|^2\\ \quad \Big\{ - \frac{(1-\tau_k)\rho_k}{2}\|K\hat{x}^k - r^{k+1} - (Kx^k - r^k)\|^2 - \frac{\tau_k\rho_k}{2}\|K\hat{x}^k - r^{k+1} - (Kx - r)\|^2\Big\}_{[\mathcal{T}_1]}, \end{array} \tag{75}$$

where we define the last line as $\mathcal{T}_1$.
Combining (73), (74), and (75), we get

$$\begin{array}{ll} \mathcal{L}_{\rho_k}(x^{k+1},\ r^{k+1}, \tilde{y}^k) = f(x^{k+1}) + g(r^{k+1}) + \phi_{\rho_k}(x^{k+1}, r^{k+1}, \tilde{y}^k)\\ \overset{(73)-(75)}{\le} (1-\tau_k)\left[g(r^k) + \phi_{\rho_k}(x^k, r^k, \tilde{y}^k)\right] + \tau_k\left[g(r) + \phi_{\rho_k}(x, r, \tilde{y}^k)\right] + \mathcal{T}_1\\ \quad + \frac{\rho_k}{2}\|K(x^{k+1} - \hat{x}^k)\|^2 + \langle \nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k),\ x^{k+1} - \hat{x}^k\rangle\\ \quad + f(x^{k+1}) + \langle \nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k), \hat{x}^k - (1-\tau_k)x^k - \tau_k x\rangle \end{array} \Big\} =: \mathcal{T}_2. \tag{76}$$

To estimate $\mathcal{T}_2$, notice that by the optimality condition of the x^{k+1}-update in (27) and the μ_f-strong convexity of f, we can show that

$$\begin{array}{lcl} f(x^{k+1}) &+& \frac{\rho_k\|K\|^2}{2}\|x^{k+1} - \hat{x}^k\|^2 + \langle \nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k), x^{k+1} - \hat{x}^k\rangle\\ &\le& f(\breve{x}^{k+1}) + \frac{\rho_k\|K\|^2}{2}\|\breve{x}^{k+1} - \hat{x}^k\|^2 + \langle \nabla_x\phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k), \breve{x}^{k+1} - \hat{x}^k\rangle\\ &&- \frac{\rho_k\|K\|^2 + \mu_f}{2}\|\breve{x}^{k+1} - x^{k+1}\|^2. \end{array} \tag{77}$$

Using the above inequality as well as (73) and (74), we can upper bound

$$
\begin{aligned}
\mathcal{T}_2 \overset{(77)}{\leq}\ & f(\breve{x}^{k+1}) + \tfrac{\rho_k\|K\|^2}{2}\|\breve{x}^{k+1}-\hat{x}^k\|^2 - \tfrac{\rho_k\|K\|^2+\mu_f}{2}\|\breve{x}^{k+1}-x^{k+1}\|^2 \\
& + \langle \nabla_x \phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k), \breve{x}^{k+1} - (1-\tau_k)x^k - \tau_k x\rangle \\
\overset{(73)-(74)}{\leq}\ & (1-\tau_k) f(x^k) + \tau_k f(x) + \tfrac{\tau_k^2}{\beta_k}\langle \tilde{x}^{k+1} - \tilde{x}^k, x - \tilde{x}^{k+1}\rangle \\
& + \langle \nabla_x \phi_{\rho_k}(\hat{x}^k, r^{k+1}, \tilde{y}^k), \breve{x}^{k+1} - (1-\tau_k)x^k - \tau_k \tilde{x}^{k+1}\rangle \\
& - \tfrac{\rho_k\|K\|^2}{2}\|\breve{x}^{k+1}-\hat{x}^k\|^2 - \tfrac{(\rho_k\|K\|^2+\mu_f)}{2}\|\breve{x}^{k+1}-x^{k+1}\|^2 \\
& - \tfrac{\tau_k\mu_f}{2}\|\tilde{x}^{k+1}-x\|^2 - \tfrac{\tau_k(1-\tau_k)\mu_f}{2}\|\tilde{x}^{k+1}-x^k\|^2 \\
=\ & (1-\tau_k) f(x^k) + \tau_k f(x) + \tfrac{\tau_k^2}{2\beta_k}\|\tilde{x}^k - x\|^2 - \tfrac{\tau_k(\tau_k+\beta_k\mu_f)}{2\beta_k}\|\tilde{x}^{k+1}-x\|^2 \\
& - \tfrac{(1-\rho_k\beta_k\|K\|^2)}{2\beta_k}\|\breve{x}^{k+1}-\hat{x}^k\|^2 - \tfrac{(\rho_k\|K\|^2+\mu_f)}{2}\|\breve{x}^{k+1}-x^{k+1}\|^2 \\
& - \tfrac{\tau_k(1-\tau_k)\mu_f}{2}\|\tilde{x}^{k+1}-x^k\|^2,
\end{aligned} \tag{78}
$$

where we have used $\breve{x}^{k+1} - \hat{x}^k = \tau_k(\tilde{x}^{k+1} - \tilde{x}^k)$. Substituting (78) into (76), we have

$$
\begin{aligned}
\mathcal{L}_{\rho_k}(x^{k+1}, r^{k+1}, \tilde{y}^k) \leq\ & (1-\tau_k)\mathcal{L}_{\rho_k}(x^k, r^k, \tilde{y}^k) + \tau_k \mathcal{L}_{\rho_k}(x, r, \tilde{y}^k) + \mathcal{T}_1 \\
& + \tfrac{\tau_k^2}{2\beta_k}\|\tilde{x}^k - x\|^2 - \tfrac{\tau_k(\tau_k+\beta_k\mu_f)}{2\beta_k}\|\tilde{x}^{k+1}-x\|^2 - \tfrac{\tau_k(1-\tau_k)\mu_f}{2\beta_k}\|\tilde{x}^{k+1}-x^k\|^2 \\
& - \tfrac{(1-\rho_k\beta_k\|K\|^2)}{2\beta_k}\|\breve{x}^{k+1}-\hat{x}^k\|^2 - \tfrac{(\rho_k\|K\|^2+\mu_f)}{2}\|\breve{x}^{k+1}-x^{k+1}\|^2.
\end{aligned} \tag{79}
$$

By the definition of $\tilde{y}^{k+1}$ from (27) and that of $\bar{y}^{k+1}$, the equations (52), (54), and (55) still hold. Substituting them into (79), and using the expression of $\mathcal{T}_1$, we get

$$
\begin{aligned}
\mathcal{L}_{\rho_k}(x^{k+1},\ r^{k+1}, y) - \mathcal{L}(x, r, \bar{y}^{k+1}) \leq\ & (1-\tau_k)\left[\mathcal{L}_{\rho_{k-1}}(x^k, r^k, y) - \mathcal{L}(x, r, \bar{y}^k)\right] \\
& \left.\begin{aligned} & + \tfrac{(1-\tau_k)}{2}(\rho_k - \rho_{k-1})\|Kx^k - r^k\|^2 - \tfrac{\tau_k\rho_k}{2}\|K\hat{x}^k - r^{k+1}\|^2 \\ & - \tfrac{(1-\tau_k)\rho_k}{2}\|K\hat{x}^k - r^{k+1} - (Kx^k - r^k)\|^2 + \tfrac{1}{2\eta_k}\|\tilde{y}^{k+1} - \tilde{y}^k\|^2 \end{aligned}\right\} =: \mathcal{T}_3 \\
& - \tfrac{(1-\rho_k\beta_k\|K\|^2)}{2\beta_k}\|\breve{x}^{k+1}-\hat{x}^k\|^2 - \tfrac{(\rho_k\|K\|^2+\mu_f)}{2}\|\breve{x}^{k+1}-x^{k+1}\|^2 \quad =: \mathcal{T}_4 \\
& + \tfrac{\tau_k^2}{2\beta_k}\|\tilde{x}^k - x\|^2 - \tfrac{\tau_k(\tau_k+\beta_k\mu_f)}{2\beta_k}\|\tilde{x}^{k+1}-x\|^2 - \tfrac{\tau_k(1-\tau_k)\mu_f}{2\beta_k}\|\tilde{x}^{k+1}-x^k\|^2 \\
& + \tfrac{1}{2\eta_k}\left[\|\tilde{y}^k - y\|^2 - \|\tilde{y}^{k+1} - y\|^2\right].
\end{aligned} \tag{80}
$$

Since $\rho_k > \eta_k$, using the same lines as (57)-(58) in the proof of Lemma 1, we have

$$
\mathcal{T}_3 \leq \frac{\rho_k\eta_k}{2(\rho_k-\eta_k)}\|K(x^{k+1}-\hat{x}^k)\|^2 - \frac{(1-\tau_k)}{2}\left[\rho_{k-1} - (1-\tau_k)\rho_k\right]\|Kx^k - r^k\|^2. \tag{81}
$$

Applying Lemma 17(b) on $\mathcal{T}_4$ with $\alpha_1 := \frac{1-\rho_k\beta_k\|K\|^2}{2\beta_k}$ and $\alpha_2 := \frac{\rho_k\|K\|^2+\mu_f}{2}$, and noting that $\rho_k\beta_k\|K\|^2 < 1$, we can further show that

$$
\mathcal{T}_4 \leq -\frac{\alpha_1\alpha_2}{\alpha_1+\alpha_2}\|x^{k+1}-\hat{x}^k\|^2 \leq -\frac{\rho_k}{2}(1-\rho_k\beta_k\|K\|^2)\|K(x^{k+1}-\hat{x}^k)\|^2. \tag{82}
$$

Substituting (81) and (82) into (80), we finally arrive at (28). ∎

C.2. The proof of Theorem 11: $\mathcal{O}\left(1/k^2\right)$ convergence rates. By the parameter update rule (29), we can easily verify that

$$
\begin{cases}
\frac{\tau_k^2}{2\beta_k} \leq \frac{(1-\tau_k)\tau_{k-1}(\tau_{k-1}+\beta_{k-1}\mu_f)}{2\beta_{k-1}}, \qquad \frac{1}{2\eta_k} = \frac{1-\tau_k}{2\eta_{k-1}}, \\
1 - \rho_k\beta_k\|K\|^2 - \frac{\eta_k}{\rho_k-\eta_k} = \frac{(1-\tau_k)}{2}\left[\rho_{k-1} - (1-\tau_k)\rho_k\right] = 0.
\end{cases}
$$

Applying these conditions to Lemma 10, we can simplify (28) as

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},r^{k+1},y) - \mathcal{L}(x,r,\bar{y}^{k+1}) + \tfrac{\tau_k(\tau_k+\beta_k\mu_f)}{2\beta_k}\|\tilde{x}^{k+1}-x\|^2 + \tfrac{1}{2\eta_k}\|\tilde{y}^{k+1}-y\|^2 \\ \le (1-\tau_k)\Big[\mathcal{L}_{\rho_{k-1}}(x^k,r^k,y) - \mathcal{L}(x,r,\bar{y}^k) \\ + \tfrac{\tau_{k-1}(\tau_{k-1}+\beta_{k-1}\mu_f)}{2\beta_{k-1}}\|\tilde{x}^k-x\|^2 + \tfrac{1}{2\eta_{k-1}}\|\tilde{y}^k-y\|^2\Big].\end{aligned}$$

By induction, the above inequality implies

$$\begin{aligned}\mathcal{L}_{\rho_{k-1}}(x^k,r^k,y) - \mathcal{L}(x,r,\bar{y}^k) &\le \left[\prod_{i=1}^{k-1}(1-\tau_i)\right]\Big[(1-\tau_0)\left(\mathcal{L}_{\rho_{-1}}(x^0,r^0,y) - \mathcal{L}(x,r,\bar{y}^0)\right) \\ &\quad + \tfrac{\tau_0^2}{2\beta_0}\|\tilde{x}^0-x\|^2 + \tfrac{1}{2\eta_0}\|\tilde{y}^0-y\|^2\Big] \\ &= \tau_{k-1}^2\left[\tfrac{\tau_0^2}{2\beta_0}\|\tilde{x}^0-x\|^2 + \tfrac{1}{2\eta_0}\|\tilde{y}^0-y\|^2\right] \\ &\le \tfrac{4}{(k+1)^2}\left[\tfrac{\rho_0\|K\|^2\|x^0-x\|^2}{2\Gamma} + \tfrac{\|y^0-y\|^2}{2(1-\gamma)\rho_0}\right],\end{aligned}$$

where we have used $1-\tau_k = \tau_k^2/\tau_{k-1}^2$, $\tau_k \le 2/(k+2)$, and the parameter initialization (30). The remaining conclusions of Theorem 11 follow the same lines as the proof of Theorem 2. Thus we omit the details here. ☐

C.3. The proof of Theorem 13: $\mathcal{O}\left(1/k^2\right)$ and $o\big(1/(k^2\sqrt{\log k})\big)$ convergence rates. Similar to the proof of Theorem 5, we abbreviate $a_k^2 := \frac{\rho_0}{2}\|Kx^k - r^k\|^2$, $b_k^2 := \frac{\rho_0\|K\|^2}{2\Gamma}\|\tilde{x}^k - x^\star\|^2$, $d_k^2 := \frac{1}{2(1-\gamma)\rho_0}\|\tilde{y}^k - y^\star\|^2$, and $\widetilde{\mathcal{G}}_k := \mathcal{L}(x^k,r^k,y^\star) - \mathcal{L}(x^\star,r^\star,\bar{y}^k) \ge 0$. We can rewrite (28) in Lemma 10 as follows:

$$\begin{aligned}\widetilde{\mathcal{G}}_{k+1} + \left(\tfrac{k+c}{c}\right)^2 a_{k+1}^2 &\le \tfrac{k}{k+c}\left[\widetilde{\mathcal{G}}_k + \left(\tfrac{k+c-1}{c}\right)^2 a_k^2\right] + (b_k^2 - b_{k+1}^2) - \tfrac{c\Gamma\mu_f}{(k+c)\rho_0\|K\|^2}b_{k+1}^2 \\ &\quad + \left(\tfrac{c}{k+c}\right)^2(d_k^2 - d_{k+1}^2) - \tfrac{k}{c^2(k+c)}\left[(k+c-1)^2 - k(k+c)\right]a_k^2.\end{aligned}$$

Multiplying both sides of this estimate by $(k+c)^2$ and rearranging the result, we get

$$\begin{aligned}R_{k+1}^2 &:= (k+c)^2\widetilde{\mathcal{G}}_{k+1} + \tfrac{(k+c)^4}{c^2}a_{k+1}^2 + \left[(k+c)^2 + \tfrac{c(k+c)\Gamma\mu_f}{\rho_0\|K\|^2}\right]b_{k+1}^2 + c^2d_{k+1}^2 \\ &\le k(k+c)\widetilde{\mathcal{G}}_k + \tfrac{k^2(k+c)^2}{c^2}a_k^2 + (k+c)^2b_k^2 + c^2d_k^2.\end{aligned} \tag{83}$$

If $c > 2$ and $0 < \rho_0 \le \frac{c(c-1)\Gamma\mu_f}{(2c-1)\|K\|^2}$, then the above right-hand-side is bounded by R_k^2, and we have

$$\begin{aligned}(c-2)\left[(k+c-1)\widetilde{\mathcal{G}}_k + \tfrac{(k+c-1)^3}{c^2}a_k^2\right] &\le \left[(c-2)k + (c-1)^2\right]\widetilde{\mathcal{G}}_k \\ &\quad + \tfrac{1}{c^2}\left[(k+c-1)^4 - k^2(k+c)^2\right]a_k^2 \\ &\le \left[(k+c-1)^2\widetilde{\mathcal{G}}_k + \tfrac{(k+c-1)^4}{c^2}a_k^2 + \left((k+c-1)^2 + \tfrac{c(k+c-1)\Gamma\mu_f}{\rho_0\|K\|^2}\right)b_k^2 + c^2d_k^2\right] \\ &\quad - \left[(k+c)^2\widetilde{\mathcal{G}}_{k+1} + \tfrac{(k+c)^4}{c^2}a_{k+1}^2 + \left((k+c)^2 + \tfrac{c(k+c)\Gamma\mu_f}{\rho_0\|K\|^2}\right)b_{k+1}^2 + c^2d_{k+1}^2\right] \\ &= R_k^2 - R_{k+1}^2.\end{aligned} \tag{84}$$

By induction and the definitions of $\widetilde{\mathcal{G}}_k$ and R_k^2, we can show that

$$\widetilde{\mathcal{L}}(x^k,y^\star) - F^\star \le \mathcal{L}(x^k,r^k,y^\star) - F^\star = \widetilde{\mathcal{G}}_k \le \frac{R_k^2}{(k+c-1)^2} \le \frac{R_0^2}{(k+c-1)^2}. \tag{85}$$

By the initialization of Algorithm 2, we have proved the first assertion of (36).

Summing up (84) from $i := 0$ to $i := k$, we get

$$(c-2)\sum_{i=0}^{k}\left[(i+c-1)\widetilde{G}_i + \tfrac{(i+c-1)^3}{c^2}a_i^2\right] \le R_0^2 - R_{k+1}^2 \le R_0^2 < +\infty.$$

Since $c-2>0$ and $i \le i+c-1$, applying Lemma 17(c) to the last expression yields

$$\liminf_{k\to\infty} k^2\log k\left[\widetilde{\mathcal{G}}_k + \frac{k^2a_k^2}{c^2}\right] = 0.$$

Since $0 \le \widetilde{\mathcal{L}}(x^k, y^\star) - F^\star \le \mathcal{L}(x^k, r^k, y^\star) - F^\star = \widetilde{\mathcal{G}}_k$, we can easily obtain the second assertion of (36) from this limit.

The remaining statements of Theorem 13 can be proved in a similar manner as of Theorem 5, but by applying Lemma 17(d, part (ii)) to prove the limit in (37). Thus we omit the details here. ∎

C.4. The proof of Corollary 15: Constrained problems. The augmented Lagrangian associated with problem (38) is $\mathcal{L}_\rho(x,w,y) := f(x) + \psi(w) + \langle y, Kx + Bw - b\rangle + \frac{\rho}{2}\|Kx + Bw - b\|^2$. Let $\tilde{w}^{k+1} := \frac{1}{\tau_k}\left[w^{k+1} - (1-\tau_k)w^k\right]$. The optimality condition of the w^{k+1}-update in (39) and the convexity of ψ imply for $w \in \mathbb{R}^q$ that

$$\begin{aligned}\psi(w^{k+1}) \le\ & (1-\tau_k)\psi(w^k) + \tau_k\psi(w)\\ &+ \tau_k\langle B^\top[\tilde{y}^k + \rho_k(K\hat{x}^k + Bw^{k+1} - b)] + \nu_0(w^{k+1} - \hat{w}^k), w - \tilde{w}^{k+1}\rangle.\end{aligned}$$

Using this estimate, we follow the same lines as the proof of Lemma 10 to derive

$$\begin{aligned}&\mathcal{L}_{\rho_k}(x^{k+1}, w^{k+1}, y) - \mathcal{L}(x, w, \bar{y}^{k+1}) \le (1-\tau_k)\left[\mathcal{L}_{\rho_{k-1}}(x^k, w^k, y) - \mathcal{L}(x, w, \bar{y}^k)\right]\\ &\quad + \tfrac{\tau_k^2}{2\beta_k}\|\tilde{x} - x\|^2 - \tfrac{\tau_k(\tau_k + \beta_k\mu_f)}{2\beta_k}\|\tilde{x}^{k+1} - x\|^2 + \tfrac{\tau_k^2\nu_0}{2}\left[\|\tilde{w}^k - w\|^2 - \|\tilde{w}^{k+1} - w\|^2\right]\\ &\quad + \tfrac{1}{2\eta_k}\left[\|\tilde{y}^k - y\|^2 - \|\tilde{y}^{k+1} - y\|^2\right] - \tfrac{(1-\tau_k)}{2}\left[\rho_{k-1} - (1-\tau_k)\rho_k\right]\|Kx^k + Bw^k - b\|^2\\ &\quad - \tfrac{\rho_k}{2}\left(1 - \rho_k\beta_k\|K\|^2 - \tfrac{\eta_k}{\rho_k - \eta_k}\right)\|K(x^{k+1} - \hat{x}^k)\|^2 - \tfrac{\nu_k}{2}\|w^{k+1} - \hat{w}^k\|^2.\end{aligned} \tag{86}$$

Plugging the parameter updates (29) and (30) into (86), we can derive (40) following the same arguments as in the proof of Corollary 7.

If we plug the parameter updates (29) and (31) into (86), then we can derive (41) following the same arguments as in the proof of Theorem 13. We omit the details. ∎

REFERENCES

[1] H. Attouch and J. Peypouquet, *The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $\mathcal{O}(1/k^2)$*, SIAM J. Optim., 26 (2016), pp. 1824–1834.

[2] H. H. Bauschke and P. Combettes, *Convex analysis and monotone operators theory in Hilbert spaces*, Springer-Verlag, 2nd ed., 2017.

[3] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sciences, 2 (2009), pp. 183–202.

[4] R. Bot, E. Csetnek, and A. Heinrich, *A primal-dual splitting algorithm for finding zeros of sums of maximally monotone operators*, SIAM J. Optim., 23 (2013), pp. 2011–2036.

[5] R. Boţ, E. Csetnek, A. Heinrich, and C. Hendrich, *On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems*, Math. Program., 150 (2015), pp. 251–279.

[6] R. I. Boţ and C. Hendrich, *Convergence analysis for a primal-dual monotone+ skew splitting algorithm with applications to total variation minimization*, Journal of mathematical imaging and vision, 49 (2014), pp. 551–568.
[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning, 3 (2011), pp. 1–122.
[8] L. Briceno-Arias and P. Combettes, *A monotone + skew splitting model for composite monotone inclusions in duality*, SIAM J. Optim., 21 (2011), pp. 1230–1250.
[9] A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schönlieb, *Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications*, SIAM J. Optim., 28 (2018), pp. 2783–2808.
[10] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vis., 40 (2011), pp. 120–145.
[11] A. Chambolle and T. Pock, *An introduction to continuous optimization for imaging*, Acta Numerica, 25 (2016), pp. 161–319.
[12] A. Chambolle and T. Pock, *On the ergodic convergence rates of a first-order primal–dual algorithm*, Math. Program., 159 (2016), pp. 253–287.
[13] P. Chen, J. Huang, and X. Zhang, *A primal-dual fixed point algorithm for minimization of the sum of three convex separable functions*, Fixed Point Theory and Applications, 2016 (2016), p. 54.
[14] Y. Chen, G. Lan, and Y. Ouyang, *Optimal primal-dual methods for a class of saddle-point problems*, SIAM J. Optim., 24 (2014), pp. 1779–1814.
[15] P. Combettes and J.-C. Pesquet, *Signal recovery by proximal forward-backward splitting*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer-Verlag, 2011, pp. 185–212.
[16] P. L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-point algorithms for inverse problems in science and engineering, Springer, 2011, pp. 185–212.
[17] P. L. Combettes and J.-C. Pesquet, *Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators*, Set-Valued Var. Anal., 20 (2012), pp. 307–330.
[18] L. Condat, *A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms*, J. Optim. Theory Appl., 158 (2013), pp. 460–479.
[19] D. Davis, *Convergence rate analysis of primal-dual splitting schemes*, SIAM J. Optim., 25 (2015), pp. 1912–1943.
[20] D. Davis, *Convergence rate analysis of the forward-Douglas-Rachford splitting scheme*, SIAM J. Optim., 25 (2015), pp. 1760–1786.
[21] D. Davis and W. Yin, *Convergence rate analysis of several splitting schemes*, in Splitting Methods in Communication, Imaging, Science, and Engineering, R. Glowinski, S. J. Osher, and W. Yin, eds., Springer, 2016, pp. 115–163.
[22] D. Davis and W. Yin, *Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions*, Math. Oper. Res., 42 (2017), pp. 577–896.
[23] D. Davis and W. Yin, *A three-operator splitting scheme and its optimization applications*, Set-valued and Variational Analysis, 25 (2017), pp. 829–858.
[24] C. Dünner, S. Forte, M. Takáč, and M. Jaggi, *Primal-dual rates and certificates*, Proc. of the 33rd International Conference on Machine Learning (ICML), (2016).
[25] J. Eckstein and D. Bertsekas, *On the Douglas - Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Program., 55 (1992), pp. 293–318.
[26] 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 (2010), pp. 1015–1046.
[27] J. E. Esser, *Primal-dual algorithm for convex models and applications to image restoration, registration and nonlocal inpainting*, PhD Thesis, University of California, Los Angeles, Los Angeles, USA, 2010.
[28] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*, vol. 1-2, Springer-Verlag, 2003.
[29] R. Glowinski, S. Osher, and W. Yin, *Splitting Methods in Communication, Imaging, Science, and Engineering*, Springer, 2017.
[30] T. Goldstein, M. Li, and X. Yuan, *Adaptive primal-dual splitting methods for statistical learning and image processing*, in Advances in Neural Information Processing Systems, 2015, pp. 2080–2088.
[31] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk, *Fast Alternating Direction Optimization Methods*, SIAM J. Imaging Sci., 7 (2012), pp. 1588–1623.

[32] B. He and X. Yuan, *Convergence analysis of primal-dual algorithms for saddle-point problem: from contraction perspective*, SIAM J. Imaging Sci., 5 (2012), pp. 119–149.
[33] Y. He and R.-D. Monteiro, *An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems*, SIAM J. Optim., 26 (2016), pp. 29–56.
[34] H. Li and Z. Lin, *Accelerated Alternating Direction Method of Multipliers: an Optimal $\mathcal{O}(1/k)$ Nonergodic Analysis*, Journal of Scientific Computing, (2016), pp. 1–29.
[35] J. Liang, J. Fadili, and G. Peyré, *Local convergence properties of Douglas–Rachford and alternating direction method of multipliers*, J. Optim. Theory Appl., 172 (2017), pp. 874–913.
[36] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Num. Anal., 16 (1979), pp. 964–979.
[37] Y. Malitsky and T. Pock, *A first-order primal-dual algorithm with linesearch*, SIAM J. Optim., 28 (2016), pp. 411–432.
[38] R. Monteiro and B. Svaiter, *On the complexity of the hybrid proximal extragradient method for the interates and the ergodic mean*, SIAM J. Optim., 20 (2010), pp. 2755–2787.
[39] R. Monteiro and B. Svaiter, *Complexity of variants of Tseng's modified F-B splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems*, SIAM J. Optim., 21 (2011), pp. 1688–1720.
[40] R. Monteiro and B. Svaiter, *Iteration-complexity of block-decomposition algorithms and the alternating minimization augmented Lagrangian method*, SIAM J. Optim., 23 (2013), pp. 475–507.
[41] MOSEK-ApS, *The MOSEK optimization toolbox for MATLAB manual, version 9.0*, 2019, http://docs.mosek.com/9.0/toolbox/index.html.
[42] I. Necoara and A. Patrascu, *Iteration complexity analysis of dual first order methods for convex programming*, Optim. Method Softw., 31 (2016), pp. 645–678.
[43] I. Necoara, A. Patrascu, and F. Glineur, *Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming*, Optim. Method Softw., 34 (2019), pp. 305–335.
[44] A. Nemirovskii, *Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM J. Op, 15 (2004), pp. 229–251.
[45] Y. Nesterov, *A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$*, Doklady AN SSSR, 269 (1983), pp. 543–547. Translated as Soviet Math. Dokl.
[46] Y. Nesterov, *Excessive gap technique in nonsmooth convex minimization*, SIAM J. Optim., 16 (2005), pp. 235–249.
[47] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.
[48] D. O'Connor and L. Vandenberghe, *Primal-dual decomposition by operator splitting and applications to image deblurring*, SIAM J. Imaging Sci., 7 (2014), pp. 1724–1754.
[49] D. O'Connor and L. Vandenberghe, *On the equivalence of the primal-dual hybrid gradient method and Douglas-Rachford splitting*, Math. Program., (2018), pp. 1–24.
[50] Y. Ouyang, Y. Chen, G. Lan, and E. J. Pasiliao, *An accelerated linearized alternating direction method of multiplier*, SIAM J. Imaging Sci., 8 (2015), pp. 644–681.
[51] T. Pock, D. Cremers, H. Bischof, and A. Chambolle, *An algorithm for minimizing the Mumford-Shah functional*, in 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 1133–1140.
[52] J. E. Spingarn, *Partial inverse of a monotone operator*, Applied mathematics and optimization, 10 (1983), pp. 247–265.
[53] K. H. L. Thi, R. Zhao, and W. B. Haskell, *An inexact primal-dual smoothing framework for large-scale non-bilinear saddle point problems*, arXiv preprint arXiv:1711.03669, (2017).
[54] Q. Tran-Dinh, *Proximal Alternating Penalty Algorithms for Constrained Convex Optimization*, Comput. Optim. Appl., 72 (2019), pp. 1–43.
[55] Q. Tran-Dinh, A. Alacaoglu, O. Fercoq, and V. Cevher, *An Adaptive Primal-Dual Framework for Nonsmooth Convex Minimization*, Math. Program. Compt. (online first), (2019), pp. 1–39.
[56] Q. Tran-Dinh, O. Fercoq, and V. Cevher, *A smooth primal-dual optimization framework for nonsmooth composite convex minimization*, SIAM J. Optim., 28 (2018), pp. 96–134.
[57] Q. Tran-Dinh, C. Savorgnan, and M. Diehl, *Combining Lagrangian decomposition and excessive gap smoothing technique for solving large-scale separable convex optimization problems*, Compt. Optim. Appl., 55 (2013), pp. 75–111.
[58] P. Tseng, *Further applications of a splitting algorithm to decomposition in variational inequal-*

ities and convex programming, Math. Program., 48 (1990), pp. 249–263.
[59] P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization*, Submitted to SIAM J. Optim, (2008).
[60] T. Valkonen, *Inertial, corrected, primal–dual proximal splitting*, SIAM J. Optim., 30 (2020), pp. 1391–1420.
[61] C. B. Vu, *A splitting algorithm for dual monotone inclusions involving co-coercive operators*, Advances in Computational Mathematics, 38 (2013), pp. 667–681.
[62] B. E. Woodworth and N. Srebro, *Tight complexity bounds for optimizing composite objectives*, in Advances in neural information processing systems (NIPS), 2016, pp. 3639–3647.
[63] Y. Xu, *Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming*, SIAM J. Optim., 27 (2017), pp. 1459–1484.
[64] M. Yan, *A new primal–dual algorithm for minimizing the sum of three functions with a linear operator*, Journal of Scientific Computing, (2018), pp. 1–20.
[65] H. E. Yazdandoost and N. S. Aybat, *A primal-dual algorithm for general convex-concave saddle point problems*, arXiv preprint arXiv:1803.01401, (2018).
[66] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, *Bregman iterative algorithms for \ell_1-minimization with applications to compressed sensing*, SIAM J. Imaging Sci., 1 (2008), pp. 143–168.
[67] X. Zhang, M. Burger, and S. Osher, *A unified primal-dual algorithm framework based on bregman iteration*, J. Sci. Comput., 46 (2011), pp. 20–46.
[68] M. Zhu and T. Chan, *An efficient primal-dual hybrid gradient algorithm for total variation image restoration*, UCLA CAM technical report, 08–34 (2008).