Accelerated Primal-Dual Algorithms for a Class of Convex-Concave Saddle-Point Problems with Non-Bilinear Coupling Term

Yuzixuan Zhu · Deyi Liu · Quoc Tran-Dinh*

Received: date / Accepted: date

Abstract We develop two new primal-dual algorithms to solve a class of convex-concave saddle-point problems involving non-bilinear coupling function, which covers many existing and brand-new applications as special cases. Our approach relies on a novel combination of non-convex augmented Lagrangian and Nesterov’s accelerated schemes, which is fundamentally different from existing works. Both algorithms are single-loop and only require one or at most two proximal operators of the objective function, one gradient of the coupling function per iteration. They do not require to solve any complex subproblem as in standard augmented Lagrangian or penalty approaches. When the objective function is merely convex, our first algorithm can achieve $O(1/k)$ convergence rates through three different criteria (primal objective residual, dual objective residual, and primal-dual gap), on either the ergodic sequence or the non-ergodic sequence. This rate can potentially be even faster than $O(1/k)$ on non-ergodic primal objective residual using a new parameter update rule. If the objective function is strongly convex, our second algorithm can boosts these convergence rates to no slower than $O(1/k^2)$. To the best of our knowledge, these are the first algorithms that achieve such fast convergence rates on non-ergodic sequences for non-bilinear convex-concave saddle-point problems. As a by-product, we specify our results to handle general conic convex problems. We test our algorithms on QCQP and a convex-concave game example to verify their performance as well as to compare them with existing methods.

Keywords Non-bilinear convex-concave saddle-point · primal-dual algorithms · optimal convergence rates · Nesterov’s accelerated schemes · cone constrained convex optimization.

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

1 Introduction
The goal of this paper is to develop novel primal-dual algorithms to solve the following convex-concave saddle-point problems involving non-bilinear coupling function:

$$\min_{x \in \mathbb{R}^p} \max_{y \in \mathbb{R}^m} \{ \tilde{L}(x, y) := F(x) + \Phi(x, y) - H^*(y) \}, \quad (\text{SP})$$
where \( F : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\} \) and \( H : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\} \) are proper, closed, and convex, but not necessarily smooth, \( H^* \) is the Fenchel conjugate of \( H \), and \( \Phi : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R} \) is a coupling function that is continuous, convex in \( x \) for all \( y \in \text{dom}(H^*) \), and linear in \( y \) for all \( x \in \text{dom}(F) \). Furthermore, \( F \) and \( \Phi \) satisfy the following conditions:

(i) The function \( F \) is \( F(x) := f(x) + h(x) \), where \( f : \mathbb{R}^p \to \mathbb{R} \) is \( L_f \)-smooth and convex, and \( h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\} \) is proper, closed, and convex, but not necessarily smooth.

(ii) The coupling function \( \Phi \) has specific form as \( \Phi(x, y) := \langle g(x), y \rangle \), the inner product between \( g(x) \) and \( y \), where \( g : \mathbb{R}^p \to \mathbb{R}^m \), and is convex in \( x \) for any \( y \in \text{dom}(H^*) \).

Under this assumption on \( \Phi \), we can formulate (SP) into the following primal composite convex minimization problem:

\[
\mathcal{P}^* := \min_{x \in \mathbb{R}^p} \left\{ \mathcal{P}(x) := F(x) + \max_{y \in \mathbb{R}^m} \{ \Phi(x, y) - H^*(y) \} \right\} \equiv F(x) + H(g(x)).
\]  

The corresponding dual problem is also convex and can be written as

\[
\mathcal{D}^* := \max_{y \in \mathbb{R}^m} \left\{ \mathcal{D}(y) := \min_{x \in \mathbb{R}^p} \{ F(x) + \Phi(x, y) \} - H^*(y) \right\}.
\]

By the convexity of \( \Phi(x, y) \) in \( x \) for any \( y \in \text{dom}(H^*) \), \( H \circ g \) is convex in \( x \). Unlike several existing works such as \([2, 11, 36, 81, 84]\), which only focus on the bilinear function \( \Phi \), the following examples are very common, and can be viewed as special cases of (1):

(a) **Linear equality constraints.** If \( K = \{0\} \), then the constraint in (1) becomes \( g(x) = 0 \).

Clearly, the convexity of \( \Phi(x, y) \) in \( x \) for any \( y \in K^* = \mathbb{R}^m \) implies the affinity of \( g \), i.e., the constraint in (1) reduces to \( Ax - b = 0 \) for some \( A \in \mathbb{R}^{m \times p} \) and \( b \in \mathbb{R}^m \).

(b) **Nonlinear inequality constraints.** If \( K = \mathbb{R}^m_+ \), then the constraint in (1) becomes \( g(x) \leq 0 \). In this case, the convexity of \( \Phi(\cdot, y) \) in \( x \) for any \( y \in K^* = \mathbb{R}^m_+ \) implies the convexity of \( g \), which can be affine or nonlinear.

(c) **Linear matrix constraints.** If \( K = \mathcal{S}^n_+ \), the cone of symmetric positive semidefinite matrices of order \( n \), and \( g(x) := A_0 + \sum_{l=1}^p x_l A_l \) for given symmetric matrices \( A_0, A_1, \cdots, A_p \in \mathcal{S}^n \), then the convexity of \( \Phi(x, y) \) in \( x \) for any \( y \in K^* = \mathcal{S}^n_+ \) requires that \( g \) is \( \mathcal{S}^n_+ \)-convex, and we obtain linear matrix constraint in (1): \( A_0 + \sum_{l=1}^p x_l A_l \leq 0 \).

If in addition \( f \) is linear and \( h = 0 \), then (1) reduces to a semidefinite program.

(d) **Product-of-cones constraints.** If \( K = \{0\}^n \times \mathbb{R}^m_+ \), then (1) becomes

\[
\begin{cases}
\min_{x \in \mathbb{R}^p} \{ F(x) := f(x) + h(x), \} \\
\text{s.t. } Ax = b, \quad g(x) \leq 0,
\end{cases}
\]

where \( A \in \mathbb{R}^{n \times p} \), and \( g : \mathbb{R}^p \to \mathbb{R}^m \) is convex. This setting is very common in classical nonlinear programming literature, see, e.g., \([10, 62]\). \[1\]

\[1\] Recall that a cone is proper if it is closed, convex, solid, and pointed.
Applications. If $\Phi$ is bilinear, then it is well-known that (SP) and its primal form (P) already cover various applications in signal and image processing, compressive sensing, machine learning, and statistics, see, e.g., [2,11,14,23,35]. Here, we only discuss a few representative applications of (SP) where $\Phi$ is non-bilinear.

(1) Kernel matrix learning. In [44], a concrete application of (SP) was proposed to learn a kernel matrix for a support vector machine problem. The underlying coupling function $\Phi$ in this model is linear in $y$ as a kernel matrix, and quadratic in $x$ as model parameters, see [44, formula (20)]. Clearly, this problem can be reformulated as a special case of (SP), and has been used in [94] for numerical illustration. Another related problem is the maximum margin clustering model studied in [88].

(2) Robust convex optimization. Various robust optimization settings rely on the well-known Wald’s max-min formulation, see, e.g., [5]. If the objective term $\Phi$ linearly depends on uncertainty $y$ as $\Phi(x,y) = \langle g(x), y \rangle$, then the robust counterpart of the underlying convex optimization problem can be cast into (SP).

(3) Wasserstein Generative Adversarial Networks. The generative adversarial networks (GANs) problem involving Wasserstein distance studied in [1] can be formulated as a special case of (SP) when the discriminator is linear and the generator is convex w.r.t. their parameters. This model has been widely studied in the machine learning community in recent years, and it is also related to optimal transport as shown in [25]. Other applications in machine learning, distributionally robust optimization, optimal transport, game theory, and image signal processing can be found in the literature, see, e.g., [24, 59, 68, 71, 76]. It is also worth noticing that problem (1) can serve as the subproblems in several non-convex optimization algorithms such as proximal-point, inner approximation, penalty-based, and DC (difference of two convex functions) algorithms, see, e.g., [7, 82].

Limitation of existing work. Methods for solving (SP) or its composite reformulation (P) when $\Phi$ is bilinear are well-developed, see, e.g., [11, 14, 23, 36, 59, 81, 84]. However, when $\Phi$ is no longer bilinear, the methods for solving (SP) remain limited, see [56, 80, 94] and their subsequent references. Existing works have the following limitations:

- Model assumptions. Gradient-based methods such as [56, 94] require $\nabla_x \Phi$ and $\nabla_y \Phi$ to be uniformly Lipschitz continuous on both $x$ and $y$, which unfortunately excludes some important cases, e.g., the cone constrained problem (1), where $\nabla_x \Phi(x,y) = g'(x)^\top y$, which may not be uniformly Lipschitz continuous on $x$ for all $y$ (see Assumption 2). In addition, if $H^*$ is not strongly convex or restricted strongly convex [19, 79], then $H \circ g$ in (P) can be nonsmooth, which creates several challenges for first-order methods.

- High per-iteration complexity and double loops. Several methods, including [48, 93], require double loops, where the inner loop approximately solves the inner problem, e.g., the maximization problem in $y$, and the outer loop handles the minimization problem. This method can be viewed as an inexact first-order scheme to solve the composite problem (P). Hence, the complexity of each outer iteration is often high. In addition, related parameters such as the number of inner iterations are often chosen based on convergence bounds, and may depend on a desired accuracy. For the cone program (1), penalty and augmented Lagrangian approaches (including alternating minimization and alternating direction methods of multipliers (ADMM)) require to solve expensive subproblems [91]. There exist very limited low per-iteration complexity methods, e.g., [56, 80, 94] for general convex-concave saddle-point problems, and [92] for a special case (2).

- Convergence rates and convergence guarantee criteria. Subgradient and mirror descent methods such as [56] often have slow convergence rates compared to gradient
and accelerated gradient-based methods \cite{58}. Hitherto, existing works can only show the best-known convergence rates on the ergodic (or, the averaging) sequences, via, e.g., a gap function (cf. \cite{0}), see, e.g., \cite{56,80,94,92}. It means that the convergence guarantee is based on an average or a weighted average sequence of all the past iterates. In practical implementation, however, researchers often report performance on the non-ergodic (or, the last-iterate) sequence, which may only have asymptotic convergence or suboptimal rate compared to the averaging one. As indicated in \cite{32}, theoretical guarantee of the last iterates can be slower than the averaging ones.

These three major limitations of the existing works motivate us to conduct this research to develop novel primal-dual algorithms, which affirmatively solve the above challenges.

**Our approach.** It is obvious that problem \cite{SP} is much more challenging to solve than its special case with bilinear $\Phi$, especially when the dual domain of $y$ is unbounded. Our approach relies on a novel combination of the following techniques:

- Firstly, we reformulate \cite{SP}, or equivalently \cite{P}, as a constrained optimization problem \cite{10}, which is non-convex in $x$ if $g$ is not affine.
- Secondly, we utilize an augmented Lagrangian function to penalize the constraints of \cite{10}. This function plays a role as a merit function to measure the optimality. Different from existing augmented Lagrangian functions, e.g., \cite{90,91}, ours is non-convex in $x$. Fortunately, it possesses an useful property as shown in Lemma 1, which allows us to still apply convex optimization techniques.
- Thirdly, we apply Nesterov’s accelerated methods \cite{57} to minimize the augmented Lagrangian function, and characterize convergence guarantees to \cite{SP}, \cite{P}, and \cite{D}.
- Finally, we exploit the homotopy strategy in \cite{81,84} to simultaneously update the penalty parameter and step-sizes, making the algorithms converge at the best-known rates.

Compared with standard augmented Lagrangian methods, e.g., in \cite{6,74,92}, our approach is fundamentally different and relies on a non-convex augmented Lagrangian form. We can view this function as a smoothed approximation of the constrained reformulation \cite{10} of \cite{P}, where the smoothness parameter is indeed the penalty parameter \cite{60}.

**Our contributions.** Our contribution can be summarized as follows (The $O$ and $o$ notations will be defined in \cite{3}), accompanied by Table 1.

(a) **New primal-dual algorithm for \cite{SP} with merely convex $F$.** We develop an accelerated primal-dual algorithm, Algorithm 1 based on low per-iteration complexity to solve \cite{SP}, where $F$ and $H$ are general convex and possibly nonsmooth.

(b) **Optimal $O(1/k)$ ergodic and non-ergodic convergence rates when $k = O(p)$.** First, we specify a parameter update rule for Algorithm 1 and establish its $O(1/k)$ ergodic convergence rate. Next, we establish $O(1/k)$ convergence rate on the primal non-ergodic and dual ergodic sequences of Algorithm 1. We call this rate semi-ergodic rate. Unlike existing work, we characterize three different criteria: gap function, primal objective residual, and dual objective residual.

(c) **Faster $\min\{O(1/k), o(1/(k\sqrt{\log k}))\}$ non-ergodic rate when $k > O(p)$.** We modify the parameter update rule of Algorithm 1 to achieve both $O(1/k)$ and $o(1/(k\sqrt{\log k}))$ non-ergodic convergence rate for \cite{P}. To the best of our knowledge, this is the first time that a first-order method for \cite{P} attains such fast convergence rates.

---

In the literature, “ergodic” and “averaging” are used interchangeably; same goes for “non-ergodic” and “last-iterate”. In this paper, we will mainly use “ergodic”, “non-ergodic”, and sometimes “semi-ergodic”.
F. With strongly convex $F$ in (SP), we develop a new accelerated primal-dual algorithm, Algorithm 2. This algorithm essentially has the same per-iteration complexity as Algorithm 1 except for one additional proximal operator of $h$.

(c) **Optimal** $\mathcal{O} (1/k^2)$ **ergodic and non-ergodic convergence rates when** $k = O (p)$.

We first specify a parameter update rule of Algorithm 2 to solve (SP) and prove its $\mathcal{O} (1/k^2)$ **ergodic** convergence rates on three criteria as described in item (b). Next, we establish that Algorithm 2 can achieve $\mathcal{O} (1/k^2)$ convergence rate, on the primal **non-ergodic** sequence and the dual **ergodic** sequence, on the same three criteria.

(f) **Faster** $\min \{ \mathcal{O} (1/k^2), \, g(1/(k^2 \log k)) \}$ **non-ergodic rate when** $k > O (p)$. By modifying the parameter update rule, we boost the convergence rate of Algorithm 2 for solving (P) to both $\mathcal{O} (1/k^2)$ and $g(1/(k^2 \log k))$ in the **non-ergodic** sense.

| Algorithm | Parameter update rule | Convergence criteria | Convergence rate | Theorem |
|-----------|-----------------------|----------------------|-----------------|---------|
| Algorithm 1 | option 1: 29 (30) | $\mathcal{P}(x^k) - \mathcal{D}(y^k), \, g_{X \times Y}(z^k, y^k)$ | $O(1/k)$ | Theorem 1 |
| | option 2: 33 (35) | $\mathcal{P}(x^k) - \mathcal{D}(y^k), \, g_{X \times Y}(z^k, y^k)$ | $O(1/k)$ | Theorem 6 |
| | option 3: 36 (37) | $\mathcal{P}(x^k) - \mathcal{P}^*$ | $\min\{O(1/k), \, g(1/(k\sqrt{\log k}))\}$ | Theorem 2 |
| Algorithm 2 | option 1: 41 (42) | $\mathcal{P}(x^k) - \mathcal{D}(y^k), \, g_{X \times Y}(z^k, y^k)$ | $O(1/k^2)$ | Theorem 4 |
| | option 2: 46 (47) | $\mathcal{P}(x^k) - \mathcal{D}(y^k), \, g_{X \times Y}(z^k, y^k)$ | $O(1/k^2)$ | Theorem 5 |
| | option 3: 49 (50) | $\mathcal{P}(x^k) - \mathcal{P}^*$ | $\min\{O(1/k^2), \, g(1/(k^2 \sqrt{\log k}))\}$ | Theorem 6 |

**Comparison.** Let us briefly compare our contributions with existing methods. Firstly, both Algorithms 1 and 2 and their variants are new and very different from existing methods, see, e.g., 36,80,94, especially, at the Nesterov’s accelerated steps, the dual update in the last step, and the parameter selection. Secondly, we do not require $\nabla_y \Phi$ to be uniformly Lipschitz continuous in $x$ for all $y$ as in 42,90,93,94, where the domain of $y$ in our setting can be unbounded. Thirdly, we do not use double loops as in 42,93, making our methods look much simpler. Fourthly, we only focus on the general convex case, and the strongly convex case of $F$, and ignore the case when both $F$ and $H^*$ are strongly convex since this condition leads to a strong monotonicity of the underlying KKT system of (SP), and linear convergence is often well-known 21,24. Finally, our convergence guarantees are on three different criteria, and in a semi-ergodic sense, which we believe is new. In addition, $\varphi(\cdot)$ convergence rates are also new, and characterize the regime $k > O (p)$ (i.e., when the number of iterations $k$ is sufficiently larger than the problem dimension $p$) as opposed to $O (\cdot)$ rates, which often reflect the regime $k = O (p)$, see, e.g., 58,84.

**Paper structure.** The rest of this paper is organized as follows. Section 2 recalls some basic concepts, and defines our augmented Lagrangian function and characterizes its property. Section 3 develops our first algorithm, Algorithm 1 for solving (SP), and discusses how its three variants lead to different types of convergence guarantees. In Section 4, we develop the second algorithm, Algorithm 2, to handle the strongly convex case and prove its convergence rates. Section 5 specifies our methods to solve cone constrained convex problem (1). Section
[6] provides several numerical examples to verify our theoretical results. Sections [7] provides a brief review on the most related works, and Section [8] is for conclusions. For the clarity of presentation, all technical proofs are deferred to the appendices.

2 Fundamental Assumptions and Mathematical Tools

Let us first recall some basic notation and concepts. Then, we describe our assumptions imposed on (SP). Finally, we reformulate (SP) into a non-convex constrained problem and introduce the associated non-convex augmented Lagrangian function. We also prove a key property of this function, which will be used in the sequel.

2.1 Basic notations and concepts

We work with Euclidean spaces $\mathbb{R}^p$ and $\mathbb{R}^m$ equipped with standard inner product $\langle u, v \rangle := \sum_i u_i v_i$ and norm $\|u\| := \langle u, u \rangle^{1/2}$. For any nonempty, closed, and convex set $\mathcal{X}$ in $\mathbb{R}^p$, $\text{ri}(\mathcal{X})$ denotes the relative interior of $\mathcal{X}$, and $\delta_X$ denotes the indicator of $\mathcal{X}$. If $K$ is a convex cone, then $K^* := \{w \in \mathbb{R}^p \mid \langle w, x \rangle \geq 0, \forall x \in K\}$ denotes its dual cone. For any proper, closed, and convex function $f : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{+\infty\}$, $\text{dom}(f) := \{x \in \mathbb{R}^p \mid f(x) < +\infty\}$ is its (effective) domain, $f^*(y) := \sup_x \{\langle x, y \rangle - f(x)\}$ denotes the Fenchel conjugate of $f$, $\partial f(x) := \{w \in \mathbb{R}^p \mid f(y) - f(x) \geq \langle w, y - x \rangle, \forall y \in \text{dom}(f)\}$ stands for the subdifferential of $f$ at $x$, and $\nabla f$ is the gradient or subgradient of $f$. We also denote $\text{prox}_f(x) := \arg\min_y \{f(y) + \frac{1}{2}\|y - x\|^2\}$ the proximal operator of $f$. If $f$ is the indicator of a convex set $\mathcal{X}$, then $\text{prox}_f$ reduces to the projection $\text{proj}_{\mathcal{X}}$ onto $\mathcal{X}$. For a vector function $g : \mathbb{R}^p \rightarrow \mathbb{R}^m$, we use $g'(\cdot) \in \mathbb{R}^{m \times p}$ to denote its Jacobian.

A function $f : \mathbb{R}^p \rightarrow \mathbb{R}^m$ is called $M_f$-Lipschitz continuous on $\text{dom}(f)$ with a Lipschitz constant $M_f \in [0, +\infty)$ if $|f(x) - f(\hat{x})| \leq M_f \|x - \hat{x}\|$ for all $x, \hat{x} \in \text{dom}(f)$. If $f$ is differentiable on $\text{dom}(f)$ and $\nabla f$ is Lipschitz continuous with a Lipschitz constant $L_f \in [0, +\infty)$, i.e., $\|\nabla f(x) - \nabla f(\hat{x})\| \leq L_f \|x - \hat{x}\|$ for $x, \hat{x} \in \text{dom}(f)$, then we say that $f$ is $L_f$-smooth. If $f(\cdot) - \frac{M_f}{2} \|\cdot\|^2$ is still convex for some $M_f > 0$, then we say that $f$ is $M_f$-strongly convex with a strong convexity parameter $M_f$. Clearly, if $M_f = 0$, then $f$ is merely convex.

For nonnegative sequence $\{u_k\}$ and positive sequence $\{v_k\}$, we recall the standard $O(\cdot)$ notation and define a new one $o(\cdot)$ as

$$
\begin{cases}
    u_k = O(v_k), & \text{if } \limsup_{k \to \infty}(u_k/v_k) < \infty; \\
    u_k = o(v_k), & \text{if } \liminf_{k \to \infty}(u_k/v_k) = 0.
\end{cases}
$$

Finally, $\mathbb{R}_+$ and $\mathbb{R}_{++}$ are the sets of nonnegative and positive real numbers, respectively, and $\mathbb{N}$ is the set of nonnegative integers.

2.2 Fundamental assumptions

Throughout this paper, we will repeatedly rely on the following two assumptions imposed on (SP) to develop our algorithms and analyze their convergence guarantees.

Assumption 1 The set of saddle-points $\mathcal{W}^* := \mathcal{X}^* \times \mathcal{Y}^*$ of (SP) is nonempty, i.e., there exists $(x^*, y^*) \in \mathcal{X}^* \times \mathcal{Y}^*$ such that:

$$
\tilde{L}(x^*, y) \leq \tilde{L}(x^*, y^*) \leq \tilde{L}(x, y^*), \quad \forall (x, y) \in \mathbb{R}^p \times \mathbb{R}^m.
$$

Assumption 2 is standard in saddle-point problems. With this, we can easily show the following connection between the primal problem (P) and its dual form (D):

$$
\mathcal{D}(y) \leq \mathcal{D}(y^*) = \mathcal{D}^* = \mathcal{P}^* = \mathcal{P}(x^*) \leq \mathcal{P}(x), \quad \forall (x, y) \in \mathbb{R}^p \times \mathbb{R}^m,
$$

(5)
where \( \mathcal{P} \) and \( \mathcal{D} \) are the primal and dual objectives defined in (\( \mathcal{P} \)) and (\( \mathcal{D} \)), respectively.

**Assumption 2** The functions \( F, H, \) and \( \Phi \) in (\( \text{SP} \)) satisfy the following conditions:

(a) The function \( F(x) = f(x) + h(x) \) is defined on \( \mathbb{R}^p \), where both \( f \) and \( h \) are proper, closed and convex. In addition, \( f \) is \( L_f \)-smooth for some Lipschitz constant \( L_f \in [0, \infty) \).

(b) The function \( H : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\} \) is proper, closed, and convex.

(c) The coupling function \( \Phi(x,y) = \langle y, g(x) \rangle \) is convex in \( x \) for any \( y \in \text{dom}(H^*) \). Moreover:

(i) For any \( y \in \text{dom}(H^*) \), \( \nabla_y \Phi(\cdot, y) = g(\cdot) \) is \( M_g \)-uniformly Lipschitz continuous with a vector of Lipschitz moduli \( M_g \in [0, \infty)^m \), i.e.:

\[
\| \nabla_y \Phi(x,y) - \nabla_y \Phi(\hat{x}, y) \| \leq \| M_g \| \| x - \hat{x} \| = M_g \| x - \hat{x} \|, \quad \forall x, \hat{x} \in \text{dom}(\mathcal{P}), \tag{6}
\]

where \( M_g := \| M_g \| \).

(ii) For any \( y \in \text{dom}(H^*) \), \( \nabla_x \Phi(\cdot, y) = \langle y, g(\cdot) \rangle \) is \( L_g(y) \)-Lipschitz continuous with a Lipschitz modulus \( L_g(y) \in (0, +\infty) \) depending on \( y \), i.e.:

\[
\| \nabla_x \Phi(x,y) - \nabla_x \Phi(\hat{x}, y) \| \leq L_g(y) \| x - \hat{x} \|, \quad \forall x, \hat{x} \in \text{dom}(\mathcal{P}).
\]

We further assume that \( L_g(y) \) satisfies \( 0 \leq L_g(y) \leq L_g \) for a fixed \( L_g \in (0, +\infty) \).

Since \( \nabla_y \Phi(x,y) = g(x) \), eqn. (6) is equivalent to the \( M_g \)-Lipschitz continuity of \( g_i \), where \( M_g \) is the \( i \)-th component of \( M_g \), and \( g_i \) is the \( i \)-th component of mapping \( g \) for \( i = 1, \cdots, m \). Clearly, if \( \Phi(x,y) = (Kx,y) \) is bilinear, then \( \nabla_x \Phi(x,y) = K^\top y \) and \( \nabla_y \Phi(x,y) = Ky \), which automatically satisfy Assumptions 2. Assumption 2 is standard in primal-dual methods for solving (\( \text{SP} \)) as used in [48, 80, 93, 94]. However, unlike these works, \( L_g(y) \) in Assumption 2 can depend on \( y \), which allows us to cover cone constrained problem (1) without requiring the boundedness of \( \text{dom}(F) \) or \( \text{dom}(H^*) \). Note that in item (c), the convexity of \( \Phi(\cdot, y) \) and the \( L_g(y) \)-smoothness of \( \nabla_x \Phi(\cdot, y) \) imply that

\[
0 \leq \langle y, g(\hat{x}) - g(x) - g'(x)(\hat{x} - x) \rangle \leq \frac{L_g(y)}{2} \| \hat{x} - x \|^2, \quad \forall x, \hat{x} \in \text{dom}(\mathcal{P}). \tag{7}
\]

In particular, if \( \text{dom}(\mathcal{P}) \) is nonempty, closed, convex, and bounded, and \( g \) is continuously differentiable on \( \text{dom}(\mathcal{P}) \), then \( g \) is \( M_g \)-Lipschitz continuous and \( \nabla_x \Phi(\cdot, y) \) is \( L_g(y) \)-smooth. Some existing works, e.g., [34, 43, 91], impose these conditions, but we do not require \( \text{dom}(\mathcal{P}) \) to be bounded.

### 2.3 Optimality condition and gap function

Now, we discuss the optimality condition of (\( \mathcal{P} \)) and (\( \mathcal{D} \)). In view of Assumption 1 and Fermat’s rule, there exists a pair of optimal solutions \((x^*, y^*) \in \mathbb{R}^p \times \mathbb{R}^m\) to the primal problem (\( \mathcal{P} \)) and its dual form (\( \mathcal{D} \)), which satisfies

\[
0 \in \partial F(x^*) + g'(x^\top) y^* \quad \text{and} \quad 0 \in g(x^*) - \partial H^*(y^*). \tag{8}
\]

Indeed, \( x^* := \max_x \tilde{\mathcal{L}}(x^*, y) \in \mathbb{R}^p \) and \( y^* = \max_y \tilde{\mathcal{L}}(x^*, y) \). The dual direction can be proved analogously.
**Gap function.** We consider two types of duality gap functions at a pair of solutions \((x, y)\). The first one is the standard primal-dual gap \(P(x) - D(y)\), which is nonnegative due to the weak duality as shown in [5], and it vanishes, i.e., \(P(x) - D(y) = 0\), if and only if \((x, y)\) is a saddle point of \((SP)\) due to strong duality.

Another gap function is defined as (see, e.g., [11, 24, 50]):

\[
G_{X \times Y}(x, y) := \sup_{x \in X, y \in Y} \left\{ \bar{L}(x, y) - \bar{L}(\hat{x}, y) \right\} = \sup_{y \in Y} \bar{L}(x, y) - \inf_{x \in X} \bar{L}(\hat{x}, y),
\]

where \(X \times Y\) contains a saddle-point. It is clear that \(G_{X \times Y}(x, y) \geq 0\) for any \((x, y)\) \(\in \mathbb{R}^p \times \mathbb{R}^m\), and when \((x, y)\) is a saddle-point, \(G_{X \times Y}(x, y) = 0\) [11]. This gap function is widely used in the literature on primal-dual convergence theory, e.g., [8, 11, 17].

It is clear that \(G_{X \times Y}(x, y) \leq G_{\mathbb{R}^p \times \mathbb{R}^m}(x, y) = P(x) - D(y)\). In our analysis, we will have convergence guarantees on both types of duality gaps. For the gap \(P(x) - D(y)\), we would require additional conditions such as the Lipschitz continuity on \(H\) and/or \(F^*\); in contrast, convergence guarantees on \(G_{X \times Y}\) do not require such conditions.

### 2.4 The augmented Lagrangian function and its properties

**Non-convex constrained reformulation.** To solve \((SP)\), we can write \((P)\) as

\[
\min_{(x, s) \in \mathbb{R}^p \times \mathbb{R}^m} \left\{ F(x) + H(-s) \quad \text{s.t.} \quad g(x) + s = 0 \right\},
\]

where \(s\) is the slack variable. If \(g\) is non-affine, then \((10)\) is non-convex. Moreover, the Lagrange function associated with \((10)\) can be written as

\[
\mathcal{L}(x, s, y) := F(x) + H(-s) + \langle y, g(x) + s \rangle,
\]

where \(y \in \mathbb{R}^m\) is a Lagrange multiplier. If \((x^*, y^*)\) is optimal to \((SP)\), i.e., satisfies \((8)\), then \((x^*, y^*, s^*)\) is optimal to \((10)\), where \(s^* = -g(x^*)\). Thus \((8)\) can be written as

\[
0 \in \partial F(x^*) + g'(x^*)^\top y^* \quad \text{and} \quad -g(x^*) = s^* \in -\partial H^*(y^*).
\]

By the Fenchel theorem, we have \(H(-s) + H^*(y) \geq -\langle s, y \rangle\), where the equality holds if and only if \(s \in -\partial H^*(y)\), or equivalently, \(y \in \partial H(-s)\). Therefore, it holds that

\[
\bar{L}(x, y) \leq \mathcal{L}(x, s, y) \quad \text{and} \quad \bar{L}(x, y) = \mathcal{L}(x, s, y) \iff s = -\partial H^*(y).
\]

Consequently, for any \((x, s, y) \in \mathbb{R}^p \times \mathbb{R}^m \times \mathbb{R}^m\), \((11)\) implies that

\[
\bar{L}(x^*, y) \leq \mathcal{L}(x^*, s^*, y) \leq \mathcal{L}(x^*, s^*, y^*) \leq \bar{L}(x, y^*) \leq \mathcal{L}(x, s, y^*).
\]

**Augmented Lagrangian function.** The augmented Lagrangian of \((10)\) is defined as

\[
\mathcal{L}_\rho(x, s, y) := \mathcal{L}(x, s, y) + \frac{\rho}{2} \|g(x) + s\|^2
\]

\[
= F(x) + H(-s) + \langle y, g(x) + s \rangle + \frac{\rho}{2} \|g(x) + s\|^2,
\]

where the scalar \(\rho > 0\) is a penalty parameter. Note that if \(g\) is not affine, then \(\mathcal{L}_\rho\) is not convex in \(x\). Some existing works [92, 89, 91] minimize \(\mathcal{L}_\rho\) over \(s\) to obtain a standard convex augmented Lagrangian function, first proposed in [73]; however, such formulation does not allow linear updates in \(y\), preventing a clear analysis when applying Nesterov’s acceleration technique. Therefore, we preserve \(s\) and keep the non-convex form of \(\mathcal{L}_\rho\), so that it is linear in \(y\). As will be shown, we only utilize the local convexity of \(\mathcal{L}_\rho\) in our analysis.

---

4 To be more specific, \(G_{X \times Y}\) could also vanish at non-saddle-points. However, if \((x, y)\) is in the interior of \(X \times Y\), then \(G_{X \times Y}(x, y) = 0\) if and only if \((x, y)\) is a saddle-point of \((SP)\), see, e.g., [11].
Augmented Lagrangian term. Let us introduce

$$\phi_p(x, s, y) := \langle y, g(x) + s \rangle + \frac{\rho}{2} \| g(x) + s \|^2. \quad (16)$$

Then, by (15), we have $\mathcal{L}_p(x, s, y) = F(x) + H^*(-s) + \phi_p(x, s, y)$. It is easy to see that at an optimal solution, i.e., a $(x^*, s^*, y^*)$-tuple that satisfies (12), we have $\phi_p(x^*, s^*, y^*) = 0$. Moreover, we can directly compute the first-order derivatives of $\phi_p$ as

$$
\begin{align*}
\nabla_x \phi_p(x, s, y) &= [g'(x)]^\top (y + \rho [g(x) + s]), \\
\nabla_s \phi_p(x, s, y) &= y + \rho [g(x) + s], \\
\nabla_y \phi_p(x, s, y) &= g(x) + s,
\end{align*}
$$

where $g'(x) \in \mathbb{R}^{m \times p}$ is the Jacobian of $g$ at $x$. For $d \in \mathbb{R}^p$, the Hessian of $\phi_p$ in $x$ to the direction of $d$ is given by

$$
\nabla^2_x \phi_p(x, s, y)[d, d] = \rho \| g'(x) [d]\|^2 + \sum_{i=1}^m (y_i + \rho [g_i(x) + s_i]) \nabla^2 g_i(x)[d, d].
$$

By Assumption 2(c), when $\hat{y} := y + \rho [g(x) + s] \in \text{dom}(H^*)$, we have that $\Phi(x, \hat{y})$ is convex in $x$, i.e., the last term in the last equality is nonnegative, and thus $\phi_p(x, s, y)$ is locally convex in $x$. Moreover, if we view $\phi_p$ as a function of $g$, then it is convex and $\rho$-smooth in $g$.

These important properties of $\phi_p$ leads to Lemma 1 which will be used to prove descent lemmas in Sections 3 and 4.

**Lemma 1** Let $\phi_p$ be defined in (16). For any $x, \hat{x} \in \mathbb{R}^p$, $s, \hat{s} \in \mathbb{R}^m$ and $y \in \mathbb{R}^m$ such that $y + \rho [g(x) + s] \in \text{dom}(H^*)$, we define the residual $\Delta_p$ of a linearization of $\phi_p$ at $(x, s, y)$ as

$$
\Delta_p(\hat{x}, \hat{s}; x, s, y) := \phi_p(\hat{x}, \hat{s}, y) - \phi_p(x, s, y) - \langle \nabla_x \phi_p(x, s, y), \hat{x} - x \rangle - \langle \nabla_s \phi_p(x, s, y), \hat{s} - s \rangle. \quad (18)
$$

Then, we have the following estimate:

$$
0 \leq \Delta_p(\hat{x}, \hat{s}; x, s, y) - \frac{\rho}{2} \| g(\hat{x}) + \hat{s} \| - \| g(x) + s \|^2 \leq \frac{L_g(y + \rho [g(x) + s])}{2} \| \hat{x} - x \|^2, \quad (19)
$$

where $L_g(y + \rho [g(x) + s])$ is the Lipschitz modulus defined by Assumption 2(c).

**Proof** See Appendix A.1.

3 Our First Algorithm: General Convex-Concave Case

In this section, we develop a novel algorithm to solve (SP) under the general convexity-concavity assumption, i.e., $F$ and $H^*$ are convex, but not necessarily strongly convex.

3.1 The derivation and the complete algorithm

Our main idea is to exploit the augmented Lagrangian $\mathcal{L}_p$ defined in (15) as a merit function to measure the progress of the iterate sequence $\{(x^k, y^k)\}$. Since this function not only involves $x$ but also the dual variables $y$, and the slack variable $s$, we also need to update them accordingly. To accelerate, we inject Nesterov’s accelerated steps [58] in $x$. Recall that $\phi_p$ is non-convex in $x$, but thanks to Lemma 1, we can still utilize its local convexity.

Step by step, we derive our scheme to solve (SP) as follows.
Step 1. We first update the slack variable $s^{k+1}$ by minimizing $L_{p_k}(\hat{x}^k, s, \tilde{y}^k)$ w.r.t. $s$:

$$s^{k+1} := \arg \min_{s \in \mathbb{R}^m} \left\{ H(-s) + \langle \tilde{y}^k, g(\hat{x}^k) + s \rangle + \frac{\rho_k}{2} \|g(\hat{x}^k) + s\|^2 \right\} = -\text{prox}_{H/p_k}(\tilde{y}^k + g(\hat{x}^k)). \tag{20}$$

Step 2. To update $x^{k+1}$, we would attempt to minimize $L_{p_k}(x, s^{k+1}, \tilde{y}^k)$ w.r.t. $x$. However, since minimizing this function directly is difficult, we instead linearize $f$ and $\phi_{p_k}(\cdot, s^{k+1}, \tilde{y}^k)$ at point $\hat{x}^k$, respectively:

$$\begin{align*}
\{ f(x) & \approx f(\hat{x}^k) + \langle \nabla f(\hat{x}^k), x - \hat{x}^k \rangle + \frac{L_f}{2} \|x - \hat{x}^k\|^2, \\
\phi_{p_k}(x, s^{k+1}, \tilde{y}^k) & \approx \phi_{p_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k) + \langle \nabla_x \phi_{p_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), x - \hat{x}^k \rangle + \frac{L_k}{2} \|x - \hat{x}^k\|^2,
\end{align*}$$

for some $L_f > 0$ and $L_k > 0$, respectively. Writing $\beta_k := 1/(L_f + L_k)$, we can combine the above two approximations and update $x^{k+1}$ as

$$x^{k+1} := \arg \min_{x \in \mathbb{R}^p} \left\{ h(x) + \langle \nabla f(\hat{x}^k) + \nabla_x \phi_{p_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), x - \hat{x}^k \rangle + \frac{1}{2p}\|x - \hat{x}^k\|^2 \right\} = \text{prox}_{\beta_k h}(\hat{x}^k + \beta_k [\nabla f(\hat{x}^k) + \nabla_x \phi_{p_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k)]). \tag{21}$$

Here, the actual value of $\beta_k > 0$ will be appropriately updated in our analysis.

Step 3. To accelerate the primal progress, we update $\hat{x}^k$ by applying Nesterov’s acceleration technique [57]:

$$\hat{x}^{k+1} := x^{k+1} + \tau_k + 1(1 - \tau_k)(x^{k+1} - x^k),$$

where the step-size $\tau_k \in (0, 1]$ will be updated appropriately.

Step 4. We update the dual variable $\tilde{y}^k$ as follows:

$$\tilde{y}^{k+1} := \text{proj}_{B_k} \left( \tilde{y}^k + \eta_k ([g(x^{k+1}) + s^{k+1}] - (1 - \tau_k)[g(x^k) + s^k]) \right), \tag{22}$$

where $B_k \subset \mathbb{R}^m$ is a norm ball, which will be specified later.

Step 5. Finally, we define the dual variable

$$\tilde{y}^{k+1} := \text{prox}_{\rho_k H^*}(\tilde{y}^k + \rho_k g(\hat{x}^k)). \tag{23}$$

However, by Moreau’s identity,

$$\rho_k s^{k+1} \text{prox}_{\rho_k H^*} \left( \tilde{y}^k + \rho_k g(\hat{x}^k) \right) = [\tilde{y}^k + \rho_k g(\hat{x}^k)] = y^{k+1} - [\tilde{y}^k + \rho_k g(\hat{x}^k)]. \tag{24}$$

Thus, we can in fact eliminate variable $s^{k+1}$ from the expression of $x^{k+1}$ in [21] by noting that $\nabla_x \phi_{p_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k) = g'(\hat{x}^k)^T \tilde{y}^{k+1}$. Similarly, the presence of $s^k$ and $s^{k+1}$ in the update of $\tilde{y}^{k+1}$ in [22] can also be eliminated. In this way, we can reformulate our algorithm into a standard primal-dual form [11, 23].

Combining the above steps, we arrive at our complete algorithm as in Algorithm 1.

**Per-iteration complexity.** We analyze the per-iteration complexity of Step 5 in Algorithm 1. The major computation includes:

(a) The first line requires one function evaluation of $g$ and one proximal operation of $H^*$.

(b) The second line needs to compute one Jacobian $g'(\hat{x}^k)$, one gradient $\nabla f$, and one proximal operation of $h$.

(c) The fourth line essentially uses one function evaluation of $g$ at $x^{k+1}$.

(d) The fifth line requires one projection on $B_k$ if necessary, i.e., when $B_k \subset \mathbb{R}^m$.

This break-down of complexity shows that Algorithm 1 essentially has the same complexity as other state-of-the-art primal-dual first-order algorithms, see, e.g., [56, 80, 94].
Algorithm 1 (Accelerated Primal-Dual Algorithm (General Convex-Concave Case))

1: **Initialization:** Choose an initial primal-dual point \((x^0, y^0) \in \mathbb{R}^p \times \mathbb{R}^m\).
2: Set \(x^0 := x^0\), \(y^0 := y^0\), and \(\theta_0 := 0\).
3: Choose appropriate initial parameters, according to \([29], [34]\), or \([36]\).
4: **For** \(k = 0\) to \(k_{\text{max}}\)
5: Update the parameters according to \([30], [35]\), or \([37]\), consistent with Step 3.
6: Update \((x^k, \hat{x}^k, y^k, \hat{y}^k)\) as follows:

\[
\begin{align*}
    y^{k+1} &:= \text{prox}_{\beta_k H^*} (\hat{y}^k + \rho_k g(\hat{x}^k)), \\
    x^{k+1} &:= \text{prox}_{\beta_{k+1}} (\hat{x}^k - \beta_k (\nabla f(\hat{x}^k) + g'(\hat{x}^k)^T y^{k+1})), \\
    \hat{x}^{k+1} &:= x^{k+1} + \frac{\tau_k(1-\tau_k)}{\tau_k} (x^{k+1} - x^k), \\
    \theta_{k+1} &:= g(x^{k+1}) - g(\hat{x}^k) + \frac{1}{\rho_k} (y^{k+1} - \hat{y}^k), \\
    \hat{y}^{k+1} &:= \text{proj}_{B_k} (\hat{y}^k + \eta_k [\theta_{k+1} - (1 - \tau_k) \theta_k]),
\end{align*}
\]

**EndFor**

3.2 Convergence rate analysis

The following lemma provides a recursive inequality based on scheme \([25]\), and will serve as a key estimate to analyze global convergence rates of Algorithm 1.

**Lemma 2** Define \(\mathcal{L}\) as in \([11]\), \(\mathcal{L}_p\) as in \([15]\), and \(L_f, M_g, \text{ and } L_g\) as in Assumption \([3]\). Let \(\{(x^k, \hat{x}^k, y^k, \hat{y}^k)\}\) be generated by \([29]\) with \(\tau_k \in (0, 1]\) and \(\rho_k > \eta_k\). Let \(\{s^k\}\) be defined in \([20]\). Further introduce

\[
L_k := L_g(y^{k+1}), \quad \bar{x}^k := \frac{1}{\tau_k} [x^k - (1 - \tau_k)x^k], \quad \text{and} \quad \bar{y}^{k+1} := (1 - \tau_k)\hat{y}^k + \tau_k y^{k+1}. \tag{26}
\]

Then, for all \(k \in \mathbb{N}\) and for any \((x, s, y) \in \mathbb{R}^p \times \mathbb{R}^m \times B_k\), it holds that

\[
\mathcal{L}_p(x^{k+1}, s^{k+1}, y) - \mathcal{L}(x, s, y^{k+1}) \leq (1 - \tau_k) \{\mathcal{L}_{p-1}(x^k, s^k, y) - \mathcal{L}(x, s, \hat{y}^k)\}
+ \frac{\tau_k}{2\tau_k} \left( \|\hat{x}^k - x\|^2 - \|x^{k+1} - x\|^2 \right) + \frac{1}{2\beta_k} \left( \|\hat{y}^k - y\|^2 - \|\hat{y}^{k+1} - y\|^2 \right) - \frac{\tau_k(1-\tau_k)}{2\tau_k} \rho_k (1 - \tau_k) \|g(x^k) + s^k\|^2
- \frac{1}{2} \left( \frac{1}{\rho_k} - L_k - L_f - \frac{\beta_k}{\rho_k - \eta_k} \right) \|x^{k+1} - \hat{x}^k\|^2. \tag{27}
\]

**Proof** See Appendix \([B.1]\) \(\square\)

Now, we analyze the convergence rates of Algorithm 1 for three parameter initialization (Step 3) and update (Step 6) options. To abbreviate the notation, given \(x^0 \in \mathbb{R}^p\), \(y^0 \in \mathbb{R}^m\), and \(\beta_0, \eta_0 > 0\), we frequently use the following quantity:

\[
\mathcal{R}_0^2(x, y) := \frac{1}{\beta_0} \|x^0 - x\|^2 + \frac{1}{\eta_0} \|y^0 - y\|^2 \tag{28}
\]

to characterize the weighted square-distance from the initial point \((x^0, y^0)\) to \((x, y)\).
3.2.1 The $O(1/k)$ ergodic convergence rate

The following theorem shows a $O(1/k)$ ergodic convergence rate of Algorithm 1.

**Theorem 1** Suppose Assumptions 1 and 2 hold for (SP). Let $\{(x^k, y^k)\}_{k \geq 0}$ be generated by Algorithm 1 with the following parameter configurations:

\begin{align*}
\text{Initialization:} & \quad \begin{cases} 
\beta := \frac{\gamma}{L_f + \rho \sqrt{C} + M_y^2}, \\
\eta := (1 - \gamma) \rho, \\
L_g \left[ \|y^*\| + (\sqrt{\eta} + \rho \sqrt{M_y}) R_0(x^*, y^*) \right] \leq \rho C,
\end{cases} \\
\text{Update:} & \quad \tau_k = 1, \quad \rho_k = \rho, \quad \beta_k = \beta, \quad \eta_k = \eta, \quad \text{and} \quad B_k = \mathbb{R}^m.
\end{align*}

Let $\{(\bar{x}^k, \bar{y}^k)\}_{k \geq 1}$ be the ergodic sequence defined as

$$ (\bar{x}^k, \bar{y}^k) := \frac{1}{k} \sum_{j=1}^{k} (x^j, y^j). $$

Then, for all $k \geq 1$, the following bounds hold:

$$
\begin{align*}
\mathcal{G}_{X \times Y}(\bar{x}^k, \bar{y}^k) & \leq \frac{1}{2k} \sup_{(x,y) \in X \times Y} \mathcal{R}_0^2(x, y), \\
\mathcal{P}(\bar{x}^k) - \mathcal{P}^* & \leq \frac{1}{2k} \left[ \frac{\|x^0 - x^*\|^2}{\beta} + \frac{(\|y^0\| + M_H)^2}{\eta} \right], \\
\mathcal{D}^* - \mathcal{D}(\bar{y}^k) & \leq \frac{1}{2k} \left[ \frac{\left(\|x^0\| + M_{F^*}\right)^2}{\beta} + \frac{\|y - y^*\|^2}{\eta} \right], \\
\mathcal{P}(\bar{x}^k) - \mathcal{D}(\bar{y}^k) & \leq \frac{1}{2k} \left[ \mathcal{R}_0^2(x^*, y^*) + \frac{\left(\|x^0\| + M_{F^*}\right)^2}{\beta} + \frac{(\|y^0\| + M_H)^2}{\eta} \right],
\end{align*}
$$

where $\mathcal{R}_0$ is defined by (28), and $M_H, M_{F^*} \in [0, \infty]$ are the Lipschitz constants of $H$ and $F^*$, respectively.

As a result, Algorithm 1 has $O(1/k)$ ergodic convergence rate on the primal objective residual, the dual objective residual, and the primal-dual gaps.

**Proof** See Appendix B.2.”

The first convergence guarantee on $\mathcal{G}_{X \times Y}$ in (32) is independent of $M_H$ and $M_{F^*}$, while the last one depends on both $M_H$ and $M_{F^*}$. Hence, the right-hand-side of the primal (resp., dual) convergence bound is finite if $M_H$ (resp., $M_{F^*}$) is finite. Note that under the update rule (29), Step 6 of Algorithm 1 can be simply written as

$$
\begin{align*}
y^{k+1} & := \text{prox}_{\rho H^*} \left( \bar{y}^k + \rho g(x^k) \right), \\
x^k & := \text{prox}_{\beta h} \left( x^k - \beta \left[ \nabla f(x^k) + g'(x^k)^\top y^{k+1} \right] \right), \\
\bar{y}^{k+1} & := \bar{y}^k + \eta [g(x^{k+1}) - g(x^k) + \frac{1}{\rho} (y^{k+1} - \bar{y}^k)].
\end{align*}
$$

This scheme requires one proximal operation of $H^*$ and $h$ each, one evaluation of $g$, one evaluation of gradient $\nabla f$ and one evaluation of Jacobian $g'$. If $H = \delta_{\mathbb{R}^n} \times \{0\}^n$, the indicator of $\mathbb{R}^n \times \{0\}^n$, then this scheme is similar to the one in [92] for solving (2). However, our dual step $\bar{y}^k$ is different from the one in [92].
Remark 1 (Initialization in (29)) In fact, for any choice of \( \rho > 0 \) and \( \gamma \in (0, 1) \), we can find \( C > 0 \) that satisfies (29). For example, we can simply set 
\[
\rho := 1, \quad \gamma := \frac{1}{2} \quad \text{and} \quad C := \max \{ L_f + 2M_g^2 + 2, \ L_gD(L_gD + 4M_g + 2) \},
\]
where \( D \geq \max \{ \| x_0 - x^* \|, \| y_0 - y^* \|, \| y^* \| \} \) is an upper estimate. As shown in Appendix B.3, the choice given in (33) is feasible to (29). Notice that \( C \) presented in (33) is not tight, since we have loosened this estimate to get simple expressions. One may choose different \( \rho \)'s and smaller \( C \)'s, which also solve (29), for better practical performance.

3.2.2 The \( O(1/k) \) semi-ergodic convergence rate

The following theorem shows \( O(1/k) \) semi-ergodic rate of Algorithm 1 for solving (SP) using the last-iterate primal sequence \( \{x^k\} \) and the averaging dual sequence \( \{\tilde{y}^k\} \).

Theorem 2 Suppose Assumptions 1 and 2 hold for (SP). In addition, assume that

(i) \( g(x) \leq B_g \) for all \( x \in \text{dom}(g) \cap \text{dom}(F) \) for some \( B_g \in [0, \infty) \) such that \( L_gB_g < +\infty \). In particular, if \( g \) is affine, then \( L_g = 0 \), and we allow \( B_g = \infty \) (i.e., no boundedness on \( g \) is required).

(ii) \( 0 \in \text{dom}(\partial H) \) and \( 0 \in \text{dom}(\partial H^*) \).

Let \( \{ (x^k, y^k) \}_{k \geq 0} \) be generated by Algorithm 1 with the following parameter configurations:

**Initialization:** Choose \( y_* \in \partial H(0), s_* \in -\partial H^*(0) \) and
\[
\rho_0 > 0 \quad \text{and} \quad \gamma \in (0, 1).
\]

**Update:** For all \( k \in \mathbb{N} \), fix \( B_k \equiv \mathbb{R}^m \), and update
\[
\tau_k := \frac{1}{k+1}, \quad \rho_k := \frac{\rho_0}{\tau_k}, \quad \eta_k := (1 - \gamma) \rho_k, \quad \text{and} \quad \beta_k := \gamma \frac{\gamma}{\gamma(L_f + 2L_g\|y_*\|) + \rho_k \left( L_g \frac{\| y_* \|}{\rho_0} + (2 - \gamma) B_g + 2(1 - \gamma) \| s_* \| + M_g^2 \right)}.
\]

Let \( \{ \tilde{y}^k \}_{k \geq 1} \) be the ergodic sequence defined in (31). Then, for \( k \geq 1 \), the following holds:

\[
\left\{ \begin{array}{l}
\mathcal{G}_{X \times Y}(x^k, \tilde{y}^k) \leq \frac{1}{2k} \sup_{(x,y) \in X \times Y} \mathcal{R}_0^2(x, y), \\
\mathcal{P}(x^k) - \mathcal{P}^* \leq \frac{1}{2k} \left[ \frac{\| x_0 - x^* \|^2}{\beta_0} + \frac{(\| y_0 \| + M_H)^2}{\eta_0} \right], \\
\mathcal{D}^* - \mathcal{D}(\tilde{y}^k) \leq \frac{1}{2k} \left[ \frac{\| x_0 \| + M_{F^*}}{\beta_0} + \frac{\| y_0 - y^* \|^2}{\eta_0} \right], \\
\mathcal{P}(x^k) - \mathcal{D}(\tilde{y}^k) \leq \frac{1}{2k} \left[ \mathcal{R}_0^2(x^k, y^k) + \frac{(\| x_0 \| + M_{F^*})^2}{\beta_0} + \frac{(\| y_0 \| + M_H)^2}{\eta_0} \right],
\end{array} \right.
\]

where \( \mathcal{R}_0 \) is defined in (28), and \( M_H, M_{F^*} \in [0, \infty) \) are the Lipschitz constants of \( H \) and \( F^* \), respectively.

As a result, Algorithm 1 has \( O(1/k) \) non-ergodic convergence rate on primal objective residual, ergodic rate on dual objective residual, and semi-ergodic rate on primal-dual gaps.
Remark 2 Condition (i) in Theorem 2 is not a strong assumption. When $H^*$ is separable in $y$, e.g., when $H^*(y) = \delta_{R^*}(y)$, the indicator of non-negative orthant, then condition $L_g B_g < +\infty$ can be relaxed to $\sum_{i=1}^{m} L_{g_i} B_{g_i} < +\infty$, where $B_{g_i}$ is the bound for $g_i$. Therefore, our condition allows both linear and bounded nonlinear constraint functions.

Again, the right-hand side of the primal (resp., dual) convergence rate bound in Theorem 2 is finite if $M_H$ (resp., $M_{F^*}$) is finite. □

3.2.3 The $\min\{O(1/k), \varrho(1/(k\sqrt{\log k}))\}$ non-ergodic convergence rate

We show in Theorem 3 below that if we modify the update rule of $\tau_k$, then we can boost the convergence rate of Algorithm 1 up to $\min\{O(1/k), \varrho(1/(k\sqrt{\log k}))\}$ in the non-ergodic sense on the primal objective residual, where $\varrho(\cdot)$ is defined in (3). Here, since $\varrho$-rate is not necessarily strictly faster than $O$-rate, we use the “min” to imply that our rate is no slower than $O(1/k)^5$. Our convergence guarantee is only on the primal problem (P).

Theorem 3 Suppose Assumptions 1 and 2 hold for (SP). In addition, assume that

(i) $\|g(x)\| \leq B_g$ for all $x \in \text{dom}(g) \cap \text{dom}(F)$ for some $B_g \in [0, \infty]$ such that $L_g B_g < +\infty$. In particular, if $g$ is affine, then $L_g = 0$, and we allow $B_g = \infty$.
(ii) $0 \in \text{dom}(\partial H)$ and $H$ is $M_H$-Lipschitz continuous with $M_H \in [0, +\infty)$.

Let $\{(x^k, y^k)\}_{k \geq 0}$ be generated from Algorithm 1 with the following parameter configurations:

| Initialization: Choose |
|------------------------|
| $\rho_0 > 0$, $\gamma \in (0, 1)$, $c > 1$, and $R_y \geq \frac{\|y^*\|}{\rho_0}$. |
| Update: Fix $y_* \in \partial H(0)$. For $k \in \mathbb{N}$, set |
| $\tau_k := \frac{c}{k + c}$, $\rho_k := \frac{\rho_0}{\tau_k}$, $\eta_k := (1 - \gamma)\rho_k$, $B_k := \{y \mid \|y\| \leq \rho_k R_y\}$, |
| and $\beta_k := \frac{\gamma}{\gamma(L_f + 2 L_g \|y_*\|) + \rho_k \left[\gamma L_g (R_y + B_g) + M_g^2\right]}$. |

Then, the following guarantees hold:

$$P(x^k) - P^* \leq \frac{R_p^2}{k + c - 1} \text{ for } \forall k \geq 1 \text{ and } \liminf_{k \to \infty} k \sqrt{\log k}[P(x^k) - P^*] = 0,$$

where $R_p := \Delta_0^2 + \sqrt{2c/\rho_0(\|y^*\| + M_H)\Delta_0}$ and $\Delta_0 := (c - 1) \left[P(x^0) - P^*\right] + \frac{c}{2} R_p^2(x^*, y^*)$ with $R_p$ defined by (28).

As a result, Algorithm 1 for solving (P) has $\min\{O(1/k), \varrho(1/(k\sqrt{\log k}))\}$ non-ergodic convergence rate on the primal objective residual.

Proof See Appendix B.5 □

---

5. Indeed, the numerical experiment in Subsection 6.1 shows that the parameter update provided in Theorem 2 greatly boosts the performance of Algorithm 1.
4 Our Second Algorithm: Strongly Convex-Concave Case

Recall that $F := f + h$ as defined in Assumption 2, where $f$ is $L_f$-smooth, and $h$ is not necessary smooth. In addition to Assumptions 1 and 2 in this section, we impose the following assumption:

**Assumption 3** The function $h$ in Assumption 2(a) is $\mu_h$-strongly convex with $\mu_h > 0$.

Note that even if $h$ is not strongly convex, but $f$ is $\mu_f$-strongly convex with $\mu_f > 0$, then we can let $\hat{h}(x) := h(x) + \frac{\mu_f}{2} \|x\|^2$, and $\hat{f}(x) := f(x) - \frac{\mu_f}{2} \|x\|^2$. In this way, we have $\mu_{\hat{h}} = \mu_f$. Hence, Assumption 3 still holds for $\hat{h}$, and we can apply the algorithms in this section to solve (SP) with the same objective term $F(x) = \hat{f}(x) + \hat{h}(x)$.

4.1 The derivation and complete algorithm

Under Assumptions 1, 2, and 3, to achieve faster convergence rates, we modify the scheme (25) by replacing Nesterov’s accelerated steps by Tseng’s steps [87], i.e., the $x^{k+1}$-update in (21) is broken into two lines of updates:

$$
\begin{align*}
\hat{x}^{k+1} &:= \text{prox}_{\tau_k h} \left( \tilde{x}^k - \frac{\partial}{\tau_k} \left[ \nabla f(\tilde{x}^k) + \nabla_x \phi_{\rho_k}(\tilde{x}^k, s^{k+1}, y^k) \right] \right), \\
x^{k+1} &:= \text{prox}_{\alpha_k} \left( \tilde{x}^k - \alpha_k \left[ \nabla f(\tilde{x}^k) + \nabla_x \phi_{\rho_k}(\tilde{x}^k, s^{k+1}, y^k) \right] \right),
\end{align*}
$$

in order to achieve the faster $O\left(1/k^2\right)$ and $O(1/(k^2 \sqrt{\log k})$ convergence rates. Here, the slack variable $s^{k+1}$ is still defined as in (20). In the meantime, the $y^{k+1}$-, $\tilde{y}^{k+1}$-, and $\Theta_{k+1}$-updates, as well as the relation $\tau_k \hat{x}^k = \tilde{x}^k - (1 - \tau_k) x^k$ in (26), are the same as before.

Using the expression of partial derivative $\nabla_x \phi$ in (17), we present the resulting algorithm as Algorithm 2.

---

**Algorithm 2** (Accelerated Primal-Dual Algorithm (Strongly Convex-Concave Case))

1: **Initialization:** Choose an initial primal-dual point $(x^0, y^0) \in \mathbb{R}^p \times \mathbb{R}^m$.
2: Set $\tilde{x}^0 = x^0$, $\tilde{y}^0 = y^0$, and $\Theta_0 := 0$.
3: Choose appropriate initial parameters, according to (41), (46), or (49).
4: **For** $k = 0$ to $k_{\text{max}}$
5: Update parameters according to (42), (47), or (50), consistent with Step 3.
6: Update $(\hat{x}^k, \tilde{x}^k, x^{k+1}, y^{k+1}, \tilde{y}^k)$ as follows:

$$
\begin{align*}
y^{k+1} &:= \text{prox}_{\rho_k H} \left( \tilde{y}^k + \rho_k g(\tilde{x}^k) \right), \\
\tilde{x}^{k+1} &:= \text{prox}_{\beta_k h} \left( \tilde{x}^k - \frac{\partial}{\beta_k} \left[ \nabla f(\tilde{x}^k) + g'(\tilde{x}^k)^\top y^{k+1} \right] \right), \\
x^{k+1} &:= \text{prox}_{\alpha_k} \left( \tilde{x}^k - \alpha_k \left[ \nabla f(\tilde{x}^k) + g'(\tilde{x}^k)^\top y^{k+1} \right] \right), \\
\hat{x}^{k+1} &:= (1 - \tau_{k+1}) x^{k+1} + \tau_{k+1} \tilde{x}^{k+1}, \\
\Theta_{k+1} &:= g(x^{k+1}) - g(\tilde{x}^{k+1}) + \frac{1}{\rho_k} (y^{k+1} - \tilde{y}^k), \\
\tilde{y}^{k+1} &:= \text{proj}_{B_h} \left( \tilde{y}^k + \eta_k \left[ \Theta_{k+1} - (1 - \tau_k) \Theta_k \right] \right).
\end{align*}
$$

**EndFor**

---

**Per-iteration complexity.** The per-iteration complexity of Algorithm 2 is the same as that of Algorithm 1 except for one additional proximal operator of $h$ at line 3 of Step 6.
4.2 Convergence rate analysis

Parallel to Subsection 3.2, let us first present a key recursive estimate to analyze the convergence of Algorithm 2.

**Lemma 3** Define $\mathcal{L}$ as in (11), $\mathcal{L}_{p}$ as in (15), and $L_{f}$, $M_{g}$, and $L_{g}$ as in Assumption 2. Let $\{x^{k}, \hat{x}^{k}, \bar{x}^{k}, y^{k}, \hat{y}^{k}\}$ be generated by (39) with $\tau_{k} \in [0, 1]$, $\rho_{k} > \eta_{k}$, and $\alpha_{k} > \beta_{k}$. Furthermore, define $\{s^{k}\}$ as in (20), and define $L_{k}$ and $\{\bar{y}^{k}\}$ as in (26). Then, for all $k \in \mathbb{N}$ and any $(x, s, y) \in \mathbb{R}^{p} \times \mathbb{R}^{m} \times \mathbb{B}_{k}$, it holds that:

\[
\begin{align*}
\mathcal{L}_{\rho_{k}}(x^{k+1}, s^{k+1}, y) - \mathcal{L}(x, s, \bar{y}^{k+1}) & \leq (1 - \tau_{k})[\mathcal{L}_{\rho_{k-1}}(x^{k}, s^{k}, y) - \mathcal{L}(x, s, \bar{y}^{k})] \\
& + \frac{\tau_{k}^{2}}{2M_{g}} \|\hat{x}^{k} - x\|^{2} - \frac{\tau_{k}(\tau_{k} + \beta_{k})}{2\rho_{k}} \|\hat{x}^{k+1} - x\|^{2} + \frac{1}{2\rho_{k}} \|\hat{y}^{k} - y\|^{2} - \|\hat{y}^{k+1} - y\|^{2} \\
& - \frac{1 - \tau_{k}}{2} \left[\rho_{k-1} - (1 - \tau_{k})\rho_{k}\right] \|g(x^{k}) + s^{k}\|^{2} \\
& - \frac{1}{2 \alpha_{k}} \left(1 - \frac{\rho_{k}}{\alpha_{k}}\right) + \frac{1}{\alpha_{k}} - L_{f} - L_{f} - \frac{\alpha_{k}M_{g}^{2}}{\mu_{k} - \eta_{k}} \|x^{k+1} - \hat{x}^{k}\|^{2}.
\end{align*}
\]

**Proof** See Appendix C.1.

Now, we establish three types of convergence rates for Algorithm 2. Each type of convergence rate is obtained by specifying the initialization and update rule for the parameters such as $\tau_{k}$, $\rho_{k}$, and $\beta_{k}$.

### 4.2.1 The $O(1/k^2)$ ergodic convergence rate

We first prove in Theorem 4 that Algorithm 2 enjoys $O(1/k^2)$ ergodic rate without assuming the boundedness of $g$ or $B_{k}$.

**Theorem 4** Suppose Assumptions 1, 2, and 3 hold for (SP). Let $\{(x^{k}, y^{k})\}_{k \geq 0}$ be generated by Algorithm 2 with the following parameter configurations:

**Initialization:** Choose $\rho_{0}$, $\beta_{0}$, $\eta_{0}$, $\hat{M} > 0$, and $\gamma$, $\Gamma \in (0, 1)$ such that

\[
\begin{align*}
\left\{\begin{array}{l}
\beta_{0} := \frac{\Gamma}{L_{f} + \rho_{0}M_{g}^{2}} \\
\eta_{0} := (1 - \gamma)\rho_{0} \\
L_{g} \left[\|y^{\ast}\| + \sqrt{\eta_{0} + \rho_{0}\sqrt{M_{g}}} \right] & \leq \rho_{0} \left[(2 - \Gamma)\hat{M}^{2} - \frac{M_{g}^{2}}{\gamma}\right].
\end{array}\right.
\end{align*}
\]

**Update:** For all $k \in \mathbb{N}$, set $B_{k} \equiv \mathbb{R}^{m}$, and update

\[
\begin{align*}
\left\{\begin{array}{l}
\tau_{k} \equiv 1, \\
\alpha_{k} := \frac{1}{L_{f} + \rho_{k}M_{g}^{2}}, \\
\eta_{k} := (1 - \gamma)\rho_{k}, \\
\theta_{k+1} := \frac{1}{1 + \tau_{k}M_{g}}, \\
\rho_{k+1} := \frac{\rho_{k}}{\theta_{k+1}}, \quad \text{and} \quad \beta_{k+1} := \theta_{k+1}\beta_{k}.
\end{array}\right.
\end{align*}
\]

Let $\{x^{k}, \hat{y}^{k}\}_{k \geq 1}$ be an ergodic sequence defined as

\[
(\hat{x}^{k}, \hat{y}^{k}) := \frac{1}{\Sigma_{k}} \sum_{j=0}^{k-1} \rho_{j}(x^{j+1}, y^{j+1}), \quad \text{where} \quad \Sigma_{k} := \sum_{j=0}^{k-1} \rho_{j}.
\]
Then, for any \( k \geq 2 \), the following bounds hold:

\[
\begin{align*}
G_{X \times Y}(\bar{x}^k, \bar{y}^k) & \leq \frac{1}{(1 + \mu_k \beta_0 - 1)k(k-1)} \sup_{(x,y) \in X \times Y} R_0^2(x,y), \\
P(\bar{x}^k) - P^* & \leq \frac{1}{(1 + \mu_k \beta_0 - 1)k(k-1)} \left[ \frac{\|x^0 - x^\star\|^2}{\beta_0} + \frac{(\|y^0\| + M_H^2)}{\eta_0} \right], \\
D^* - D(\bar{y}^k) & \leq \frac{1}{(1 + \mu_k \beta_0 - 1)k(k-1)} \left[ \frac{(\|x^0\| + M_F^2)}{\beta_0} + \frac{\|y^0 - y^\star\|^2}{\eta_0} \right], \\
P(\bar{x}^k) - D(\bar{y}^k) & \leq \frac{1}{(1 + \mu_k \beta_0 - 1)k(k-1)} \left[ \frac{R_0^2(x^\star, y^\star)}{\beta_0} + \frac{(\|x^0\| + M_F^2)}{\beta_0} + \frac{(\|y^0\| + M_H^2)}{\eta_0} \right],
\end{align*}
\]

where \( R_0 \) is defined in (28), and \( M_H, M_F^2 \in [0, \infty) \) are the Lipschitz constants of \( H \) and \( F^\star \), respectively.

As a result, Algorithm 2 for solving (SP) has \( O(1/k^2) \) ergodic convergence rate on the primal objective residual, the dual objective residual, and the primal-dual gaps.

**Proof** See Appendix C.2

**Remark 3** (Initial parameters in (41)) As shown in Appendix C.3, the following parameter values are feasible to (41):

\[
\rho_0 := 1, \quad \gamma := \Gamma := \frac{1}{2}, \quad \text{and} \quad M^2 := \max \left\{ L_f + 1, \frac{8L_D^2}{9}, \frac{4(2M_D^2 + L_D \sqrt{2}M_D^2)}{3} \right\},
\]

where \( D \) is defined in Remark 1. This bound is relatively loose in pursuit of a simple expression. Thus one can choose tighter values for these parameters that satisfy (41) to achieve better practical performance.

**Remark 4** If we let \( \tau_k \equiv 1 \) and \( \alpha_k \equiv \beta_k \) (i.e., \( \Gamma := 1 \)), then scheme (39) is simplified as

\[
\begin{align*}
y^{k+1} & := \text{prox}_{\rho_k H^*}(\hat{y}^k + \rho_k g(x^k)), \\
x^{k+1} & := \text{prox}_{\beta_k h}(x^k - \beta_k (\nabla f(x^k) + g(x^k)^\top y^{k+1})), \\
\hat{y}^{k+1} & := \hat{y}^k + \eta_k [g(x^{k+1}) - g(x^k) + \frac{1}{\rho_k}(y^{k+1} - \hat{y}^k)],
\end{align*}
\]

with only one proximal operation. Now, if we combine the initialization condition (29) (with \( (\rho, \beta, \eta) \) there replaced by \( (\alpha_0, \beta_0, \eta_0) \)) and the update rule (42) (except for the absence of \( \alpha_k \)), then we would still achieve the same \( O(1/k^2) \) ergodic convergence rates. This guarantee can be proved using similar lines as the proofs of Lemma 3 and Theorem 4.

### 4.2.2 The \( O(1/k^2) \) semi-ergodic convergence rate

Next, we analyze the semi-ergodic convergence rate of Algorithm 2 in Theorem 5 below.

**Theorem 5** Suppose Assumptions 1, 2 and 3 hold for (SP). In addition, assume that

(i) \( \|g(x)\| \leq B_g \) for all \( x \in \text{dom}(g) \cap \text{dom}(F) \) for some \( B_g \in [0, \infty) \) such that \( L_F B_g < +\infty \).

(ii) There exists \( s_0 \in -\partial H^*(0) \) and there exists \( y_0 \in \partial H(0) \) such that \( \|y_0\| \leq \frac{(1-\Gamma)L_f}{2L_g} \), where \( \Gamma \) will be chosen below.

Let \( \{(x^k, y^k)\}_{k \geq 1} \) be generated by Algorithm 2 with \( y^0 := 0 \) and the following configurations:
4.2.3 Suppose that Assumptions 1, 2, and 3 hold for (SP). Again, if $H$ is affine, then the condition (i) in Theorem 5 automatically holds since $\gamma(2 - \Gamma) > 0$. The condition (ii) in Theorem 5 is not restrictive. We can always shift the function $H$ by a linear term to obtain this condition. If $g$ is affine, then the condition (i) in Theorem 5 automatically holds since $L_g = 0$. □

Remark 5 The condition $\gamma_\ast \in \partial H(0)$ such that $\|\gamma_\ast\| \leq \frac{(1 - \Gamma)L_f}{2L_g}$ in Theorem 5 is not restrictive. We can always shift the function $H$ by a linear term to obtain this condition. Again, if $g$ is affine, then the condition (i) in Theorem 5 automatically holds since $L_g = 0$. □

4.2.3 The non-ergodic convergence rate

Finally, using a different update rule for parameters, we establish a potentially faster non-ergodic convergence rate of Algorithm 2 in Theorem 6 below.

Theorem 6 Suppose that Assumptions 1, 2, and 3 hold for (SP). In addition, assume that

(i) $\|g(x)\| \leq B_g$ for all $x \in \text{dom}(g) \cap \text{dom}(F)$ for some $B_g \in [0, \infty]$ such that $L_g B_g < \infty$.
(ii) There exists $y_\ast \in \partial H(0)$ such that $\|y_\ast\| \leq \frac{(1 - \Gamma)L_f}{2L_g}$, where $\Gamma$ will be chosen below.

Let $\{(x^k, y^k)\}_{k \geq 0}$ be generated by Algorithm 2 using the following parameter configurations:
In this section, we specify our algorithms and their convergence results to handle (1). One important special case of (SP) is the cone constrained convex optimization problem (CP).

Initialization: Choose $\rho_0$, $M$, $R_y > 0$, $c > 2$, and $\gamma$, $\Gamma \in (0, 1)$, such that

$$\frac{M^2 + \gamma L_y(R_y + B_y)}{\gamma (2 - \Gamma)} \leq M^2 \leq \frac{c^2 \Gamma \mu_h}{(2c + 1) \rho_0} \quad \text{and} \quad \rho_0 R_y \geq \|y^*\|. \quad (49)$$

Update: For all $k \in \mathbb{N}$, update

$$\begin{align*}
\tau_k &:= \frac{c}{k + c}, \\
\rho_k &:= \frac{\rho_0}{\tau_k^2}, \\
\alpha_k &:= \frac{1}{L_f + \rho_k M^2}, \\
\beta_k &:= \Gamma \alpha_k, \\
\eta_k &:= (1 - \gamma) \rho_k, \quad \text{and} \quad B_k := \{y \mid \|y\| \leq \rho_{k-1} R_y\}.
\end{align*} \quad (50)$$

Then, one has

$$\mathcal{P}(x^k) - \mathcal{P}^* \leq \frac{R^2_p}{(k + c - 1)^2} \quad \text{for all } k \geq 0, \quad \text{and} \quad \lim_{k \to \infty} k^2 \sqrt{\log k} [\mathcal{P}(x^k) - \mathcal{P}^*] = 0.$$  

$R^2_p := \Delta_0^2 + c \sqrt{2/\rho_0} (\|y^*\| + M_H) \Delta_0$ and $\Delta_0^2 := (c - 1)^2 [\mathcal{P}(x^0) - \mathcal{P}^*] + \frac{c^2}{2} R^2_0(x^*, y^*)$. Here, $M_H \in [0, \infty]$ is the Lipschitz constant of $H$, which is assumed to be finite.

As a result, Algorithm 2 for solving (P) has $\min \{O(1/k^2), g(1/(k^2 \sqrt{\log k}))\}$ non-ergodic convergence rate on the primal objective residual.

Proof See Appendix C.5

Remark 6 (Initial parameters in (46) and (49)) The initializations in Theorems 5 and 6 are both feasible. For simplicity, one may set $\gamma := \Gamma := \frac{1}{2}$, then choose $M^2$ such that the first inequality in (46) or (49) is tight. For (49), one can then easily solve a quadratic equation for $c > 2$. However, the user may choose other feasible initial parameters for better practical performance.

5 Application to Cone Constrained Convex Optimization

One important special case of (SP) is the cone constrained convex optimization problem (1). In this section, we specify our algorithms and their convergence results to handle (1).

For our convenience of reference, let us recall (11) as follows:

$$\min_{x \in \mathbb{R}^p} \left\{ F(x) := f(x) + h(x) \quad \text{s.t.} \quad g(x) \in -\mathcal{K} \right\}. \quad \text{(CP)}$$

By Assumption 2(c), since $\langle y, g(x) \rangle$ is convex in $x$ for any $y \in \mathcal{K}^*$, $g$ is $\mathcal{K}$-convex, i.e., for all $x$, $\hat{x} \in \text{dom}(g)$ and $t \in [0, 1]$, it holds that $(1 - t)g(x) + tg(\hat{x}) - g((1 - t)x + t\hat{x}) \in \mathcal{K}$. Thus the constraint in (CP) is convex. Some special cases of (CP) have been listed in Section 1.

To develop special variants of Algorithms 1 and 2 for solving (CP) and establish their convergence guarantees, we redefine the associated Lagrange function as

$$\mathcal{L}(x, s, y) := F(x) + \langle y, g(x) + s \rangle, \quad \text{(51)}$$

where $s \in \mathcal{K}$ is again the slack variable, and $y \in \mathcal{K}^*$ is the Lagrange multiplier.

The following theorem tailors both Algorithms 1 and 2 to (CP), and provides their convergence rate guarantees on both the primal objective value and the feasibility gap.
Theorem 7 To specify Algorithms 1 and 2 for solving (CP), we replace the update of $y^{k+1}$ at Step 6 by

$$y^{k+1} := \text{proj}_{K^c} \left( \hat{y}^k + \rho_k g(\hat{x}^k) \right).$$

In addition, we define

$$\mathcal{E}(x) := \max \left\{ |F(x) - F^*|, \text{dist}_{K^c}(g(x)) \right\} \quad \text{and} \quad C_0 := \frac{\|x^0 - x^*\|^2}{\beta_0} + \left( \|x^0\| + \|y^0\| + 1 \right)^2,$$

where $\mathcal{E}(x)$ denotes the combined primal objective residual and primal feasibility violation at $x$. Then, for all $k \geq 1$, the following statements hold:

(a) Under the conditions of Theorem 1, we have $\mathcal{E}(\hat{x}^k) \leq \frac{C_0}{2k}$.

(b) Under the conditions of Theorem 2, we have $\mathcal{E}(x^k) \leq \frac{C_0}{2k}$.

(c) Under the conditions of Theorem 3, we have

$$\mathcal{E}(x^k) \leq \frac{\Delta_0}{k + c - 1} \quad \text{and} \quad \liminf_{k \to \infty} k \sqrt{\log k} \cdot \mathcal{E}(x^k) = 0,$$

where $\Delta_0 := (c - 1)[F(x^0) - F^*] + \frac{\sqrt{2}}{2} R_0(x^*, y^*)$.

(d) Under the conditions of Theorem 4, we have $\mathcal{E}(\hat{x}^k) \leq \frac{C_0}{2k}$.

(e) Under the conditions of Theorem 5, we have $\mathcal{E}(x^k) \leq \frac{C_0}{2(k + 1)^2}$.

(f) Under the conditions of Theorem 6, we have

$$\mathcal{E}(x^k) \leq \frac{\Delta_0}{k + c - 1} \quad \text{and} \quad \liminf_{k \to \infty} k^2 \sqrt{\log k} \cdot \mathcal{E}(x^k) = 0,$$

where $\Delta_0 := (c - 1)^2[F(x^0) - F^*] + \left( c - 1 + \frac{\sqrt{2} \beta_0}{2} \right) \frac{e_0^2}{\beta_0} \|x^0 - x^*\|^2 + \frac{\sqrt{2}}{2 \beta_0} \|y^0 - y^*\|^2$.

As a result, Algorithm 1 for solving (CP) is convergent on both the objective residual and the feasibility violation, with convergence rate $O(1/k)$ in ergodic sense, and rate $\min\{O(1/k), g(1/k \sqrt{\log k})\}$ in non-ergodic sense. Alternatively, Algorithm 2 for solving (CP) boosts these rates to $O(1/k^2)$ and $\min\{O(1/k^2), g(1/k^2 \sqrt{\log k})\}$, respectively.

Proof See Appendix D

6 Numerical Experiments

In this section, we aim at testing our algorithms on two numerical examples. The first one is a special case of quadratically constrained quadratic programming (QCQP) in Subsection 6.1. We use this example to verify the theoretical convergence rates of our algorithms. The second example is a convex-concave min-max game in Subsection 6.2.

We suggest the following tips when implementing our algorithms in order to obtain faster performance. These tips are guided by our theoretical results.

- As briefly discussed in Remark 2, when $H^*$ is separable (or block-separable) in $y$, which is often the case, such as QCQP, instead of using the product such as $L_g M_g$ in $\bigoplus_{m=1}^m L_g M_g$. In this case, Theorem 4 still holds. Similarly, the product $L_g B_g$ can be replaced by $\sum_{g=1}^G L_g B_g$ in the expressions of parameter initialization updates, e.g., in (45) and (46), and the theorems still hold true. Therefore, it is useful to use such replacements in implementation.
• One can tune the initial parameters, such as $\rho_0$ and $\beta_0$, in order to improve the performance. These parameters trade-off the dependence of the right-hand side convergence bounds on the primal and dual initial points $x^0$ and $y^0$, respectively.

• We can directly use $L_k := L_f(y^{k+1})$ in the parameter update, i.e., adaptively update

$$\beta_k := \frac{1}{L_f + L_k + \gamma^{-1} \rho_k M_g^2}$$

in Algorithm 1. Under this update, the last term in (27) of Lemma 2 diminishes with the largest possible $\beta_k$, which often improves the algorithm’s practical performance by taking more aggressive primal steps. Similarly, in Algorithm 2, we can let

$$\alpha_k := \frac{1}{L_f + (2 - \Gamma)^{-1} (L_k + \gamma^{-1} \rho_k M_g^2)}.$$

• Restarting the parameters by periodically setting, e.g., $x^0 := x_k$, and $r_k := 1$, in the context of Theorems 2, 3, 5, and 6. In this way, we can avoid the primal step-sizes $\beta_k$ and $\alpha_k$ from becoming too small after many iterations. While restarting technique can significantly boost the algorithms’ performance [65,26], we did not implement it in this section due to the lack of theoretical guarantee.

6.1 Verifying theoretical guarantees via a special case of QCQP

We consider the following problem of computing the square distance from a given point $a_0 \in \mathbb{R}^p$ to the intersection of $m$ given balls centered at $a_i$ of radius $r_i$ ($i = 1, \ldots, m$):

$$\begin{cases}
\min_{x \in \mathbb{R}^p} & \|x - a_0\|^2, \\
\text{s.t.} & \|x - a_i\|^2 \leq r_i^2, \quad i = 1, \ldots, m,
\end{cases} \quad (53)$$

where $a_i \in \mathbb{R}^p$ for $i = 0, 1, \ldots, m$, and $r_i > 0$ is a scalar for each $i = 1, \ldots, m$. Problem (53) fits the special case (CP) of our template with $f(x) := 0$, $h(x) := \|x - a_0\|^2$, $g_i(x) := \|x - a_i\|^2 - r_i^2$, and $K := \mathbb{R}^p_+$. Here, $h$ is strongly convex with $\mu_h = 2$.

We first fix the problem size as $p := 400$ and $m = 1000$. Next, we generate problem instances of (53) by drawing all entries of $a_i$'s from uniform distribution in $(-1,1)$, where $i = 0, 1, \ldots, m$. Then we define $r_i^2 := \|a_i\|^2 + \varepsilon_i$, where $\varepsilon_i > 0$ is a scalar draw from uniform distribution in $(0,1)$. Clearly, 0 is a strictly feasible solution to (53).

To test our algorithms, we generate 30 different random problem instances of the same size. For each instance, we run all six algorithmic variants up to $10^4$ iterations. Here, Algorithm 1 (v1) denotes the variant combining Algorithm 1 and parameter initialization/update rules specified in (29-30) in Theorem 1. Similarly, we call the other two variants Algorithm 1 (v2) and Algorithm 1 (v3), respectively. The three variants of Algorithm 2 are named accordingly. Without over-tuning, we simply set $\rho_0 := 5 \times 10^{-4}$ for all three variants of Algorithm 1, as well as Algorithm 2 (v1); we set $\rho_0 := 5 \times 10^{-5}$, and $M := 10^3$ for Algorithm 2 (v1) and (v2). Furthermore, we set $c := 2$ for Algorithm 1 (v3), and $c := 4$ for Algorithm 2 (v3).

The performance of six algorithmic variants is shown in Figure 1, where the relative objective residual and the relative feasibility gap, defined by

$$\frac{|F(x) - F^*|}{\max\{1, |F^*|\}} \quad \text{and} \quad \frac{\|g(x)\|}{\max\{1, \|g(x^*)\|\}},$$
are shown on the left and right, respectively. Here, \( x^* \) and \( F^* \) appearing in the figure is computed by CVX \([34]\) with the MOSEK solver \([52]\) at the highest precision. For each algorithmic variant, what we are plotting here is the theoretically convergent sequence:

- For Algorithm 1 (v1), the blue curve is based on the averaging (ergodic) sequence \( \{\bar{x}^k\} \) defined by (31) in Theorem 1.
- For Algorithm 1 (v2), the red curve is simply based on the last-iterate (non-ergodic) sequence \( \{x^k\} \).
- For Algorithm 1 (v3), the green curve is based on the so-called “best-iterate” sequence \( \{x^k\} \), defined as the minimizer of \( F(x^j) + \frac{1}{2}\|g(x^j)\| \) over \( 0 \leq j \leq k \), guided by definition of \( E \) in (52).
- The curves (black, pink, and yellow) of Algorithm 2 are similarly based on their respective iterate sequences.

Since we generate 30 different random problem instances, we use the thick line to indicate the mean value, and use the shaded area to describe the statistics range over all problems.

![Image](image-url)

**Fig. 1** Average performance of our six algorithmic variants on 30 instances of (53). Left: relative objective residual in log-scale. Right: relative feasibility residual in log-scale.

From Figure 1, we observe that Algorithm 1 indeed behaves with \( \mathcal{O}(1/k) \) convergence rate, in terms of both the objective value and the conic constraint violation. Among the three variants, Algorithm 1 (v3), with theoretical \( \min\{\mathcal{O}(1/k), o(1/k^{3/2}\log k)\} \) convergence rate, is the fastest. On the other hand, Algorithm 1 (v1), whose theoretical rate is based on the averaging iterate, has the worst performance. Moreover, since problem (53) is strongly convex, the three variants of Algorithm 2 indeed took advantages of this property, and boosted the performance to \( \mathcal{O}(1/k^2) \). Again, as theoretically predicted, the yellow curve for the “best-iterate” sequence is the best, which exhibits a empirically faster \( \min\{\mathcal{O}(1/k^2), o(1/k^{3/2}\log k)\} \) rate.

### 6.2 Convex-Concave Min-Max Game

We consider a convex-concave min-max game between two players, where Player 1 chooses her strategy \( x \in \Delta_p := \{x \in \mathbb{R}^p \mid \sum_{j=1}^p x_j = 1\} \) to minimize cost function \( F(x) \), and

---

6 We note here that the first-order methods with such constant stepsizes usually perform well empirically using the last-iterate (non-ergodic) sequence. However, only the averaging (ergodic) sequence possesses known theoretical convergence rate guarantees.
simultaneously, Player 2 chooses her strategy $y \in \Delta_n := \{y \in \mathbb{R}^n \mid \sum_{i=1}^n y_i = 1\}$ to minimize cost function $H^*$. In addition, Player 1 has to pay $\Phi(x, y)$ loss to Player 2.

Let $p = m$, and define the following functions:

$$
\begin{aligned}
 f(x) &:= \sum_{j=1}^n \log(1 + e^{a_j^\top x}), & h(x) &:= \delta_{\Delta_m}(x), \\
 F(x) &:= f(x) + h(x), & H^*(y) &:= \delta_{\Delta_n}(y), \\
 g_i(x) &:= \frac{1}{1+x_i}, & g(x) &:= (g_1(x), \ldots, g_m(x))^\top, \\
 \Phi(x, y) &:= (y, g(x)),
\end{aligned}
$$

where $A = (a_1, \ldots, a_n) \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. We can model the two-player min-max game model into the following problem, which fits well our template (SP):

$$
\min_{x \in \Delta_m} \max_{y \in \Delta_n} \left\{ \sum_{j=1}^n \log(1 + e^{a_j^\top x}) + \sum_{i=1}^m b_i y_i \right\}.
$$

Problem (55) is similar to [13, Section 4.3], but our coupling term is linear in $y$. It is easy to compute that $L_f = \|A\|^2/4$, and $L_{g_i} = 2|b_i|$, $M_{g_i} = B_{g_i} = |b_i|$ for each $i \in \{1, \ldots, m\}$.

Since $f$ in (54) is not strongly convex, we solve (55) using two variants of Algorithm 1 (v1) and Algorithm 1 (v2), both having $O(1/k)$ convergence guarantees on the primal-dual gap. Consistent with the previous subsection, for the first variant, the gap is computed based on the averaging sequences; while for the second variant, the gap $\mathcal{P}(x^k) - \mathcal{D}(y^k)$ is based on primal last-iterate sequence and the dual averaging sequence.

We compare our algorithmic variants with two existing algorithms: the Accelerated Primal-Dual (APD) algorithm proposed by [94], and the Mirror descent method in [56]. Similar to Algorithm 1 (v1), they both have the $O(1/k)$ rate on duality gap based on averaging sequences. Note that APD does not write $F(x)$ as two separate functions as in (54), thus it has to solve a non-trivial subproblem at each iteration $k$ to update $x^k$:

$$
x^{k+1} := \arg \min_{x \in \Delta_m} \left\{ \beta \sum_{j=1}^n \log(1 + e^{a_j^\top x}) + \frac{1}{2} \|x - \beta(x^k - [g'(x^k)]^\top y^{k+1})\|^2 \right\}.
$$

We have implemented restarted FISTA [78] to solve this problem, with a stopping criterion: $\|x_{k+1}^x - x_k^x\| < \varepsilon \max\{1,\|x_0^x\|\}$, where $\{x_k^x\}_{k=0}^\infty$ is the iterates for the subproblem to solve (56), and we set $\varepsilon := 10^{-6}$. On the other hand, note that Mirror descent is double-loop, and at each inner iteration, it solves two subproblems that are slightly easier to solve than (56), where we again employ a restarted FISTA routine.

To generate problem instances, we set $p = m := 1000$, and $n := 500$, and simply draw all entries of $A$ and $b$ from standard Gaussian distribution. For APD, we set the primal size as $\beta := 1/(L_b + M_b^2)$, and the dual stepsize as $\rho := 1$, as suggested in [94] Remarks 2.3 and 2.4. For Mirror descent, we set the primal-dual stepsize as $(\beta, \rho) := \left(\frac{1}{\sqrt{2L_b}}, \frac{1}{\sqrt{2(L_f + L_q)}}\right)$, as suggested in [56 eqn. (3.2)]. For both of our variants Algorithm 1 (v1) and (v2), we simply set $\gamma := \frac{1}{2}$ and $p_0 := 1$, without over-tuning.

We generate 30 problem instances, and for each instance, we run each algorithm up to 500 iterations, and the performance is shown in Figure 2. On the left, we plot the duality gap against the number of iterations; and on the right, we plot the duality against time in seconds. The curves are all based on each method’s theoretical iterations, i.e., they are
Fig. 2 Average performance of four methods on 30 instances of (55) with the problem size $p = m = 1000$ and $n = 500$. Left: duality gap against iteration counter. Right: duality gap against time.

Based on averaging (ergodic) iterates $\{\bar{x}^k\}$ and $\{\bar{y}^k\}$ for Algorithm 1 (v1), APD, and Mirror descent. However, for Algorithm 1 (v1), we use the last (non-ergodic) iterates $\{x^k\}$.

As in Section 6.1, we take the mean over all 30 instances to plot into thick curves; while we take the range of duality gaps over all 30 iterations to plot them as shaded areas. We make the following comments:

- Algorithm 1 (v1) and Mirror descent have relatively similar behavior, while our method, Algorithm 1 (v1), is still slightly faster.
- Algorithm 1 (v2) converges fastest, since the duality gap reduces below $10^{-3}$ using the smallest number of iterations and using the shortest time. Moreover, it exhibits the most oscillation, shown through both the mean curve and the shaded range area. This is a normal behavior since it uses the last-iterate sequence, and thus is less smooth than other curves, which use an averaging sequence.
- APD takes the longest time to run, since it has to solve the expensive subproblem (56). It is approximately 100 times slower than Algorithm 1 (v2) as can be seen on the time axis of the right plot of Figure 2.

In order to further solidify our conclusions, we also conducted experiment on another 30 problem instances with a larger size $(m, n) := (1500, 750)$. Indeed, the resulting performance shown in Figure 3 verified the fast speed of our proposed methods in terms of both number of iterations and the CPU time as seen in Figure 2.

7 A Brief Overview on Related Work

The convex-concave saddle-point problem (SP) presents as a unified tool to cope with several applications in convex optimization and related fields \cite{2,72,75}. Representative applications include, but not limited to, signal and image processing, game theory, robust optimization, machine learning, and most recently, generative adversarial nets (GANs) \cite{5,24,33,39,76}. Problem (SP) also covers both constrained and composite convex optimization problems as special cases. The most common solution method to solve (SP) is primal-dual methods, which are powerful, flexible, and efficient \cite{2,23}. They can be cast into a class of operator splitting schemes applied to a monotone inclusion, see, e.g., \cite{23,36,64}. Let us briefly review the most related works to the problem settings and algorithms we studied in this paper.
Convex-concave saddle-point with bilinear cost. One special case of (SP) is the bilinear case, i.e., $\Phi(x, y) = \langle Kx, y \rangle$, which has been extensively studied in the literature, including \cite{11,12,13,14,15,16,17,18,19,20,21,22,23,24,25} and the references quoted therein. Several solution methods have been proposed to solve this special case, where various variants revolve around operator splitting frameworks. For instance, as shown in \cite{64}, the well-known Chambolle-Pock as well as the Primal-Dual Hybrid Gradient algorithms can be cast equivalently to the Douglas-Rachford method \cite{18,50} applied to an appropriate reformulation of (P). Due to the equivalence between (SP) and the primal-dual pair (P)–(D), other approaches have been proposed to solve (P) when $g(x) = Kx$. For example, \cite{59,60} introduced a combination between smoothing technique and accelerated gradient methods to solve (P), which opens up a new research direction for handling large-scale applications.

Another approach relies on constrained reformulation and augmented Lagrangian framework, where the alternating direction method of multipliers (ADMM) gains the most popularity. ADMM has attracted huge attention in the last decades, and it can be viewed as special case of the Douglas-Rachford method \cite{21,50,86}. Some recent works in this direction include \cite{9,37,51,67,77}. The bilinear case has found broad applications in signal and image processing as well as machine learning, see, e.g., \cite{11,20,23,31,63}, just to name a few.

Convex-concave saddle-point with non-bilinear cost. Unlike the bilinear case, problem (SP) with non-bilinear cost $\Phi$ is more challenging to solve. One common approach is to reformulate its optimality condition into a monotone inclusion or a variational inequality as in \cite{38,40,56,61}. Based on a monotone inclusion, operator splitting techniques can be exploited to solve (SP), see, e.g., \cite{15,56,61}. Some recent work relies on extensions of the bilinear case such as Chambolle-Pock’s variant \cite{94} or Nesterov’s smoothing technique \cite{80}.

Another approach exploits the primal formulation (P) and applying inexact gradient methods, where the inner maximization problem is solved inexactly by existing methods such as Nesterov’s accelerated gradient-based algorithms \cite{43,88}. This approach often leads to complicated algorithms with double loops and several involved parameters, which are often hard to tune in practice. In terms of convergence guarantees, \cite{80,43,50,61} achieved the known best ergodic convergence rates only under certain restrictive assumptions.
**Constrained convex optimization.** One common special case of (SP) is the cone constrained convex program (1) and its special case (2). Classical approaches for solving (1) often rely on sequential linear and quadratic programming, interior-point, penalty, and augmented Lagrangian-based methods, see, e.g., [6,10,27,28,62]. Recently, [46] studied a class of quadratic penalty methods to solve conic instances of (1). The authors combined classical quadratic penalty methods with Nesterov’s accelerated schemes to develop two-loop algorithms and then characterized their worst-case iteration-complexity. These authors then extended this approach to augmented Lagrangian methods in [42]. Many authors studied instances of (1) by exploiting duality framework, smoothing techniques, and augmented Lagrangian schemes combining with Nesterov’s accelerated ideas, see, e.g., [29,51,53,55,70,83,85]. Subgradient and mirror descent-based methods have also been studied in the literature starting from [69], and most recently in [3,4,43]. ADMM, which has gained its great popularity in the past decades, see, e.g., [9,30,37,49,66,67,77], is another popular approach to solve (1). However, these works have mainly focused on the affine constraints, and rarely tackled the general form (1) with nonlinear cone constraints.

Hitherto, efficient first-order methods for solving (1) with nonlinear functional constraints remain limited. Along this line, recent works include [45,46,47,90,91,92]. While [46,47] rely on bundle methods, [90,91,92] utilize the augmented Lagrangian framework. The authors in these works have characterized worst-case iteration complexity of their methods. The most notable convergence rate in these works is $O(1/k)$ in ergodic sense. [45] is a very recent preprint, where the authors propose two double loop augmented Lagrangian algorithms to solve (2), and achieves $O(\log(1/\varepsilon)/\sqrt{k})$ worst-case iteration-complexity to reach an $\varepsilon$-accuracy iterate in non-ergodic sense.

**8 Conclusions**

In this paper, we have studied a class of convex-concave saddle-point problems (SP) involving non-bilinear coupling function. We have developed two novel primal-dual algorithms to solve (SP) and its primal-dual pair reformulation (1)-(4). Our algorithms have single-loop, where all the parameters are updated with explicit formulas. The first algorithm, Algorithm 1 achieves both ergodic and semi-ergodic optimal $O(1/k)$ convergence rates on the duality gap, and can be boosted up to $\min\{O(1/k), g(1/(k\sqrt{\log k})\}$ non-ergodic primal convergence rate. Under strong convexity of $F$, our second algorithm, Algorithm 2 can be accelerated to have $O(1/k^2)$ and $\min\{O(1/k^2), g(1/(k^2\sqrt{\log k})\}$ convergence rates. To the best of our knowledge, this is the first algorithms that achieve such non-ergodic as well as fast rates for non-bilinear saddle-point problems. We believe that our results can be further extended to a general non-bilinear function $\Phi$ under appropriate assumptions.

**A Preliminary Lemmas**

This appendix provides the full proof of Lemma 1 and some elementary results, which will be used for our analysis in the sequel.

**A.1 The proof of Lemma 1** key property of the augmented Lagrangian term

By definition of $\Delta_\rho$ in (17), we can use the definition of $\phi_\rho(x,y)$ in (16) and its partial gradients w.r.t. $x$ and $s$ in (17) to explicitly write $\Delta_\rho$ as

$$
\Delta_\rho(\hat{x}, \hat{s}; x, s, y) = \langle y, [g(\hat{x}) + \hat{s}] - [g(x) + s]\rangle + \frac{\rho}{2} \left[\|g(\hat{x}) + \hat{s}\|^2 - \|g(x) + s\|^2\right] \\
- \langle y + \rho[g(x) + s], [g'(x)](\hat{x} - x) + (\hat{s} - s)\rangle \\
= \langle y + \rho[g(x) + s], g(\hat{x}) - g(x) - [g'(x)](\hat{x} - x)\rangle \\
+ \frac{\rho}{2} \|g(\hat{x}) + \hat{s}\|^2 - \|g(x) + s\|^2. 
$$

(57)
By the $L_p(\cdot)$-smoothness of $\nabla_x \Phi(x, \cdot) = \langle g'(x), \cdot \rangle$ w.r.t. $x$, and that $y + \rho (g(x) + s)$ is in $\text{dom}(H^*)$, we can apply (7) with $y \leftarrow y + \rho (g(x) + s)$ to get

$$0 \leq (y + \rho (g(x) + s), g(\hat{x}) - g(x) - [g'(x)](\hat{x} - x)) \leq \frac{L_y(y + \rho (g(x) + s))}{2} \|\hat{x} - x\|^2. \tag{58}$$

Combining (57) and (58), we immediately get

$$\begin{cases}
\Delta_p(\hat{x}, \hat{s}; x, s, y) & \geq \frac{\rho}{2} \|g(\hat{x}) + \hat{s} - [g(x) + s]\|^2 \\
\Delta_p(\hat{x}, \hat{s}; x, s, y) & \leq \frac{\rho}{2} \|g(\hat{x}) + \hat{s} - [g(x) + s]\|^2 + \frac{L_y(y + \rho (g(x) + s))}{2} \|\hat{x} - x\|^2,
\end{cases}$$

which is exactly (19). \qed

### A.2 Additional mathematical tools

We will repeatedly use the following elementary facts in Lemma 4 in our analysis.

**Lemma 4 ([84, Lemma 18])** The following statements hold:

(a) For any $u, v, w \in \mathbb{R}^p$ and $t_1, t_2 \in \mathbb{R}$ with $t_1 + t_2 \neq 0$, it holds that

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

(b) Let $\{u_k\}$ be a nonnegative sequence. If $\sum_{k=1}^{\infty} u_k < \infty$, then $\lim inf_{k \to \infty} (k \log k) u_k = 0$.

(c) Let $\{u_k\}$ and $\{v_k\}$ be two nonnegative sequences and $t_1, t_2 > 0$ be two constants.

(i) If $\lim inf_{k \to \infty} (k \log k)(u_k + t_1 v_k^2) = 0$, then $\lim inf_{k \to \infty} k \sqrt{\log k} (u_k + t_2 v_k^2) = 0$.

(ii) If $\lim inf_{k \to \infty} (k^2 \log k)(u_k + t_1 k^2 v_k^2) = 0$, then $\lim inf_{k \to \infty} k^2 \sqrt{\log k} (u_k + t_2 v_k^2) = 0$.

### B Technical Proofs in Section 3: General Convex-Concave Case

This appendix provides the full proofs of technical results in Section 3.

#### B.1 Proof of Lemma 2: One-iteration analysis of Algorithm 1

For readability, the full proof of Lemma 2 is broken into Lemma 5 and its own proof. The proof of Lemma 5 is right after the proof of Lemma 2.

**Lemma 5** Let $\{(x^k, \tilde{y}^k)\}$ be generated by scheme (25), and $\{s^k\}$ be given by (20). Then, for any $(x, s) \in \mathbb{R}^p \times \mathbb{R}^m$, one has

$$\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \tilde{y}^k) \leq \mathcal{L}_{\rho_k}(x, s, \tilde{y}^k) + \frac{1}{\rho_k} (x^{k+1} - \tilde{x}^k, x - x^{k+1}) - \frac{\rho_k}{2} \|[g(x) + s] - [g(\tilde{x}^k) + s^{k+1}]\|^2 + \frac{1}{2} (L_k + L_f + \rho_k M_f^2) \|x^{k+1} - \tilde{x}^k\|^2. \tag{59}$$

**Proof (Proof of Lemma 2)** Plugging $(x, s) := (x^k, s^k)$ in (59) of Lemma 5, we obtain

$$\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \tilde{y}^k) \leq \mathcal{L}_{\rho_k}(x^k, s^k, \tilde{y}^k) + \frac{1}{\rho_k} (x^{k+1} - \tilde{x}^k, x^k - x^{k+1}) - \frac{\rho_k}{2} \|[g(x^k) + s^k] - [g(\tilde{x}^k) + s^{k+1}]\|^2 + \frac{L_k + L_f + \rho_k M_f^2}{2} \|x^{k+1} - \tilde{x}^k\|^2.$$

Now, multiplying the above estimate above by $1 - \tau_k \in (0, 1)$, and (59) by $\tau_k \in (0, 1]$, and then summing up the results, we get

$$\begin{align*}
\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \tilde{y}^k) & \leq (1 - \tau_k) \mathcal{L}_{\rho_k}(x^k, s^k, \tilde{y}^k) + \tau_k \mathcal{L}(x, s, \tilde{y}^k) + \frac{\tau_k}{\rho_k} (x^{k+1} - \tilde{x}^k, x - x^{k+1}) \\
& \quad - \frac{(1 - \tau_k) \rho_k}{2} \|[g(x^k) + s^k] - [g(\tilde{x}^k) + s^{k+1}]\|^2 - \frac{\tau_k \rho_k}{2} \|[g(x) + s] - [g(\tilde{x}^k) + s^{k+1}]\|^2 + \frac{L_k + L_f + \rho_k M_f^2}{2} \|x^{k+1} - \tilde{x}^k\|^2,
\end{align*} \tag{60}$$
where we have used \( \tau_k(\bar{x}^{k+1} - \bar{x}^k) = x^{k+1} - \bar{x}^k \) and \((1 - \tau_k)x^k + \tau_k x - x^{k+1} = \tau_k(x - \bar{x}^{k+1}) \) derived from the definition of \( \bar{x}^k \) in (26).

Next, by the definition of \( \mathcal{L}_{\rho_k} \), for any \( y \in \mathcal{B}_k \), we have
\[
\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y) - (1 - \tau_k)\mathcal{L}_{\rho_k}(x^k, s^k, y)
= \mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) - (1 - \tau_k)\mathcal{L}_{\rho_k}(x^k, s^k, \hat{y}^k) + \langle y - \hat{y}^k, [g(x^{k+1}) + s^{k+1}] - (1 - \tau_k)[g(x^k) + s^k]\rangle. \quad (=: T_1)
\]

To analyze the last term, \( T_1 \), in (61), we denote
\[
u^{k+1} := \eta_k \left( [g(x^{k+1}) + s^{k+1}] - (1 - \tau_k)[g(x^k) + s^k] \right) \eta_k [\Theta_{k+1} - (1 - \tau_k)\Theta_k]. \quad (62)
\]

Then, by the update of \( \hat{y}^{k+1} \) in (25) and the fact that \( y \in \mathcal{B}_k \), we can use the non-expansive property of the projection \( \text{proj}_{\mathcal{B}_k} \) to get
\[
\| \hat{y}^{k+1} - y \| = \| \text{proj}_{\mathcal{B}_k}(\hat{y}^k + \nu^{k+1}) - \text{proj}_{\mathcal{B}_k}(y) \| \leq \| \hat{y}^k + \nu^{k+1} - y \|.
\]

Therefore, \( T_1 \) becomes
\[
T_1 \overset{(62)}{=} \frac{1}{\eta_k} \langle y - \hat{y}^k, \nu^{k+1} \rangle = \frac{1}{\eta_k} \langle \hat{y}^k - y, (y - \nu^{k+1}) - y \rangle = \frac{1}{\eta_k} \left[ \| \hat{y}^k - y \|^2 + \| y - \nu^{k+1} - y \|^2 - \| \hat{y}^k + \nu^{k+1} - y \|^2 \right] \leq \frac{1}{\eta_k} \left[ \| \hat{y}^2 - y \|^2 - \| \hat{y}^{k+1} - y \|^2 + \| \nu^{k+1} \|^2 \right]. \quad (63)
\]

Substituting (63) into (61), and then combining with (60), we can further derive
\[
\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y) \leq (1 - \tau_k)\mathcal{L}_{\rho_k}(x^k, s^k, y) + \tau_k \mathcal{L}(x, s, \hat{y}^k - \tau_k \rho_k \|g(x) + s\| - [g(\hat{x}^k) + s^{k+1}] \|^2 \quad (=: T_2)
+ \frac{\tau_k}{\rho_k}[\| \hat{y}^k - y \|^2 - \| \hat{y}^{k+1} - y \|^2] + \frac{L_k}{2}\| \nu^{k+1} \|^2 \quad (=: T_3)
\]

We now estimate the terms \( T_2, T_3, \) and \( T_4 \) above. It is easy to see that
\[
T_2 = (1 - \tau_k) \left[ \mathcal{L}_{\rho_k - 1}(x^k, s^k, y) + \frac{\rho_k - \rho_{k-1}}{2} \| g(x^k) + s^k \| \right]. \quad (65)
\]

By definition of \( \hat{y}^{k+1} \) in (26), we have
\[
T_3 = \mathcal{L}(x, s, \hat{y}^{k+1}) - (1 - \tau_k)\mathcal{L}(x, s, \hat{y}^k) - \frac{\tau_k \rho_k}{2} \| g(\hat{x}^k) + s^{k+1} \|^2. \quad (66)
\]

Using the relation \( \bar{x}^{k+1} - \bar{x}^k = \frac{1}{\tau_k} (x^{k+1} - \bar{x}^k) \), we further have
\[
T_4 = \frac{\tau_k^2}{2\rho_k} (\| \bar{x}^{k+1} - x \|^2 - \| \bar{x}^{k+1} - x \|^2 - \| \bar{x}^{k+1} - \bar{x}^k \|^2) + \frac{L_k + L_f + \rho_k M_g^2}{2} \| x^{k+1} - \bar{x}^k \|^2
\]
\[
= \frac{\tau_k^2}{2\rho_k} (\| \bar{x}^{k+1} - x \|^2 - \| \bar{x}^{k+1} - x \|^2) - \frac{L_k}{2\rho_k} \left( \frac{1}{\tau_k} L_k - L_f - \rho_k M_g^2 \right) \| x^{k+1} - \bar{x}^k \|. \quad (67)
\]
Substituting (65)-(67) into (64), we get
\[
\mathcal{L}_{pk}(x^{k+1}, s^{k+1}, y) = \mathcal{L}(x, s, \tilde{y}^{k+1}) \leq \left(1 - \tau_k\right)\left[\mathcal{L}_{pk-1}(x^k, s^k, y) - \mathcal{L}(x, s, \tilde{y}^k)\right] \\
+ \frac{\eta^2}{4\tau_k}(\|\tilde{x}^k - x\|^2 - \|\tilde{x}^{k+1} - x\|^2) + \frac{1}{2\tau_k}(\|\tilde{y}^k - y\|^2 - \|\tilde{y}^{k+1} - y\|^2) \\
- \frac{1}{\tau_k} \left(\tilde{L}_k - L_f - \rho_k M^2_\beta\right)\|x^{k+1} - \tilde{x}^k\|^2 + \mathcal{T}_5,
\]
where
\[
\mathcal{T}_5 := \frac{1}{2\eta_k} \|u^{k+1}\|^2 + \frac{(1 - \tau_k)(\rho_k - \rho_{k+1})}{2} \|g(x^k) + s^k\|^2 - \frac{\tau_k}{2\eta_k} \|g(\tilde{x}^k) + r^{k+1}\|^2 \\
- \frac{(1 - \tau_k)\rho_k}{2} \|[g(x^k) + s^k] - [g(\tilde{x}^k) + s^{k+1}]\|^2 \\
= \frac{1}{2\eta_k} \|u^{k+1}\|^2 - \frac{\eta_k}{2} \|[g(x^k) + s^k] - [g(\tilde{x}^k) + s^{k+1}]\|^2 \\
- \frac{(1 - \tau_k)\rho_k}{2} \|[g(x^k) + s^k] - [g(\tilde{x}^k) + s^{k+1}]\|^2 \\
\leq \frac{\rho_k \eta_k}{2(\rho_k - \rho_{k+1})} \|x^{k+1} - \tilde{x}^k\|^2 - \frac{(1 - \tau_k)\rho_k}{2(\rho_k - \rho_{k+1})} \|g(x^k) + s^k\|^2 \\
\leq \frac{\rho_k \eta_k}{2(\rho_k - \rho_{k+1})} \|x^{k+1} - \tilde{x}^k\|^2 - \frac{(1 - \tau_k)\rho_k}{2(\rho_k - \rho_{k+1})} \|g(x^k) + s^k\|^2,
\]
where in the first inequality (second to last line) above, we have used Lemma 4(a) and \(\rho_k > \eta_k\). Finally, substituting (69) into (68), we eventually get
\[
\mathcal{L}_{pk}(x^{k+1}, s^{k+1}, y) - \mathcal{L}(x, s, \tilde{y}^{k+1}) \leq \left(1 - \tau_k\right)\left[\mathcal{L}_{pk-1}(x^k, s^k, y) - \mathcal{L}(x, s, \tilde{y}^k)\right] \\
+ \frac{\eta^2}{4\tau_k}(\|\tilde{x}^k - x\|^2 - \|\tilde{x}^{k+1} - x\|^2) + \frac{1}{2\tau_k}(\|\tilde{y}^k - y\|^2 - \|\tilde{y}^{k+1} - y\|^2) \\
- \frac{1}{\tau_k} \left(\tilde{L}_k - L_f - \rho_k M^2_\beta\right)\|x^{k+1} - \tilde{x}^k\|^2 \\
- \frac{1}{\tau_k} \left(\rho_k - (1 - \tau_k)\rho_k\right)\|g(x^k) + s^k\|^2,
\]
which is exactly (27).

\[\square\]

**Proof (Proof of Lemma 5)** First, the optimality condition of the \(x^{k+1}\)-subproblem in the second line of (25), which is equivalent to (21), can be written as
\[
0 = \beta_k \nabla h(x^{k+1}) + \beta_k \nabla f(x^k) + \beta_k \nabla x \phi_{\rho_k}(\tilde{x}^k, x^{k+1}, \tilde{y}^k) + x^{k+1} - \tilde{x}^k,
\]
for some \(\nabla h(x^{k+1}) \in \partial h(x^{k+1})\). Next, by convexity of \(h\) and \(L_f\)-smoothness of \(f\), for any \(x \in \text{dom}(P)\) we have
\[
\begin{align*}
 h(x^{k+1}) & \leq h(x) + \langle \nabla h(x^{k+1}), x^{k+1} - x \rangle, \\
 f(x^{k+1}) & \leq f(x) + \langle \nabla f(\tilde{x}^k), x^{k+1} - x \rangle + \frac{L_f}{2}\|x^{k+1} - \tilde{x}^k\|^2.
\end{align*}
\]
Combining these two inequalities and then using (70) and \(F := f + h\), we can derive
\[
F(x^{k+1}) \leq F(x) - \langle \nabla x \phi_{\rho_k}(\tilde{x}^k, x^{k+1}, y^k), x^{k+1} - x \rangle + \frac{L_f}{2}\|x^{k+1} - \tilde{x}^k\|^2 \\
- \frac{1}{\beta_k} \langle x^{k+1} - \tilde{x}^k, x^{k+1} - x \rangle.
\]
Similarly, by the \(s^{k+1}\)-subproblem (20) and the convexity of \(H\), we have
\[
H(-s^{k+1}) \leq H(-s) + \langle \nabla s \phi_{\rho_k}(\tilde{x}^k, s^{k+1}, y^k), -s + s^{k+1} \rangle.
\]
Furthermore, combining (24) and the definition of \(L_k\) in (26), we have \(L_k = L_g(\hat{y}^k + \rho_k [g(\hat{x}^k) + s^{k+1}])\). Thus we can use (19) in Lemma 1 for any \((x, s) \in \mathbb{R}^p \times \mathbb{R}^m\), to get

\[
\begin{align*}
\phi_{p_k}(x^{k+1}, s^{k+1}, \hat{y}^k) &\leq \phi_{p_k}(\hat{x}^k, s^{k+1}, \hat{y}^k) + \langle \nabla_x \phi_{p_k}(\hat{x}^k, s^{k+1}, \hat{y}^k), x^{k+1} - \hat{x}^k \rangle \\
&\quad + \frac{\rho_k}{2} \|g(x^{k+1}) - g(\hat{x}^k)\|^2 + \frac{L_k}{2} \|s^{k+1} - \hat{s}\|^2,
\end{align*}
\]

\[
\begin{align*}
\phi_{p_k}(x, s, \hat{y}^k) &\geq \phi_{p_k}(\hat{x}^k, s^{k+1}, \hat{y}^k) + \langle \nabla_x \phi_{p_k}(\hat{x}^k, s^{k+1}, \hat{y}^k), x - \hat{x}^k \rangle \\
&\quad + \langle \nabla_s \phi_{p_k}(\hat{x}^k, s^{k+1}, \hat{y}^k), s - s^{k+1} \rangle \\
&\quad + \frac{\rho_k}{2} \|g(x) + s\| - \|g(\hat{x}^k) + s^{k+1}\|^2.
\end{align*}
\]

By (6), the above two inequalities imply

\[
\phi_{p_k}(x^{k+1}, s^{k+1}, \hat{y}^k) \leq \phi_{p_k}(x, s, \hat{y}^k) + \langle \nabla_x \phi_{p_k}(\hat{x}^k, s^{k+1}, \hat{y}^k), x^{k+1} - x \rangle \\
+ \langle \nabla_s \phi_{p_k}(\hat{x}^k, s^{k+1}, \hat{y}^k), s^{k+1} - s \rangle
+ \frac{L_k + \rho_k M^2}{2} \|s^{k+1} - \hat{s}\|^2
\]

\[
- \frac{\rho_k}{2} \|g(x) + s\| - \|g(\hat{x}^k) + s^{k+1}\|^2 \tag{73}
\]

Now, combining (71)-(73), we can derive

\[
\mathcal{L}_{p_k}(x^{k+1}, s^{k+1}, \hat{y}^k) = F(x^{k+1}) + H(-s^{k+1}) + \phi_{p_k}(x^{k+1}, s^{k+1}, \hat{y}^k)
\]

\[
\leq F(x) + H(-s) + \phi_{p_k}(x, s, \hat{y}^k) + \frac{1}{\rho_k} \langle x^{k+1} - \hat{x}^k, x - x^{k+1} \rangle
\]

\[
- \frac{\rho_k}{2} \|g(x) + s\| - \|g(\hat{x}^k) + s^{k+1}\|^2 + \frac{L_k + L_f + \rho_k M^2}{2} \|s^{k+1} - \hat{s}\|^2,
\]

which proves (59).

\[
\square
\]

B.2 Proof of Theorem 1: \(O(1/k)\) ergodic convergence rate of Algorithm 1

Let us first show the boundedness of \(\{\|\hat{y}^k - y^*\|\}\) and \(\{\|x^k - x^*\|\}\), whose proof will be given right after the proof of Theorem 1

Lemma 6 Let \(\{(x^k, \hat{y}^k)\}\) be generated by Algorithm 1, where the parameters, including \(\rho\) and \(C\), are set as in (29) and (50). Then, for all \(k \in \mathbb{N}\), we have

\[
L_g[\|y^*\| + \|\hat{y}^k - y^*\| + \rho M\|x^k - x^*\|] \leq \rho C. \tag{74}
\]

Proof (Proof of Theorem 1) By (74) of Lemma 6, we can follow the same lines as (76) and (77) to show that \(\frac{1}{\beta} - L_f - L_k - \frac{\rho M^2}{\rho - \eta} \geq 0\). Therefore, similar to (78), for any \(y \in \mathbb{R}^m\) and any \(j \in \mathbb{N}\), we have

\[
\mathcal{L}(x^{j+1}, s^{j+1}, y) - \mathcal{L}(x, s, \hat{y}^j) \leq \frac{1}{2\eta} \left[\|x^{j+1} - x\|^2 - \|x^{j+1} - x\|^2\right]
\]

\[
+ \frac{1}{2\eta} \left[\|\hat{y}^j - y\|^2 - \|\hat{y}^j - y\|^2\right].
\]

Summing up this inequality from \(j := 0\) to \(j := k - 1\), we get

\[
\sum_{j=0}^{k-1} [\mathcal{L}(x^{j+1}, s^{j+1}, y) - \mathcal{L}(x, s, \hat{y}^j)] \leq \frac{1}{2} \left[\frac{1}{\beta\eta}[\|x^0 - x\|^2 + \frac{1}{\eta}[\|y^0 - y\|^2]ight] = \frac{R^2_0(x, y)}{2}.
\]
Dividing the above inequality by \( k \geq 1 \), and using the convexity of \( \mathcal{L} \) in \( x \) and \( -s \), and its concavity in \( y \), with \( \bar{x}^k \) and \( \bar{y}^k \) defined in (31) and \( \bar{s}^k := \frac{1}{k} \sum_{j=1}^{k} s^j \), we get

\[
\mathcal{L}(\bar{x}^k, \bar{s}^k, y) - \mathcal{L}(x, s, \bar{y}^k) \leq \frac{1}{k} \sum_{j=1}^{k} [\mathcal{L}(x^j, s^j, y) - \mathcal{L}(x, s, y^j)] \leq \frac{R^2_0(x, y)}{2k}. \tag{75}
\]

Now, by (33), we have \( \bar{L}(\bar{x}^k, y) \leq \bar{L}(\bar{x}^k, \bar{s}^k, y) \) and \( \bar{L}(x, \bar{y}^k) = \mathcal{L}(x, \bar{s}^k, \bar{y}^k) \) for \( \bar{s}^k \in -\partial H^*(\bar{y}^k) \). Hence, \( \bar{L}(\bar{x}^k, y) - \bar{L}(x, \bar{y}^k) \leq \bar{L}(\bar{x}^k, \bar{s}^k, y) - \mathcal{L}(x, \bar{s}^k, \bar{y}^k) \). Substituting \( s := \bar{s}^k \) and this inequality into (73), we obtain \( \bar{L}(\bar{x}^k, y) - \bar{L}(x, \bar{y}^k) \leq \frac{R^2_0(x, y)}{2k} \). Taking the supremum on both sides of this estimate over \( \mathcal{X} \times \mathcal{Y} \) and recalling the definition of \( \mathcal{G}_{\mathcal{X} \times \mathcal{Y}} \) in (9), we prove the first assertion of (32).

Next, if \( H \) is \( M_H \)-Lipschitz continuous, then we let \( \bar{y}^k := \frac{M_H}{\|g(x^k)\|} [g(\bar{x}^k) + \bar{s}^k] \), and substitute \( (x, s, y) := (x^*, s^*, \bar{y}^*) \) in (75) to get

\[
\mathcal{P}(\bar{x}^k) - \mathcal{P}^* \leq F(\bar{x}^k) + H(g(\bar{x}^k)) - \mathcal{P}^* \leq F(\bar{x}^k) + H(-\bar{s}^k) + M_H |g(\bar{x}^k)| + M_H |\bar{s}^k| - \mathcal{P}^* \\
= F(\bar{x}^k) + H(-\bar{s}^k) + \langle \bar{g}^k, g(\bar{x}^k) + \bar{s}^k \rangle - \mathcal{P}^* \leq \mathcal{L}(\bar{x}^k, \bar{s}^k, \bar{y}^*) - \mathcal{P}^* \leq \frac{R^2_0(x^*, \bar{y}^*)}{2k}. \tag{76}
\]

Using \( \|y^0 - \bar{y}^k\|^2 \leq (\|y^0\| + \|\bar{y}^k\|)^2 = (\|y^0\| + M_H)^2 \) to upper bound \( R^2_0(x^*, \bar{y}^k) \) in the last estimate, we obtain the second assertion of (32).

On the other hand, let \( \bar{x}^k \) satisfy \( 0 \in g(\bar{x}^k) + \partial F(\bar{x}^k) \), then by the form of (D), we have \( D(\bar{y}^k) = \bar{L}(\bar{x}^k, \bar{y}^k) = \mathcal{L}(\bar{x}^k, \bar{s}^k, \bar{y}^k) \) for \( \bar{s}^k \in -\partial H^*(\bar{y}^k) \). Moreover, notice that \( \mathcal{D}^* = \mathcal{L}(x^*, s^*, y^*) \) in (14). Therefore, substituting \( (x, s, y) := (\bar{x}^k, \bar{s}^k, \bar{y}^*) \) into (75), we can derive

\[
\mathcal{D}^* - D(\bar{y}^k) \leq \mathcal{L}(\bar{x}^k, \bar{s}^k, y^*) - \mathcal{L}(\bar{x}^k, \bar{s}^k, \bar{y}^k) \leq \frac{R^2_0(x^*, \bar{y}^*)}{2k}. \tag{77}
\]

By \( 0 \in g(\bar{x}^k)^T \bar{y}^k + \partial F(\bar{x}^k) \), we have \( \bar{x}^k \in \partial F^*(-g(\bar{x}^k)^T \bar{y}^k) \). If \( F^* \) is \( M_F \)-Lipschitz continuous, then \( \|\bar{x}^k\| = \|\nabla F^*(-\bar{g}^k)^T \bar{y}^k\| \leq M_F, \) thus \( \|x^0 - \bar{x}^k\|^2 \leq (\|x^0\| + M_F)^2 \). Substituting this into \( R^2_0(\bar{x}^k, y^*) \) of the last inequality leads to the third assertion of (32).

Finally, combining the second and third assertions of (32), we have immediately proved the last assertion on the primal-dual gap \( \mathcal{P}(\bar{x}^k) - D(\bar{y}^k) \).

**Proof (Proof of Lemma 9)** We prove (71) by induction. For \( k = 0 \), (71) holds due to the choice of \( C \) in (29). Suppose that (71) holds for all \( k \in \{0, 1, \ldots, K\} \) for some \( K \geq 0 \), i.e., \( L_g(\|y^k - y^*\| + \rho M_g \|x^k - x^*\|) \leq \rho C \). We now prove that (74) also holds for \( K + 1 \). Indeed, using \( y^* = \text{prox}_{\rho H^*} (y^* + \rho g(x^*)) \) from (12), for \( 0 \leq k \leq K \) we have

\[
L_k \leq L_g(y^{k+1}) \leq L_g \left( \text{prox}_{\rho H^*} (\bar{y}^k + \rho g(x^k)) \right) \leq L_g \left( \text{prox}_{\rho H^*} (\bar{y}^k + \rho g(x^k)) - \text{prox}_{\rho H^*} (y^* + \rho g(x^*)) + y^* \right) \leq L_g \left( \|y^* - \text{prox}_{\rho H^*} (\bar{y}^k + \rho g(x^k))\| + \|y^*\| \right) \leq C \tag{76}
\]

where in the third line we applied Assumption (2)(c), in the fourth line we used the non-expansiveness of proximal operators, and the last inequality is due to induction assumption.
By definitions of $\beta$ and $\eta$ in (29), for $0 \leq k \leq K$, we have
\[
\frac{1}{\beta} - L_k - L_f - \frac{\rho^2 M_g^2}{\rho - \eta} \geq \frac{\gamma L_f + \rho (\gamma C + M_g^2)}{\gamma} - \rho C - L_f - \frac{\rho^2 M_g^2}{\gamma \rho} = 0.
\] (77)

Using this estimate, we substitute $\tau_k := 1$ and $\tilde{x}^k := x^k$ into (27) of Lemma 2 to obtain for any $(x, s, y) \in \mathbb{R}^p \times \mathbb{R}^m \times \mathbb{R}^m$ that
\[
L_{\rho}(x^{k+1}, s^{k+1}, y) - \mathcal{L}(x, s, y^{k+1}) \leq \frac{1}{2\beta} \left( \|x^k - x\|^2 - \|x^{k+1} - x\|^2 \right) + \frac{1}{\eta} \left( \|y^k - y\|^2 - \|y^{k+1} - y\|^2 \right).
\] (78)

By (14), we have $L_{\rho}(x^{k+1}, s^{k+1}, y^*) - \mathcal{L}(x^*, s^*, y^k) \geq 0$. Hence, (78) implies that
\[
\frac{1}{\beta} \|x^{k+1} - x^*\|^2 + \frac{1}{\eta} \|y^{k+1} - y^*\|^2 \leq \frac{1}{\beta} \|x^k - x^*\|^2 + \frac{1}{\eta} \|y^k - y^*\|^2.
\]

Since the above inequality holds for all $0 \leq k \leq K$, we can show that
\[
\frac{1}{\beta} \|x^{K+1} - x^*\|^2 + \frac{1}{\eta} \|y^{K+1} - y^*\|^2 \leq \frac{1}{\beta} \|x^K - x^*\|^2 + \frac{1}{\eta} \|y^K - y^*\|^2 \leq \frac{1}{\beta} \|x^0 - x^*\|^2 + \frac{1}{\eta} \|y^0 - y^*\|^2 \leq \mathcal{R}_0^2(x^*, y^*).
\]

The last inequality leads to $\|x^{K+1} - x^*\| \leq \sqrt{\beta} \mathcal{R}_0(x^*, y^*)$ and $\|y^{K+1} - y^*\| \leq \sqrt{\eta} \mathcal{R}_0(x^*, y^*)$. Using these bounds and (29), we can derive
\[
L_g \left[ \|y^*\| + \|y^{K+1} - y^*\| + \rho M_g \|x^{K+1} - x^*\| \right] \leq L_g \left[ \|y^*\| + (\sqrt{\eta} + \sqrt{\beta} M_g) \mathcal{R}_0(x^*, y^*) \right] \leq \rho C.
\] (28)

Hence, we prove that (74) also holds for $K + 1$. By induction, it holds for all $k \in \mathbb{N}$. \hfill \Box

B.3 Proof of Remark 1: Initializing parameters in Theorem 1
For simplicity, we set $\rho := 1$ and $\gamma := \frac{1}{2}$. Substituting them into the expression of $\beta, \eta$, and $\mathcal{R}_0^2(x^*, y^*)$, we get
\[
\beta := \frac{1}{L_f + C + 2M_g^2}, \quad \eta := \frac{1}{2}, \quad \text{and} \quad \mathcal{R}_0^2(x^*, y^*) \leq (L_f + C + 2M_g^2 + 2)D^2,
\] (79)
where $D \geq \max\{\|x_0 - x^*\|, \|y_0 - y^*\|, \|y^*\|\}$ as defined in Remark 1. Substituting the above expressions for $\rho, \beta, \eta$, and $\mathcal{R}_0^2(x^*, y^*)$ into the second line of (29), we only need the following inequality in order for (29) to hold:
\[
1 + \sqrt{\frac{L_f + C}{2} + M_g^2 + 1} + M_g \sqrt{1 + \frac{2}{L_f + C + 2M_g^2}} = 1 + \left( \frac{1}{\sqrt{2}} + \frac{M_g}{\sqrt{L_f + C + 2M_g^2}} \right) \sqrt{L_f + C + 2M_g^2 + 2} \leq \frac{C}{L_g D}.
\] (80)

Let
\[
C \geq L_f + 2M_g^2 + 2,
\] (81)
then $\frac{L_f + C}{2} + M_g^2 + 1 \leq C$ and $\frac{2}{L_f + C + 2M_g^2} \leq 3$. Substituting them into the left-hand-side above, we only need the following inequality in order for (80) to hold:
\[
1 + \sqrt{C} + 2M_g \leq \frac{C}{L_g D}.
\]
which can be solved as
\[ \sqrt{C} \geq \sqrt{L_gD(L_gD + 4M_g + 2)}, \]
where we have used \( \frac{t_1 + t_2}{2} \leq \sqrt{\frac{t_1^2 + t_2^2}{2}} \) to simplify the expression. Combining (81) and (82), we finally get (33).

**B.4 Proof of Theorem 2:** \( O(1/k) \) semi-ergodic convergence rate of Algorithm 1

Before proving Theorem 2, let us use the lemma below to bound the term \( \{L_k/\rho_k\} \). Its proof is right after the proof of Theorem 2.

**Lemma 7** Let \( \{(x^k, \hat{y}^k)\} \) be generated by Algorithm 1 where the parameters, including \( \rho_k \) and \( \gamma \), are defined in (34) and (35). Let \( B_g \) and \( s_* \) be defined in Theorem 2. Then for \( k \in \mathbb{N} \),
\[
\frac{\|\hat{y}^k\|}{\rho_k} \leq \frac{1}{\gamma} \left[ \frac{\|\nabla y\|}{\rho_0} + 2(1 - \gamma)(B_g + \|s_*\|) \right].
\]

**Proof (Proof of Theorem 2)** From the assumption that 0 \( \in \text{dom}(\partial H^*) \), we have \( \partial H^*(0) \neq 0 \). Take an arbitrary \( y_s \in \partial H^*(0) \), then \( y_s = \text{prox}_{\rho_k H^*}(y_s) \) for any \( \rho_k > 0 \). Using this relation, and the non-expansiveness of \( \text{prox}_{\rho_k H^*} \), we can prove that
\[
L_k \leq L_g\|\text{prox}_{\rho_k H^*}(\hat{y}^k + \rho_k g(\hat{x}^k))\| = L_g\|\text{prox}_{\rho_k H^*}(\hat{y}^k + \rho_k g(\hat{x}^k)) - \text{prox}_{\rho_k H^*}(y_s) + y_s\| \leq L_g(\|y_s\| + \|\hat{y}^k + \rho_k g(\hat{x}^k)\|)
\]
(83)
\[
\leq L_g \left( \frac{\|y_s\|}{\rho_k} + \rho_k B_g + \frac{\rho_k}{\gamma} \left[ \frac{\|\nabla y\|}{\rho_0} + 2(1 - \gamma)(B_g + \|s_*\|) \right] \right).
\]
Therefore, by the update rule of \( \beta_k \) and \( \eta_k \) in (35), and (84), we can easily show that
\[
\frac{1}{\beta_k} - L_k - \frac{\rho_k^2 M^2}{\rho_k - \eta_k} \geq \frac{\rho_k}{\gamma} \left[ \frac{\|\nabla y\|}{\rho_0} + (2 - \gamma)B_g + 2(1 - \gamma)\|s_*\| \right] + M^2
\]
(85)
\[
L_f + 2L_g\|y_s\| - L_g \left[ 2\|y_s\| + \frac{\rho_k}{\gamma} \left( \frac{\|\nabla y\|}{\rho_0} + (2 - \gamma)B_g + 2(1 - \gamma)\|s_*\| \right) \right] - L_f - \frac{\rho_k M^2}{\gamma} = \frac{\rho_k}{\gamma} \left[ \frac{\|\nabla y\|}{\rho_0} + (2 - \gamma)B_g \right] + M^2 - L_g \left( \frac{\|\nabla y\|}{\rho_0} + (2 - \gamma)B_g \right) - M^2 = 0.
\]
Furthermore, other conditions in (35) ensure that
\[
\rho_k > \eta_k, \quad \frac{1}{2\eta_k} = \frac{1 - \tau_k}{2\eta_k - 1}, \quad \rho_{k-1} - (1 - \tau_k)\rho_k = 0, \quad \text{and} \quad \frac{\tau_k^2}{2\beta_k} \leq \frac{(1 - \tau_k)\tau_k^2}{2\beta_{k-1}}.
\]
Utilizing these relations, we can simplify estimate (27) of Lemma 2 to get
\[
\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y) - \mathcal{L}(x, s, \hat{y}^{k+1}) + \frac{\tau_k^2}{2\beta_k} \|\hat{x}^{k+1} - x\|^2 + \frac{\tau_{k-1}^2}{2\eta_{k-1}} \|\hat{y}^{k+1} - y\|^2
\]
\[
\leq (1 - \tau_k) \left[ \mathcal{L}_{\rho_{k-1}}(x^k, s^k, y) - \mathcal{L}(x, s, \hat{y}^k) + \frac{\tau_{k-1}^2}{2\beta_{k-1}} \|\hat{x}^k - x\|^2 + \frac{1}{2\eta_{k-1}} \|\hat{y}^k - y\|^2 \right],
\]
for any \((x, s, y) \in \mathbb{R}^p \times \mathbb{R}^m \times \mathbb{R}^m\). Here, we have used the fact that \(\tilde{g}^k\) as defined in (61) is equal to \(\tilde{y}^k\) as defined in (20), since \(\tau_k = \frac{1}{k+1}\). By induction, this inequality implies that
\[
\mathcal{L}(x^k, s^k, y) - \mathcal{L}(x, s, \tilde{y}^k) \leq \mathcal{L}_{\rho_{k-1}}(x^k, s^k, y) - \mathcal{L}(x, s, \tilde{y}^k)
\]
\[
\leq \prod_{j=1}^{k-1} (1 - \tau_j) \left[ \mathcal{L}_{\rho_0}(x^1, s^1, y^*) - \mathcal{L}(x, s, \tilde{y}^1) + \frac{x^2}{2\rho_0} \|\tilde{x}^1 - x\|^2 + \frac{1}{2\rho_0} \|\tilde{y}^1 - y\|^2 \right]
\]
\[
\leq \frac{1}{k} \left[ (1 - \tau_0) \left( \mathcal{L}_{\rho_0}(x^0, s^0, y^*) - \mathcal{L}(x, s, y^0) \right) + \frac{x^2}{2\rho_0} \|\tilde{x}^0 - x\|^2 + \frac{1}{2\rho_0} \|y^0 - y\|^2 \right]
\]
\[
\tau_0 = 1 - \frac{1}{2k} \left( \frac{1}{\rho_0} \|x^0 - x\|^2 + \frac{1}{\rho_0} \|y^0 - y\|^2 \right) = \frac{\mathcal{R}_0^2(x,y)}{2k}.
\]
Take \(\tilde{x}^k \in -\partial H^*(\tilde{y}^k)\). Using the argument immediately following (75), we can show that
\[
\tilde{L}(x^k, y) - \tilde{L}(x, \tilde{y}^k) \leq \mathcal{L}(x^k, s^k, y) - \mathcal{L}(x, s^k, \tilde{y}^k) \leq \frac{\mathcal{R}_0^2(x,y)}{2k}.
\]
The rest of the proof of Theorem 2 is similar to the lines after (75) in the proof of Theorem 1 except that we replace \(\tilde{x}^k\) there by \(x^k\). Thus we omit the verbatim here. \(\square\)

**Proof (Proof of Lemma 2)** Since \(B_k \equiv \mathbb{R}^m\) by (55), i.e., there is no projection, the \(\tilde{y}^{k+1}\)-update in Algorithm 1 becomes \(\tilde{y}^{k+1} := \tilde{y}^k + \eta_k (\Theta_{k+1} - (1 - \tau_k) \Theta_k)\). Thus
\[
\tilde{y}^{k+1} - \eta_k \Theta_{k+1} = \tilde{y}^k - (1 - \tau_k) \eta_k \Theta_k \quad \text{by (35), i.e., there is no projection, the } \tilde{y}^{k+1}\text{-update in Algorithm 1 becomes } \tilde{y}^{k+1} := \tilde{y}^k + \eta_k (\Theta_{k+1} - (1 - \tau_k) \Theta_k)\text{.}
\]

By induction, for all \(k \in \mathbb{N}\), we obtain
\[
\tilde{y}^{k+1} - \eta_k \Theta_{k+1} = \tilde{y}^k - (1 - \tau_0) \eta_k \Theta_0 = \tilde{y}^0.
\]
Next, since \(0 \in \text{dom}(\partial H^*)\), we have \(\partial H(0) \neq 0\). Take an arbitrary \(s^* \in -\partial H^*(0)\), then \(-s^* = \text{prox}_{\rho_k H}(s^*)\) for any \(\rho_k > 0\). By the update of \(s^{k+1}\) in (20), the definition of \(B_g\), and the non-expansiveness of \(\text{prox}_{\rho_k H}\), we have
\[
\|s^{k+1}\| = \|\text{prox}_{\rho_k H} \left( \frac{\tilde{g}^k + g(\tilde{x}^k)}{\rho_k} \right) - \text{prox}_{\rho_k H} \left( -s^* \right) \| \\
\leq \left\| \frac{\tilde{g}^k + g(\tilde{x}^k)}{\rho_k} - s^* \right\| + \|s^*\| \leq \frac{\|\tilde{g}^k\|}{\rho_k} + B_g + 2\|s^*\|.
\]
Furthermore, by the \(\Theta_{k+1}\)-update in (25) and the connection between \(y^{k+1}\) and \(s^{k+1}\) described by (27), we have
\[
\|\Theta_{k+1}\| = \|g(x^{k+1}) + s^{k+1}\| \leq B_g + \|s^{k+1}\| \leq 2B_g + 2\|s^*\| + \frac{\|\tilde{g}^k\|}{\rho_k}.
\]
Now, we can prove (83) by induction. For \(k = 0\), it is true since \(\gamma \in (0,1)\) and \(\tilde{y}^0 = y^0\). Suppose that (83) holds for some \(K \geq 0\). We prove that it also holds for \(K + 1\). Indeed, using (85), (86), \(\rho_{K+1} \geq \rho_0\), and the induction hypothesis, we have
\[
\frac{\|\tilde{g}^{K+1}\|}{\rho_{K+1}} \leq \frac{1}{\rho_{K+1}} \|\tilde{g}^0\| + (1 - \gamma) \|\Theta_{K+1}\| \leq \frac{1}{\rho_{K+1}} \|\tilde{g}^0\| + (1 - \gamma) \left( 2B_g + 2\|s^*\| + \frac{\|\tilde{g}^K\|}{\rho_k} \right) \leq \frac{\|\tilde{g}^0\|}{\rho_{K+1}} + (1 - \gamma) \left( 2B_g + 2\|s^*\| + \frac{1}{\gamma} \left( \frac{\|\tilde{g}^0\|}{\rho_{K+1}} + 2(1 - \gamma) (B_g + \|s^*\|) \right) \right) \leq \frac{1}{\gamma} \left( \frac{\|\tilde{g}^0\|}{\rho_{K+1}} + 2(1 - \gamma) (B_g + \|s^*\|) \right) \cdot
\]
This shows that (83) also holds for \(K + 1\). Therefore, by induction, we conclude that (83) holds for all \(k \in \mathbb{N}\). \(\square\)
B.5 Proof of Theorem 3

\[ \min \{O(1/k) , \rho(1/(k\sqrt{\log k}))\} \] non-ergodic convergence rate of Algorithm 1

By the definition of \( B_k \) in (37) and the projection step of \( \tilde{y}^k \), we have \( \|\tilde{y}^k\| \leq \rho_k R_y \).

Similar to (84), by definition of \( y_s \), we can show that

\[
L_k \leq L_g (2\|y_s\| + \|\tilde{y}^k + \rho_k g(\hat{x}^k)\|) \leq L_g (2\|y_s\| + \rho_k R_y + \rho_k B_g)
\]

Thus, by the update of \( \beta_k \) and \( \eta_k \) in (37), one can show that

\[
\frac{1}{\beta_k} - L_k - L_f - \frac{\rho_k M_g^2}{\rho_k - \eta_k} \geq \left[ L_f + 2L_g \|y_s\| + \rho_k L_g (R_y + B_g) + \frac{\rho_k M_g^2}{\gamma} \right] - L_g (2\|y_s\| + \rho_k (R_y + B_g)) - \frac{L_f - \rho_k M_g^2}{\gamma} = 0.
\]

Using this inequality and the update rules from (37) onto (27) of Lemma 2, we can derive

\[
L(x^{k+1}, s^{k+1}, y) - L(x, s, \tilde{y}^{k+1}) + \frac{k+c}{e} \cdot \frac{\rho_0}{2} \|g(x^{k+1}) + s^{k+1}\|^2 
\]

\[
\leq \frac{k}{k+c} \left[ L(x, s, \tilde{y}^k) - L(x, s, \tilde{y}^{k+1}) + \frac{k+c-1}{c} \cdot \frac{\rho_0}{2} \|g(x^k) + s^k\|^2 \right]
\]

\[
+ \frac{c}{k+c} \cdot \frac{7}{2\lambda_k} \left[ \|\tilde{x}^k - x^*\|^2 - \|\tilde{x}^{k+1} - x^*\|^2 \right] - \left( \frac{(c-1)k}{c(k+c)} \cdot \frac{\rho_0}{2} \|g(x^k) + s^k\|^2 \right).
\]

Note that \( y^* \in B_0 \subseteq B_k \) by (59) and (37). Thus, we can substitute \((x, s, y) := (x^*, s^*, y^*)\) into (88) while introducing the following notations:

\[
\begin{align*}
\tilde{G}_k := L(x^k, s^k, y^*) - L(x^*, s^*, y^*) = L(x^k, s^k, y^*) - \mathcal{P}^* \geq 0.
\end{align*}
\]

Then, we can simplify (88) as:

\[
\tilde{G}_{k+1} + \frac{k+c}{e} \cdot a_{k+1}^2 \leq \frac{k}{k+c} \left( \tilde{G}_k + \frac{k+c-1}{c} a_k^2 \right) - \frac{(c-1)k}{c(k+c)} a_k^2 
\]

\[
+ \frac{c}{k+c} \left[ b_k^2 - b_{k+1}^2 + \frac{c(L_f + 2L_g \|y_s\|)}{2} \left( \frac{1}{k+c} - \frac{1}{k+c+1} \right) \|\tilde{x}^{k+1} - x^*\|^2 \right] 
\]

\[
\leq \frac{k}{k+c} \left( \tilde{G}_k + \frac{k+c-1}{c} a_k^2 \right) - \frac{(c-1)k}{c(k+c)} a_k^2 + \frac{c}{k+c} \left( b_k^2 - b_{k+1}^2 \right).
\]

Multiplying both sides of the last inequality by \( k + c \) and rearranging the result, we get

\[
(c - 1) \left( \tilde{G}_k + \frac{k+c-1}{c} a_k^2 \right) \leq (c - 1) \left( \tilde{G}_k + \frac{2k+c-1}{c} a_k^2 \right) 
\]

\[
\leq \left[ (k+c-1) \tilde{G}_k + \frac{(k+c-1)^2}{c} a_k^2 + c b_k^2 \right] - \left[ (k+c) \tilde{G}_{k+1} + \frac{(k+c)^2}{c} a_{k+1}^2 + c b_{k+1}^2 \right].
\]

Since \( c > 1, \tilde{G}_k \geq 0, \) and \( a_k^2 \geq 0, \) the inequality (89) implies that

\[
(k + c) \tilde{G}_{k+1} + \frac{(k+c)^2}{c} a_{k+1}^2 + c b_{k+1}^2 \leq (k + c - 1) \tilde{G}_k + \frac{(k+c-1)^2}{c} a_k^2 + c b_k^2.
\]

By induction, we can show that

\[
(k + c - 1) \tilde{G}_k + \frac{(k+c-1)^2}{c} a_k^2 + c b_k^2 \leq (c - 1) \tilde{G}_0 + \frac{(c-1)^2}{c} a_0^2 + c b_0^2 = (c - 1)\mathcal{P}(x^0) - \mathcal{P}^* + \frac{c\mathcal{R}^2(x^0, y^*)}{2} = \Delta_0^2,
\]
where in the second line we have used \( \|g(x^0) + s^0\| = \|\Theta_0\| = 0 \) as initialized in Step 2 of Algorithm 1. As a result,
\[
\mathcal{L}(x^k, s^k, y^*) - \mathcal{P}^* = \tilde{G}_k \leq \frac{\Delta_0^2}{k + c - 1} \quad \text{and} \quad \|g(x^k) + s^k\| = a_k \sqrt{\frac{2}{\rho_0}} \leq \sqrt{\frac{2c}{k + c - 1}}. \tag{90}
\]
Consequently, if \( H \) is \( M_H \)-Lipschitz continuous, then we can show that
\[
\mathcal{P}(x^k) - \mathcal{P}^* = F(x^k) + H(g(x^k)) - \mathcal{P}^* \leq F(x^k) + H(-s^k) + M_H \|g(x^k) + s^k\| - \mathcal{P}^*
\leq \mathcal{L}(x^k, s^k, y^*) - \mathcal{P}^* + (\|y^\ast\| + M_H) \|g(x^k) + s^k\|
\leq \frac{\Delta_0^2}{k + c - 1} + \sqrt{\frac{2c}{\rho_0 \Delta_0} (\|y^\ast\| + M_H)},
\tag{91}
\]
which is the first assertion of (38).
Next, summing up (89) from \( j := 0 \) to \( j := k \), we get
\[
(c - 1) \sum_{j=0}^{k} \left[ G_j + \frac{j+c-1}{c}a_j^2 \right] \leq \left[ (c - 1)\tilde{G}_0 + \frac{(c-1)^2}{c}a_0^2 + cb_0^2 \right] - \left[ (k + c)\tilde{G}_{k+1} + \frac{(k+c)^2}{c}a_{k+1}^2 + cb_{k+1}^2 \right] \leq \Delta_0^2.
\]
Since \( c > 1 \) and \( \tilde{G}_j \geq 0 \), we can apply Lemma 4(b) to show that
\[
\liminf_{k \to \infty} (k \log k) \left( \tilde{G}_k + \frac{ka_k^2}{c} \right) = 0, \tag{92}
\]
Combining this limit with (90) and (91), and applying Lemma 4(c(i)), we can easily prove the second assertion of (38). \( \square \)

C Technical Proofs in Section 4: Strongly Convex-Concave Case
This section provides the full proofs of technical results in Section 4.

C.1 Proof of Lemma 3: One-iteration analysis of Algorithm 2
To prove Lemma 3, we need the following two lemmas, whose proofs are given right after the proof of Lemma 3.

Lemma 8 Let \( \{(x^k, \tilde{y}^k)\} \) be generated by (39) with \( \tau_k \in [0, 1] \), and \( \{s^k\} \) be defined in (20). Let us define
\[
\hat{x}^{k+1} := (1 - \tau_k)x^k + \tau_k \tilde{x}^{k+1}. \tag{93}
\]
Then, for any \((x, s) \in \mathbb{R}^p \times \mathbb{R}^m\),
\[
F(x^{k+1}) + H(-s^{k+1}) \leq (1 - \tau_k)[F(x^k) + H(-s^k)] + \tau_k[F(x) + H(-s)] + \langle \nabla_x \phi_{\tau_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), (1 - \tau_k)x^k + \tau_k x - x^{k+1} \rangle \\
+ \langle \nabla_s \phi_{\tau_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), (1 - \tau_k)s^k + \tau_k s - s^{k+1} \rangle \\
+ \frac{\tau_k^2}{2\beta_x} \|\hat{x}^k - x\|^2 + \frac{\tau_k^2}{2\beta_{\hat{y}}} \|\hat{x}^{k+1} - x\|^2 - \frac{\tau_k}{\alpha_k - L_f} \|\hat{x}^{k+1} - \hat{x}^k\|^2 \\
- \frac{1}{2} \left( \frac{1}{\alpha_k} + \mu_h \right) \|\hat{x}^{k+1} - x^{k+1}\|^2 - \frac{1}{2} \left( \frac{1}{\beta_k} - \frac{1}{\alpha_k} \right) \|\hat{x}^{k+1} - \hat{x}^k\|^2. \tag{94}
\]
Lemma 9 Let \( \{(x^k, y^k)\} \) be generated by (39) with \( \tau_k \in [0, 1] \), and \( \{s^k\} \) be defined in (20), then, for any \((x, s) \in \mathbb{R}^p \times \mathbb{R}^m\), we have

\[
\phi_{p_k}(x^{k+1}, s^{k+1}, y^k) \leq (1 - \tau_k)\phi_{p_k}(x^k, s^k, y^k) + \tau_k \phi_{p_k}(x, s, y^k) + \frac{L_k + \rho_k M_y^2}{2} \|x^{k+1} - \hat{x}^k\|^2
\]

\[
+ \langle \nabla_x \phi_{p_k}(x^k, s^{k+1}, y^k), x^{k+1} - (1 - \tau_k)x^k - \tau_k x \rangle \\
+ \langle \nabla_s \phi_{p_k}(x^k, s^{k+1}, y^k), s^{k+1} - (1 - \tau_k)s^k - \tau_k s \rangle \\
- \frac{1 - \tau_k}{2} \|g(x^k) + s^k\|^2 - \|g(\hat{x}^k) + s^{k+1}\|^2 \\
- \frac{\tau_k \rho_k}{2} \|g(x) + s - [g(\hat{x}^k) + s^{k+1}]\|^2.
\]

Proof (Proof of Lemma 9) Summing up the estimates (94) from Lemma 8 and (95) from Lemma 9 we get

\[
\mathcal{L}_{p_k}(x^{k+1}, s^{k+1}, y^k) \leq (1 - \tau_k)\mathcal{L}_{p_k}(x^k, s^k, y^k) + \tau_k \mathcal{L}_{p_k}(x, s, y^k) \\
+ \frac{\tau_k^2}{2\alpha_k} \|\hat{x}^k - x\|^2 - \frac{\tau_k (\tau_k + \beta_k \mu_k)}{2\alpha_k} \|\hat{x}^{k+1} - x\|^2 \\
- \frac{1}{2} \left( \frac{1}{\alpha_k} - L_k - L_f - \rho_k M_y^2 \right) \|x^{k+1} - \hat{x}^k\|^2 \\
- \frac{1}{2} \left( \frac{1}{\alpha_k} - \frac{1}{\tau_k} \right) \|\hat{x}^{k+1} - \hat{x}^k\|^2 - \frac{1}{2} \left( 1 + \frac{\mu_k}{\alpha_k} \right) \|\hat{x}^{k+1} - x^{k+1}\|^2 \\
- \frac{1}{2} \left( 1 + \frac{\mu_k}{\alpha_k} \right) \|\hat{x}^{k+1} - x^{k+1}\|^2 \\
- \frac{1}{2} \left( \frac{1}{\alpha_k} - L_k - L_f - \rho_k M_y^2 \right) \|x^{k+1} - \hat{x}^k\|^2 + \mathcal{T}_1,
\]

where

\[
\mathcal{T}_1 := \frac{\eta_k}{2} \|g(x^{k+1}) + s^{k+1}\|^2 - (1 - \tau_k) \|g(x^k) + s^k\|^2 - \frac{1}{2} \left( \frac{1}{\tau_k} + \frac{1}{\alpha_k} \right) \|\hat{x}^{k+1} - \hat{x}^k\|^2 \\
- \frac{1}{2} \left( \frac{1}{\alpha_k} - L_k - L_f - \rho_k M_y^2 \right) \|x^{k+1} - \hat{x}^k\|^2 \\
- \frac{1}{2} \left( \frac{1}{\alpha_k} - L_k - L_f - \rho_k M_y^2 \right) \|x^{k+1} - \hat{x}^k\|^2 + \mathcal{T}_1,
\]

To upper bound \( \mathcal{T}_1 \), we can use the same line as (69) in proof of Lemma 2 to derive

\[
\mathcal{T}_1 \leq \frac{\rho_k \eta_k M_y^2}{2(\rho_k - \eta_k)} \|x^{k+1} - \hat{x}^k\|^2 - \frac{1}{2} (1 - \tau_k) \|\rho_{k-1} - (1 - \tau_k) \rho_k\| \|g(x^k) + s^k\|^2.
\]
Moreover, applying Lemma 4(a) with $t_1 := \frac{1}{2} \left( \frac{1}{\beta_k} - \frac{1}{\alpha_k} \right) > 0$ and $t_2 := \frac{1}{2} \left( \frac{1}{\alpha_k} + \mu_h \right)$, we can show that

$$- \frac{1}{2} \left( \frac{1}{\beta_k} - \frac{1}{\alpha_k} \right) \| \hat{x}^{k+1} - \hat{x}^k \|^2 - \frac{1}{2} \left( \frac{1}{\alpha_k} + \mu_h \right) \| \hat{x}^{k+1} - x^{k+1} \|^2 \leq - \frac{1}{2} \frac{t_1 t_2}{\alpha_k} \| x^{k+1} - \hat{x}^k \|^2 \leq - \frac{1}{2} \left( \frac{1}{\beta_k} - \frac{1}{\alpha_k} \right) \left( \frac{1}{\alpha_k} + \mu_h \right) \| x^{k+1} - \hat{x}^k \|^2$$

$$\leq - \frac{1}{2 \alpha_k} \left( 1 - \frac{\beta_k}{\alpha_k} \right) \| x^{k+1} - \hat{x}^k \|^2,$$

where in the last inequality we used $\alpha_k > \beta_k$. Substituting (98) and (99) into (97), we eventually obtain (100).

Proof (Proof of Lemma 8) Firstly, from the optimality condition of the $\hat{x}^{k+1}$-subproblem in the second line of (39), there exists $\nabla h(\hat{x}^{k+1}) \in \partial h(\hat{x}^{k+1})$ such that

$$\nabla h(\hat{x}^{k+1}) = - \frac{\tau_k}{\beta_k} (\hat{x}^{k+1} - \hat{x}^k) - \left[ \nabla f(\hat{x}^k) + \nabla_x \phi_{p_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k) \right],$$

where we have used the expression of $\nabla x \phi$ in (17). Combining this expression and (93), and using the $\mu_h$-strong convexity of $h$, we can derive

$$h(x^{k+1}) \leq (1 - \tau_k) h(x^k) + \tau_k h(x) + \tau_k \left( \nabla h(\hat{x}^{k+1}), \hat{x}^{k+1} - x \right)$$

$$= - \frac{\tau_k \mu_h}{2} \| \hat{x}^{k+1} - x \|^2 - \frac{\tau_k (1 - \tau_k) \mu_h}{2} \| \hat{x}^{k+1} - x \|^2$$

$$= (1 - \tau_k) h(x^k) + \tau_k h(x) - \tau_k \left( \nabla f(x^k) + \nabla_x \phi_{p_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), \hat{x}^{k+1} - x \right)$$

$$- \frac{\tau_k \mu_h}{2} \| \hat{x}^{k+1} - x \|^2 - \frac{\tau_k (1 - \tau_k) \mu_h}{2} \| \hat{x}^{k+1} - x \|^2$$

(100)

Next, by the $x^{k+1}$-subproblem in the third line of (39) and the $\mu_h$-strong convexity of $h$, we can show that

$$h(x^{k+1}) + \frac{1}{2 \alpha_k} \| x^{k+1} - \hat{x}^k \|^2 + \langle \nabla f(\hat{x}^k) + \nabla_x \phi_{p_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), x^{k+1} - \hat{x}^k \rangle$$

$$\leq h(\hat{x}^{k+1}) + \frac{1}{2 \alpha_k} \| \hat{x}^{k+1} - \hat{x}^k \|^2$$

$$+ \frac{1}{2 \alpha_k} \| \hat{x}^{k+1} - \hat{x}^k \|^2 - \| x^{k+1} - \hat{x}^k \|^2 - \| x^{k+1} - x^{k+1} \|^2$$

$$- \frac{\tau_k \mu_h}{2} \| \hat{x}^{k+1} - x^{k+1} \|^2$$

$$\leq (1 - \tau_k) h(x^k) + \tau_k h(x) - \frac{\tau_k^2}{\alpha_k} (\hat{x}^{k+1} - \hat{x}^k, \hat{x}^{k+1} - x) - \frac{\tau_k \mu_h}{2} \| \hat{x}^{k+1} - x \|^2$$

$$+ \langle \nabla f(\hat{x}^k) + \nabla_x \phi_{p_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), \hat{x}^{k+1} - \hat{x}^k \rangle$$

$$+ \frac{1}{2 \alpha_k} \| \hat{x}^{k+1} - \hat{x}^k \|^2 - \| x^{k+1} - \hat{x}^k \|^2 - \| \hat{x}^{k+1} - x^{k+1} \|^2$$

$$+ \tau_k (1 - \tau_k) \mu_h \| \hat{x}^{k+1} - x \|^2 - \frac{\mu_h}{2} \| \hat{x}^{k+1} - x \|^2$$

(101)

Combining (93), (100), and (101), and using $\hat{x}^{k+1} - \hat{x}^k = \tau_k (\hat{x}^{k+1} - \hat{x}^k)$, we further derive

$$h(x^{k+1}) \leq h(\hat{x}^{k+1}) + \langle \nabla f(\hat{x}^k) + \nabla_x \phi_{p_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), \hat{x}^{k+1} - x^{k+1} \rangle$$

$$+ \frac{1}{2 \alpha_k} \| \hat{x}^{k+1} - \hat{x}^k \|^2 - \| x^{k+1} - \hat{x}^k \|^2 - \| \hat{x}^{k+1} - x^{k+1} \|^2$$

$$+ \frac{\tau_k \mu_h}{2} \| \hat{x}^{k+1} - x \|^2 - \frac{\mu_h}{2} \| \hat{x}^{k+1} - x \|^2$$

(102)
On the other hand, by the $L_f$-smoothness and the convexity of $f$, one can show that

\[
 f(x^{k+1}) \leq f(\hat{x}^k) + \langle \nabla f(\hat{x}^k), x^{k+1} - \hat{x}^k \rangle + \frac{L_f}{2} \| x^{k+1} - \hat{x}^k \|^2 \\
 \leq (1 - \tau_k) f(x^k) + \tau_k f(x) + \langle \nabla f(\hat{x}^k), x^{k+1} - (1 - \tau_k) x^k - \tau_k x \rangle \\
 - \frac{(1 - \tau_k) \tau_k \mu \| x^k - x \|^2 + L_f / 2 \| x^{k+1} - \hat{x}^k \|^2.}
\]

Moreover, by the $s^{k+1}$-subproblem in (20), we get exactly (72) again, which implies

\[
 H(-s^{k+1}) \leq (1 - \tau_k) H(-s^k) + \tau_k H(-s) \\
 + \langle \nabla \phi_\rho_k(\hat{x}^k, s^{k+1}, \tilde{y}^k), (1 - \tau_k) s^k + \tau_k s - s^{k+1} \rangle. \tag{103}
\]

Finally, combining (102)–(103), we obtain (94).

**Proof (Proof of Lemma 9)** Combining (24) and the definition of $L_k$ in (26), we have $L_k = L_g \left( \tilde{y}^k + \rho_k [g(\hat{x}^k) + s^k] \right)$. Thus we can use (19) in Lemma 1 and the $\mathcal{M}_g$-Lipschitz continuity of $g$ to get

\[
 \begin{cases}
 \phi_{\rho_k}(x^{k+1}, s^{k+1}, \tilde{y}^k) \leq \phi_{\rho_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k) + \langle \nabla_x \phi_{\rho_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), x^{k+1} - \hat{x}^k \rangle \\
 + \frac{L_g + \rho_k \mathcal{M}_g^2}{2} \| x^{k+1} - \hat{x}^k \|^2, \\
 \phi_{\rho_k}(x, s, \tilde{y}^k) \\
 \geq \phi_{\rho_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k) + \langle \nabla_x \phi_{\rho_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), x - \hat{x}^k \rangle \\
 + \langle \nabla_s \phi_{\rho_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), s - s^{k+1} \rangle + \frac{l_k}{2} \| g(x) + s - [g(\hat{x}^k) + s^{k+1}] \|^2. \tag{104}
\end{cases}
\]

Letting $(x, s) := (x^k, s^k)$ in the second inequality of (104), we get

\[
 \phi_{\rho_k}(x^k, s^k, \tilde{y}^k) \geq \phi_{\rho_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k) + \langle \nabla_x \phi_{\rho_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), x^k - \hat{x}^k \rangle \\
 + \langle \nabla_s \phi_{\rho_k}(\hat{x}^k, s^{k+1}, \tilde{y}^k), s^k - s^{k+1} \rangle \tag{105}
\]

Multiplying the second inequality of (104) by $\tau_k$, multiplying (105) by $1 - \tau_k$, and then adding them to the first inequality of (104), we arrive at (95). \qed}

### C.2 Proof of Theorem 4: $O(1/k^2)$ ergodic convergence rate of Algorithm 2

To prove Theorem 4, we need the following two technical lemmas.

**Lemma 10** Let $\rho_k$, $\beta_k$, and $\eta_k$ be defined by (41) and (42) of Theorem 4. Then, for all $k \in \mathbb{N}$, we have

\[
 \begin{cases}
 \beta_k \leq \frac{\Gamma}{L_f + \rho_k \mathcal{M}_f^2}, & \rho_k \geq \rho_0 + (\sqrt{1 + \mu_k \beta_0} - 1) \rho_0 k, \\
 1 + \rho_k \beta_k / \rho_k = \frac{1}{\gamma_{k+1}}, & \frac{1}{\rho_k} = \frac{1}{\gamma_{k+1}} \frac{1}{\rho_{k+1}}, \quad \text{and} \quad \frac{1}{\eta_k} = \frac{1}{\gamma_{k+1}} \frac{1}{\rho_{k+1}}. \tag{106}
\end{cases}
\]

**Lemma 11** Let $\rho_k$, $\beta_k$, and $\eta_k$ be defined by (41) and (42) of Theorem 4. Then for $k \in \mathbb{N},$

\[
 L_g \left[ \| y^* \| + \| \tilde{y}^k - y^* \| + \rho_k \mathcal{M}_g \| x^k - x^* \| \right] \leq \rho_k \left[ (2 - \Gamma) \mathcal{M}_f^2 - \frac{M_f^2}{\gamma} \right]. \tag{107}
\]
Proof (Proof of Theorem 14) By Lemma 11, holds for all \(k \in \mathbb{N}\). Thus, using the same lines from (108) to (110), we have for all \(j \in \mathbb{N}\) any \((x, s, y) \in \mathbb{R}^p \times \mathbb{R}^m \times \mathbb{R}^m\) that

\[
\mathcal{L}(x^{j+1}, s^{j+1}, y) - \mathcal{L}(x, s, y^{j+1}) \geq \left[ \frac{1}{2\theta_j} \|\tilde{x}^j - x\|^2 + \frac{1}{\eta_j} \|\tilde{y}^j - y\|^2 \right] - \frac{1}{2\theta_{j+1}} \|\tilde{x}^j - x\|^2 - \frac{1}{2\eta_{j+1}} \|\tilde{y}^j - y\|^2 \right].
\]

Multiplying the last inequality by \(2\rho_j\) and noticing that \(\frac{\rho_j}{\theta_{j+1}} = \rho_{j+1}\), it leads to

\[
2\rho_j \mathcal{L}(x^{j+1}, s^{j+1}, y) - \mathcal{L}(x, s, y^{j+1}) \leq \rho_j \left[ \frac{1}{\eta_j} \|\tilde{x}^j - x\|^2 + \frac{1}{\eta_j} \|\tilde{y}^j - y\|^2 \right] - \rho_{j+1} \left[ \frac{\rho_j}{\theta_j} \|\tilde{x}^j - x\|^2 + \frac{1}{\eta_{j+1}} \|\tilde{y}^j - y\|^2 \right].
\]

Summing up this inequality from \(j := 0\) to \(j := k - 1\), we obtain

\[
\sum_{j=0}^{k-1} \rho_j [\mathcal{L}(x^{j+1}, s^{j+1}, y) - \mathcal{L}(x, s, y^{j+1})] \leq \frac{\rho_0}{2} \left[ \frac{1}{\beta_0} \|x^0 - x\|^2 + \frac{1}{\eta_0} \|y^0 - y\|^2 \right] = \rho_0 \mathcal{R}_0^2(x, y) / 2.
\]

Dividing it by \(\sum_{j=0}^{k-1} \rho_j\), and using the convexity of \(\mathcal{L}\) in \(x\) and \(s\), and the concavity in \(y\), and \((\bar{x}^k, \bar{y}^k)\) defined by (43) and \(\bar{s}^k := \left( \sum_{j=0}^{k-1} \rho_j \right)^{-1} \sum_{j=0}^{k-1} \rho_j s^{j+1}\), we get

\[
\mathcal{L}(\bar{x}^k, \bar{s}^k, y) - \mathcal{L}(x, s, \bar{y}^k) \leq \frac{1}{\sum_{j=0}^{k-1} \rho_j} \sum_{j=0}^{k-1} \rho_j \left[ \mathcal{L}(x^{j+1}, s^{j+1}, y) - \mathcal{L}(x, s, y^{j+1}) \right] \leq \frac{\rho_0 \mathcal{R}_0^2(x, y)}{2 \sum_{j=0}^{k-1} \rho_j}.
\]

By the second inequality in (106) of Lemma 10 we have

\[
\sum_{j=0}^{k-1} \rho_j \geq k \rho_0 + \frac{1}{2} (\sqrt{1 + \mu_h \beta_0} - 1) \rho_0 k (k - 1) \geq \frac{1}{2} (\sqrt{1 + \mu_h \beta_0} - 1) \rho_0 k (k - 1).
\]

Combining the above two inequalities, we eventually get

\[
\mathcal{L}(\bar{x}^k, \bar{s}^k, y) - \mathcal{L}(x, s, \bar{y}^k) \leq \frac{\mathcal{R}_0^2(x, y)}{(\sqrt{1 + \mu_h \beta_0} - 1) k (k - 1)}.
\]

Therefore, we can use the same lines as in the proof of Theorem 1 to prove (44).

Proof (Proof of Lemma 10) We prove the first inequality in (106) by induction. First, it holds with equality for \(k = 0\) due to the definition of \(\beta_0\). Furthermore, if it holds for \(k \geq 0\), then we can show that it holds for \(k + 1\). Indeed, we have

\[
\beta_{k+1} \frac{\theta_{k+1} \Gamma}{L_f + \rho_k M^2} \leq \frac{\theta_{k+1} \Gamma}{L_f + \rho_k M^2} \leq \frac{\Gamma}{L_f + \rho_{k+1} M^2} \leq \frac{\Gamma}{L_f + \rho_{k+1} M^2}.
\]

Thus the first inequality of (106) is also true for \(k + 1\). By induction, the first inequality in (106) holds for all \(k \in \mathbb{N}\).

To prove the second inequality of the first line in (106), we notice that (42) implies

\[
\rho_{k+1} \rho_k \sqrt{1 + \mu_h \beta_k} = \rho_k \left( 1 + \frac{\mu_h \beta_k}{1 + \sqrt{1 + \mu_h \beta_k}} \right) \rho_k + \frac{\mu_h \beta_k \rho_0}{1 + \sqrt{1 + \mu_h \beta_k}} \rho_k + \left( \sqrt{1 + \mu_h \beta_k} - 1 \right) \rho_0.
\]
The desired inequality is then achieved via induction.

The first statement of the second line in (106) holds since

\[
1 + \mu_k \beta_k \frac{1}{\beta_k} \frac{1}{\theta_{k+1}^2 \beta_k} \frac{1}{\theta_{k+1} \beta_{k+1}}.
\]

The last two equations of (106) directly follow from the update of \(\eta_k\) and \(\rho_k\) in (42).

**Proof (Proof of Lemma 11)** We prove (107) by induction. For \(k = 0\), the inequality (107) holds due to the second line in (41). Suppose that (107) holds for all \(k \in \{0, 1, \cdots, K\}\) for some \(K \in \mathbb{N}\). Then by the definition of \(L_k\) in Lemma 3 and the same lines as (76), for all \(k \in \{0, 1, \cdots, K\}\), we can show that

\[
L_k \leq L_g [\|y^*\| + \|\hat{y}^k - y^*\| + \rho_k M_g \|\hat{x}^k - x^*\|] \leq \rho_k \left(2 - \Gamma \hat{M}^2 - \frac{M_g^2}{\gamma}\right). \tag{108}
\]

By the update of \(\alpha_k\), \(\eta_k\) and \(\beta_k\) in (42), the first inequality of (106), and (108), we have

\[
\frac{1}{\alpha_k} \left(1 - \frac{\beta_k}{\alpha_k}\right) - L_k - L_f - \frac{\rho_k^2 M_g^2}{\rho_k \gamma} \leq \frac{1}{\alpha_k} \left(1 - \Gamma \right)(L_f + \rho_k \hat{M}^2) + \rho_k \hat{M}^2 - L_k - \frac{\rho_k M_g^2}{\gamma} \leq 0.
\]

Using this inequality, \(\tau_k := 1\), and \((x, s, y) := (x^*, s^*, y^*)\) into (40) of Lemma 3 yields

\[
0 \leq \mathcal{E}_{\rho_k}(x^{k+1}, x^{k+1}, y^*) - p^*
\]

\[
\leq \frac{1}{2\tau_{k+1}} \|\hat{x}^{k+1} - x^*\|^2 + \frac{1}{2\tau_{k+1}} \|\hat{y}^{k+1} - y^*\|^2 + \frac{1}{2\tau_{k+1}} \|\hat{y}^{k+1} - y^{k+1} - y^*\|^2
\]

\[
- \frac{1}{\theta_{k+1}} \left[\frac{1}{2\tau_{k+1}} \|\hat{x}^{k+1} - x^*\|^2 + \frac{1}{2\tau_{k+1}} \|\hat{y}^k - y^*\|^2\right]. \tag{110}
\]

Multiplying (110) by \(2\rho_k\), and noticing that \(\rho_k \leq \frac{\rho_k}{\rho_{k+1}} = \rho_{k+1}\), we get

\[
\rho_k \left[\frac{1}{\tau_{k+1}} \|\hat{x}^{k+1} - x^*\|^2 + \frac{1}{\tau_{k+1}} \|\hat{y}^{k+1} - y^*\|^2\right] \leq \rho_{k+1} \left[\frac{1}{\tau_{k+1}} \|\hat{x}^{k+1} - x^*\|^2 + \frac{1}{\tau_{k+1}} \|\hat{y}^{k+1} - y^*\|^2\right]
\]

By induction, the above holds for \(k \in \{0, 1, \cdots, K\}\). Consequently, one has

\[
\frac{1}{\beta_{k+1}} \|\hat{x}^{k+1} - x^*\|^2 + \frac{1}{\beta_{k+1}} \|\hat{y}^{k+1} - y^*\|^2 \leq \frac{1}{\rho_{k+1}} \|x^0 - x^*\|^2 + \frac{1}{\rho_{k+1}} \|y^0 - y^*\|^2 = \mathcal{R}_0(x^*, y^*),
\]

which implies that

\[
\|\hat{x}^{k+1} - x^*\| \leq \sqrt{\beta_{k+1}} \mathcal{R}_0(x^*, y^*)\quad \text{and} \quad \|\hat{y}^{k+1} - y^*\| \leq \sqrt{\beta_{k+1}} \mathcal{R}_0(x^*, y^*).
\]

Finally, using the above estimates, we can easily deduce

\[
\frac{1}{\rho_{k+1}} \left[\|x^0 - x^*\|^2 + \|y^{k+1} - y^*\|^2\right] \leq \frac{1}{\rho_{k+1}} \mathcal{R}_0(x^*, y^*) + \frac{1}{\rho_{k+1}} M_g \mathcal{R}_0(x^*, y^*)
\]

\[
\leq \frac{1}{\rho_{k+1}} \mathcal{R}_0(x^*, y^*) + \frac{1}{\rho_{k+1}} \mathcal{R}_0(x^*, y^*) \leq \frac{1}{\rho_{k+1}} \mathcal{R}_0(x^*, y^*) \leq \frac{1}{\tau_{k+1}} \left(2 - \Gamma \right) \hat{M}^2 - \frac{M_g^2}{\gamma}.
\]

This inequality shows that (107) also holds for \(K + 1\). By induction, we have thus proved that (107) holds for all \(k \in \mathbb{N}\).
C.3 Proof of Remark 3: Initializing parameters in Theorem 4

For simplicity, we set \( \rho_0 := 1, \gamma := \Gamma := \frac{1}{2} \). Using the same lines as (79) and (80) in the proof of Remark 1, it is clear that we only need the following inequality for (41) to hold:

\[
L_g D \left( 1 + \sqrt{L_f + \hat{M}^2 + 1} + M_g \sqrt{1 + \frac{1}{L_f + \hat{M}^2}} \right) \leq \frac{3\hat{M}^2}{2} - 2M_g^2. \tag{111}
\]

Let

\[
\hat{M}^2 \geq L_f + 1. \tag{112}
\]

Then, \( L_f + \hat{M}^2 + 1 \leq \sqrt{2\hat{M}} \) and \( \frac{1}{L_f + \hat{M}^2} \leq 1 \). Substituting these into (111), we can see that (111) holds if

\[
L_g D \left( 1 + \sqrt{\hat{M}} + \sqrt{2M_g} \right) \leq \frac{3\hat{M}^2}{2} - 2M_g^2,
\]

which can be solved as

\[
\hat{M} \geq \frac{2}{3} \sqrt{2L_g^2D^2 + 3(2M_g^2 + L_g D + \sqrt{2L_g D M_g})} \tag{113}
\]

Combining (112) and (113), we finally get (41).

C.4 Proof of Theorem 5: \( \mathcal{O}(1/k^2) \) semi-ergodic rate of Algorithm 2

Firstly, the update of \( \{\tau_k\} \) in (47) leads to

\[
\frac{\tau_k^2}{2\beta_k} = (1 - \tau_k)\tau_{k-1}^2 \quad \text{and} \quad \frac{1}{k+1} \leq \tau_k \leq \frac{2}{k+2}. \tag{114}
\]

Thus \( \rho_{k-1} = \frac{\rho_0}{\tau_{k-1}^2} = \frac{(1 - \tau_k)\rho_0}{\tau_k^2} = (1 - \tau_k)\rho_k \), which implies

\[
\rho_k \geq \eta_k, \quad \rho_k \geq \rho_0, \quad \frac{1}{2\eta_k} = \frac{1 - \tau_k}{2\tau_{k-1}}, \quad \text{and} \quad \rho_{k-1} - (1 - \tau_k)\rho_k = 0. \tag{115}
\]

By the equality in (46), the updates in (47), and (114), for \( k \geq 1 \), we have

\[
\begin{align*}
\tau_k^2 \left( \frac{L_f + \rho_k \hat{M}^2}{2f} \right) & = \tau_k^2 \left( L_f + \rho_k \hat{M}^2 \right) + \tau_k \rho_k \hat{M}^2 \tag{116} \\
& = \frac{\tau_k^2}{2\beta_{k-1}} + \frac{\tau_k \rho_k \hat{M}^2}{2f} \tag{47} \\
& = \frac{(1 - \tau_k)\tau_{k-1}^2}{2\beta_{k-1}},
\end{align*}
\]

Moreover, by definition of \( s_* \) and \( y_* \), it is easily shown that (84) of Lemma 7 still holds. Therefore, by the first inequality of (46) and \( y^0 := 0 \), we can derive that

\[
\begin{align*}
\frac{1}{\alpha_k} \left( 1 - \frac{\beta_k}{\alpha_k} \right) & + \frac{1}{\alpha_k} - L_k - L_f - \frac{\rho_k^2 M_g^2}{\rho_k - \eta_k} \\
& \geq (1 - \Gamma)(L_f + \rho_k \hat{M}^2) + \rho_k \hat{M}^2 - \frac{\rho_k M_g^2}{\gamma} \tag{117} \\
& - L_g \left( 2\|y_*\| + \frac{\rho_k}{\gamma} \left( (2 - \gamma) B_g + 2(1 - \gamma)\|s_*\| \right) \right) \tag{118} \\
& = \rho_k \left( (2 - \Gamma)\hat{M}^2 - \frac{1}{\gamma} M_g^2 + (2 - \gamma) B_g L_g + 2(1 - \gamma)\|s_*\|\|L_g\| \right) \tag{119} \\
& + (1 - \Gamma)L_f - 2L_g \|y_*\| \geq 0,
\end{align*}
\]
Combining (115)–(116), we can simplify the relation (40) in Lemma 3 as
\[
\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y) - \mathcal{L}(x, s, y^{k+1}) + \frac{\tau_k(\tau_k+\beta_k\mu_k)}{2\beta_k} \|\tilde{x}^{k+1} - x\|^2 + \frac{1}{2\eta_k} \|y^{k+1} - y\|^2
\leq (1 - \tau_k) \left[ \mathcal{L}_{\rho_{k-1}}(x^k, s^k, y) - \mathcal{L}(x, s, y^k) + \frac{\tau_{k-1}(\tau_{k-1}+\beta_{k-1}\mu_{k-1})}{2\beta_{k-1}} \|\tilde{x}^k - x\|^2 + \frac{1}{2\eta_{k-1}} \|y^k - y\|^2 \right].
\]

Here, we have used the fact that \(\tilde{y}^k\) defined in (20) is equal to the ergodic iterate \(\bar{y}^k\) defined in the statement of Theorem 5, thus we replace \(\tilde{y}^k\) by \(\bar{y}^k\). By induction, this inequality implies that
\[
\mathcal{L}(x^k, s^k, y) - \mathcal{L}(x, s, \bar{y}^k) \leq \mathcal{L}_{\rho_{k-1}}(x^k, s^k, y) - \mathcal{L}(x, s, y^k) \leq \left[ \prod_{j=1}^{k-1} (1 - \tau_j) \right]
\times \left[ \mathcal{L}_{\rho_0}(x^1, s^1, y) - \mathcal{L}(x, s, y^1) + \frac{\tau_0(\tau_0+\beta_0\mu_0)}{2\beta_0} \|\tilde{x}^0 - x\|^2 + \frac{1}{2\eta_0} \|\tilde{y}^0 - y\|^2 \right]
\leq \frac{2}{(k+1)^2} \left[ \frac{\beta_0}{\eta_0} \|x^0 - x\|^2 + \frac{1}{\eta_0} \|y^0 - y\|^2 \right] = \frac{2}{(k+1)^2} \mathcal{R}_0^2(x, y).
\]

Using the last estimate we can prove (48) in a similar manner as in the proof of Theorem 2. We therefore omit the details here.

C.5 Proof of Theorem 6: \min(\mathcal{O}(1/k^2), \mathcal{O}(1/(k^2\sqrt{\log k})) \) non-ergodic rate

Since \(\tilde{y}^k\) is projected onto \(B_{k-1}\), we can use the definition of \(y^k\) and similar arguments as (87) to show that \(L_k \leq L_g[2\|y^k\| + \rho_{k-1}(R_y + B_g)]\). Now, by the first inequality in (49), and the update of \(\alpha_k\) and \(\beta_k\) in (50), we can show that
\[
\frac{1}{\alpha_k} \left(1 - \frac{\beta_k}{\alpha_k} \right) + \frac{1}{\alpha_k} - L_f - \frac{\rho_k^2M_g^2}{\rho_k - \eta_k}
\geq (1 - \Gamma) (L_f + \rho_k M^2) + \rho_k M^2 - L_g \left[ 2\|y^k\| + \rho_k(R_y + B_g) \right] - \frac{\rho_k M_g^2}{\rho_k - \eta_k}
= \rho_k \left[ (2 - \Gamma) \hat{M}^2 - \frac{M^2}{\gamma} - L_g(R_y + B_g) \right] + (1 - \Gamma) L_f - 2L_g\|y^k\| \geq 0.
\]

Utilizing this inequality and the update rules (50) into (40) of Lemma 3 we can derive
\[
\mathcal{L}(x^{k+1}, s^{k+1}, y) - \mathcal{L}(x, s, y^{k+1}) + \frac{(k+c)^2}{k+c} \cdot \frac{\rho_k}{2} \|g(x^{k+1}) + s^{k+1}\|^2
\leq \frac{k}{k+c} \left[ \mathcal{L}(x^k, s^k, y) - \mathcal{L}(x, s, y^k) + \frac{(k+c-1)^2}{c} \cdot \frac{\rho_k}{2} \|g(x^k) + s^k\|^2 \right]
+ \frac{\tau_k^2}{2\beta_k} \|\tilde{x}^k - x\|^2 - \frac{\tau_k(\tau_k+\beta_k\mu_k)}{2\beta_k} \|\tilde{x}^{k+1} - x\|^2
+ \left( \frac{c}{k+c} \right)^2 \cdot \frac{1}{2\eta_k} \left[ \|\tilde{y}^k - y\|^2 - \|\bar{y}^{k+1} - y\|^2 \right]
- \frac{k(k+c-1)^2}{c(k+c)} \cdot \frac{\rho_k}{2} \|g(x^k) + s^k\|^2.
\]

Since \(y^* \leq \rho_0 R_y\), we have \(y^* \in B_0 \subseteq B_k\). We can thus substitute \((x, s, y) := (x^*, s^*, y^*)\) into (117), and then abbreviate
\[
\begin{aligned}
a_k^2 := \frac{\rho_k}{2} \|g(x^k) + s^k\|^2, \quad b_k^2 := \frac{(k+c)^2\tau_k^2}{2\beta_k} \|\tilde{x}^k - x^*\|^2, \quad d_k^2 := \frac{1}{2\eta_k} \|\tilde{y}^k - y^*\|^2, \\
G_k := \mathcal{L}(x^k, s^k, y^*) - \mathcal{L}(x^*, s^*, y^*) = \mathcal{L}(x^k, s^k, y^*) - P^* \geq 0.
\end{aligned}
\]
Then, we can simplify (117) as:
\[
\tilde{G}_{k+1} + \left(\frac{k+c}{c}\right)^2 a_{k+1}^2 \leq \frac{k}{k+c} \left[ \tilde{G}_k + \left(\frac{k+c-1}{c}\right)^2 a_k^2 \right] + \frac{b_k^2}{(k+c)^2} \tau_k (\tau_k + \beta_k \mu_k) \frac{\|x^k - x^*\|^2}{2} + \left(\frac{c}{k+c}\right)^2 (a_k^2 - d_{k+1}^2) - \frac{k}{c\tau(k+c)} [(k + c - 1)^2 - k(k + c)] a_k^2
\]
which implies that (118) and (122) are the parallel counterparts of (90) and (92) in proof Theorem 3. Therefore, the remaining proof of Theorem 6 is similar to the one in Theorem 3, and we omit the verbatim here.

\[
\Delta_{k+1}^2 := (k+c)^2 \tilde{G}_{k+1} + \left(\frac{k+c}{c}\right)^4 a_{k+1}^2 + b_{k+1}^2 + c^2 d_{k+1}^2 \\
\leq k(k+c) \tilde{G}_k + \frac{k^2(k+c)^2}{c^2} a_k^2 + b_k^2 + c^2 d_k^2 \\
\leq \left[ (k + c - 1)^2 - (c - 2)(k + c - 1) \right] \tilde{G}_k + \left(\frac{k+c-1}{c}\right)^4 - \frac{(c-2)(k+c-1)^2}{c^2} a_k^2 + b_k^2 + c^2 d_k^2 \\
= \Delta_k^2 - (c-2) \left[ (k+c-1) \tilde{G}_k + \left(\frac{k+c-1}{c}\right)^2 a_k^2 \right].
\]
where we have used \(c > 2\) and the following elementary facts:

\[
\begin{align*}
(k(k+c)) &\leq (k+c-1)^2 - (c-2)(k+c-1), \\
k^2(k+c)^2 &\leq (k+c-1)^4 - (c-2)(k+c-1)^3.
\end{align*}
\]
Using (120), and by induction, we can deduce \((k + c - 1)^2 \tilde{G}_k + \left(\frac{k+c-1}{c}\right)^2 a_k^2 \leq \Delta_k^2 \leq \Delta_0^2\). In particular, we obtain
\[
\mathcal{L}(x^k, s^k, y^*) - \mathcal{P}^* \leq \tilde{G}_k \leq \frac{\Delta_k^2}{(k+c-1)^2} \quad \text{and} \quad \|g(x^k) + s^k\| = a_k \sqrt{\frac{2}{\rho_0}} \leq \sqrt{\frac{2}{\rho_0} \Delta_0}.
\]
Furthermore, summing up (120) from \(j := 0\) to \(j := k\), we also get
\[
(c-2) \sum_{j=0}^k \left[ (j+c-1) \tilde{G}_j + \left(\frac{j+c-1}{c}\right)^3 a_j^2 \right] \leq \sum_{j=0}^k (\Delta_j^2 - \Delta_{j+1}^2) \leq \Delta_0^2 < +\infty.
\]
Since \(c > 2\) and \(\tilde{G}_j \geq 0\), if we apply Lemma 4(b), then we can easily show that
\[
\liminf_{k \to \infty} (k^2 \log k) \left[ \tilde{G}_k + \left(\frac{k a_k}{c}\right)^2 \right] = 0.
\]
Notice that (121) and (122) are the parallel counterparts of (90) and (92) in proof Theorem 3. Therefore, the remaining proof of Theorem 6 is similar to the one in Theorem 3, and we omit the verbatim here.
D Proof of Theorem 7: Convergence rates of Algorithms 1 and 2 to solve (CP)

Using the same lines as in the proof of Theorem 1 in Appendix B.2, we can see that (75) still holds with $L$ defined in (51). Substituting $(x, s) := (x^\star, s^\star)$ into (75), we get

$$F(\bar{x}^k) + \langle y, g(\bar{x}^k) + \bar{s}^k \rangle - F^\star \leq \frac{R_0^2(x^\star, y^\star)}{2k}.$$  

Let $R_0(y) := R_0^2(x^\star, y)$, then for any fixed $r > 0$, we have

$$F(\bar{x}^k) - F^\star \leq F(\bar{x}^k) - F^\star + r\|g(\bar{x}^k) + \bar{s}^k\| \leq \frac{1}{2k} \sup \{ R_0(y) : \|y\| \leq r \}. \quad (123)$$

On the other hand, by the saddle-point relation (14), we have

$$F(\bar{x}^k) + \langle y^\star, g(\bar{x}^k) + \bar{s}^k \rangle = L(\bar{x}^k, \bar{s}^k, y^\star) \geq L(x^\star, s^\star, y^\star) = F^\star.$$  

By the Cauchy-Schwarz inequality, the last estimate leads to

$$F(\bar{x}^k) - F^\star \geq -(y^\star, g(\bar{x}^k) + \bar{s}^k) \geq -\|y^\star\| \|g(\bar{x}^k) + \bar{s}^k\|. \quad (124)$$

Substituting (124) into (123), we get

$$(r - \|y^\star\|) \|g(\bar{x}^k) + \bar{s}^k\| \leq \frac{1}{2k} \sup \{ R_0(y) : \|y\| \leq r \}.$$  

Let us choose $r := \|y^\star\| + 1$. Notice that $\bar{s}^k \in K$ due to (20), dom($H$) = $-K$, and that $K$ is convex. Therefore, the last inequality becomes

$$\text{dist}_{-K} (g(\bar{x}^k)) = \inf_{s \in K} \|g(\bar{x}^k) + s\| \leq \|g(\bar{x}^k) + \bar{s}^k\| \leq \frac{1}{2k} \sup \{ R_0^2(y) : \|y\| \leq \|y^\star\| + 1 \} \leq \frac{1}{2k} \left[ \frac{1}{\bar{c}} \|x^0 - x^\star\|^2 + \frac{1}{\bar{c}} (\|y^0\| + \|y^\star\| + 1)^2 \right], \quad (125)$$

Combining (123), (124), and (125), we arrive at the conclusion in Statement (a).

Statement (b) can be proved in a similar way as above, thus we omit the verbatim.

The first part of Statement (c) follows from (90). For the second part, notice that (92) still holds. Applying Lemma 4(c, part(i)) with $u_k := G_k$, $v_k := a_k$, $t_1 := \frac{1}{c}$, and $t_2 := \|y^\star\| + 1$, we get

$$\liminf_{k \to \infty} k \sqrt{\log k} [F(x^k) - F^\star] + \|g(x^k) + s^k\| \leq \liminf_{k \to \infty} k \sqrt{\log k} [L(x^k, s^k, y^\star) - F^\star + (\|y^\star\| + 1)\|g(x^k) + s^k\|] = 0,$$

which is exactly the second part of Statement (c).

The last three statements: Statements (d), (e), and (f), can be proved the same way as the first three statements. We therefore omit the details. \qed
References

1. M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning, pages 214–223, 2017.

2. H. H. Bauschke and P. Combettes. Convex Analysis and Monotone Operators Theory in Hilbert Spaces. Springer-Verlag, 2nd edition, 2017.

3. A. Bayandina, P. Dvurechensky, A. Gasnikov, F. Stonyakin, and A. Titov. Mirror descent and convex optimization problems with non-smooth inequality constraints. In Large-Scale and Distributed Optimization, pages 181–213. Springer, 2018.

4. A. Beck, A. Ben-Tal, N. Guttmann-Beck, and L. Tetruashvili. The CoMirror algorithm for solving nonsmooth constrained convex problems. Oper. Res. Lett., 38(6):493–498, 2010.

5. A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. Robust Optimization. Princeton University Press, 2009.

6. Dimitri P. Bertsekas. Constrained Optimization and Lagrange Multiplier Methods. Athena Scientific, 1996.

7. D. Boob, Q. Deng, and G. Lan. Proximal point methods for optimization with nonconvex functional constraints. arXiv preprint arXiv:1908.02734, 2019.

8. 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(2):251–279, 2015.

9. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends® Mach. Learn., 3(1):1–122, 2011.

10. S. Boyd and L. Vandenberghe. Convex Optimization. University Press, Cambridge, 2004.

11. A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis., 40(1):120–145, 2011.

12. A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. Math. Program., 159(1-2):253–287, 2016.

13. Y. Chen, G. Lan, and Y. Ouyang. Accelerated schemes for a class of variational inequalities. Math. Program., 165(1):113–149, 2017.

14. P. Combettes and J.-C. Pesquet. Signal recovery by proximal forward-backward splitting. In Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pages 185–212. Springer-Verlag, 2011.

15. 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(2):307–330, 2012.

16. L. Condat. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl., 158(4):460–479, 2013.

17. D. Davis. Convergence rate analysis of primal-dual splitting schemes. SIAM J. Optim., 25(3):1912–1943, 2015.

18. J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. Trans. Am. Math. Soc., 82(2):421–439, 1956.

19. S. S. Du and W. Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. arXiv preprint arXiv:1802.01504, 2019.

20. C. Dümmer, S. Forte, M. Takáč, and M. Jaggi. Primal-dual rates and certificates. In Proceedings of the 33rd International Conference on Machine Learning, 2016.

21. J. Eckstein. Parallel alternating direction multiplier decomposition of convex programs. J. Optim. Theory Appl., 80(1):39–62, 1994.

22. 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 Sci., 3(4):1015–1046, 2010.

23. J. E. Esser. Primal-Dual Algorithm for Convex Models and Applications to Image Restoration, Registration and Nonlocal Inpainting. PhD Thesis, University of California, Los Angeles, Los Angeles, USA, 2010.

24. F. Facchinei and J.-S. Pang. Finite-Dimensional Variational Inequalities and Complementarity Problems, volume 1-2. Springer-Verlag, 2003.

25. F. Farnia and D. Tse. A convex duality framework for GANs. In Advances in Neural Information Processing Systems, pages 5248–5258, 2018.

26. O. Fercoq and Z. Qu. Restarting accelerated gradient methods with a rough strong convexity estimate. arXiv preprint arXiv:1608.07338, pages 1–23, 2016.

27. A. V. Fiacco and G. P. McCormick. Nonlinear Programming: Sequential Unconstrained Minimization Techniques. SIAM publications, 1990.

28. R. Fletcher. Practical Methods of Optimization. Wiley, Chichester, 2nd edition, 1987.
29. P. Giselsson, D. M. Doan, T. Keviczky, B. De Schutter, and A. Rantzer. Accelerated gradient methods and dual decomposition in distributed model predictive control. *Automatica*, 49(3):829–833, 2013.
30. T. Goldstein, B. O’Donoghue, S. Setzer, and R. Baraniuk. Fast Alternating Direction Optimization Methods. *SIAM J. Imaging Sci.*, 7(3):1588–1623, 2012.
31. Tom Goldstein, Min Li, and Xiaoming Yuan. Adaptive primal-dual splitting methods for statistical learning and image processing. In *Advances in Neural Information Processing Systems*, pages 2080–2088, 2015.
32. C. Golowich, S. Pattathil, C. Daskalakis, and A. Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. *arXiv preprint arXiv:2002.00557*, 2020.
33. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.
34. M. Grant, S. Boyd, and Y. Ye. Disciplined convex programming. In L. Liberti and N. Maculan, editors, *Global Optimization: From Theory to Implementation, Nonconvex Optimization and its Applications*, pages 155–210. Springer, 2006.
35. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2nd edition, 2009.
36. B. He and X. Yuan. Convergence analysis of primal-dual algorithms for saddle-point problem: from contraction perspective. *SIAM J. Imaging Sci.*, 5:119–149, 2012.
37. B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM J. Numer. Anal.*, 50:700–709, 2012.
38. Y. He and R.-D. C Monteiro. An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems. *SIAM J. Optim.*, 26(1):29–56, 2016.
39. S.-J. Kim and S. Boyd. A minimax theorem with applications to machine learning, signal processing, and finance. *SIAM J. Optim.*, 19(3):1344–1367, 2008.
40. G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.
41. G. Lan and R.-D.C. Monteiro. Iteration complexity of first-order penalty methods for convex programming. *Math. Program.*, 138(1):115–139, 2013.
42. G. Lan and R.-D.C. Monteiro. Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Math. Program.*, 155(1-2):511–547, 2016.
43. G. Lan and Z. Zhou. Algorithms for stochastic optimization with functional or expectation constraints. *arXiv preprint arXiv:1604.00887*, 2019.
44. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5(Jan):27–72, 2004.
45. Z. Li and Y. Xu. Augmented Lagrangian based first-order methods for convex and nonconvex programs: nonergodic convergence and iteration complexity. *arXiv preprint arXiv:2003.08880*, 2020.
46. Q. Lin, R. Ma, and T. Yang. Level-set methods for finite-sum constrained convex optimization. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3112–3121, 2018.
47. Q. Lin, S. Nadarajah, and S. Soheili. A level-set method for convex optimization with a feasible solution path. *SIAM J. Optim.*, 28(4):3290–3311, 2018.
48. T. Lin, C. Jin, and M. Jordan. Near-optimal algorithms for minimax optimization. *arXiv preprint arXiv:2009.08417*, 2020.
49. T. Lin, S. Ma, and S. Zhang. On the global linear convergence of the admm with multiblock variables. *SIAM J. Optim.*, 25(3):1478–1497, 2015.
50. P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Num. Anal.*, 16:964–979, 1979.
51. R.-D.C. Monteiro and B.F. Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM J. Optim.*, 23(1):475–507, 2013.
52. MOSEK-ApS. The MOSEK Optimization Toolbox for MATLAB Manual, version 9.0, 2019.
53. A. Nemirovskii. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.*, 15(1):229–251, 2004.
54. Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983. Translated as Soviet Math. Dokl.
58. Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, 2004.

59. Y. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM J. Optim.*, 16(1):235–249, 2005.

60. Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.

61. Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Math. Program.*, 109(2–3):319–344, 2007.

62. J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering, Springer, 2 edition, 2006.

63. D. O’Connor and L. Vandenberghe. Primal-dual decomposition by operator splitting and applications to image deblurring. *SIAM J. Imaging Sci.*, 7(3):1724–1754, 2014.

64. D. O’Connor and L. Vandenberghe. On the equivalence of the primal-dual hybrid gradient method and Douglas-Rachford splitting. *Math. Program.*, pages 1–24, 2018.

65. B. O’Donoghue and E. Candes. Adaptive restart for accelerated gradient schemes. *Found. Comput. Math.*, 15:715–732, 2015.

66. H. Ouyang, N. He, L. Q. Tran, and A. Gray. Stochastic alternating direction method of multipliers. *JMLR W&CP*, 28:80–88, 2013.

67. Y. Ouyang, Y. Chen, G. Lan, and E. Jr. Pasiliao. An accelerated linearized alternating direction method of multiplier. *SIAM J. Imaging Sci.*, 8(1):644–681, 2015.

68. G. Peyré and M. Cuturi. Computational optimal transport. *Found. Trends® Mach. Learn.*, 11(5-6):355–607, 2019.

69. B. T. Polyak. A general method of solving extremum problems. *Doklady Akademii Nauk SSSR*, 174(1):33–49, 1967.

70. R. A. Polyak, J. Costa, and J. Neyshabouri. Dual fast projected gradient method for quadratic programming. *Optim. Lett.*, 7(4):631–645, 2013.

71. H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

72. R. T. Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematics Series*. Princeton University Press, 1970.

73. R. T. Rockafellar. The multiplier method of Hestenes and Powell applied to convex programming. *J. Optim. Theory Appl.*, 12(6):555–562, 1973.

74. R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.*, 1:97–116, 1976.

75. R.T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer-Verlag, 1997.

76. A. Shapiro and S. Ahmed. On a class of minimax stochastic programs. *SIAM J. Optim.*, 14(4):1237–1249, 2004.

77. R. Shefi and M. Teboulle. On the rate of convergence of the proximal alternating linearized minimization algorithm for convex problems. *EURO J. Comput. Optim.*, 4(1):27–46, 2016.

78. W. Su, S. Boyd, and E. Candes. A differential equation for modeling Nesterov’s accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

79. K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems*, pages 12659–12670, 2019.

80. K. H. Le 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*, 2019.

81. Q. Tran-Dinh, S. Gumussoy, W. Michiels, and M. Diehl. Combining convex-concave decompositions and linearization approaches for solving BMIs, with application to static output feedback. *IEEE Trans. Autom. Control*, 57(6):1377–1390, 2012.

82. Q. Tran-Dinh, I. Necoara, and M. Diehl. Fast inexact decomposition algorithms for large-scale separable convex optimization. *Optimization*, 66:325–356, 2016.

83. Q. Tran-Dinh and Y. Zhu. Non-stationary first-order primal-dual algorithms with faster convergence rates. *arXiv preprint arXiv:1903.05282*, 2020.

84. Quoc Tran-Dinh. Adaptive smoothing algorithms for nonsmooth composite convex minimization. *Comput. Optim. Appl.*, 66(3):425–451, 2016.

85. P. Tseng. Applications of splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control Optim.*, 29:119–138, 1990.

86. P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Submitted to SIAM J. Optim.*, 2008.
88. L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *Advances in Neural Information Processing Systems*, pages 1537–1544, 2005.

89. Y. Xu. Global convergence rates of augmented Lagrangian methods for constrained convex programming. *arXiv preprint arXiv:1711.05812*, 10, 2017.

90. Y. Xu. Primal-dual stochastic gradient method for convex programs with many functional constraints. *arXiv preprint arXiv:1802.02724*, 2019.

91. Y. Xu. Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Math. Program.*, pages 1–46, 2019.

92. Y. Xu. First-order methods for constrained convex programming based on linearized augmented Lagrangian function. *INFORMS J. Optim.*, 2020.

93. Y. Yan, Y. Xu, Q. Lin, W. Liu, and T. Yang. Sharp analysis of epoch stochastic gradient descent ascent methods for min-max optimization. *arXiv preprint arXiv:2002.05509*, 2020.

94. H. E. Yazdandoost and N. S. Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2019.

95. X. Zhang, M. Burger, and S. Osher. A unified primal-dual algorithm framework based on Bregman iteration. *J. Sci. Comput.*, 46(1):20–46, January 2011.