SIAM J. OPTIM.
Vol. 32, No. 4, pp. 2580–2611

NEW PRIMAL-DUAL ALGORITHMS FOR A CLASS OF NONSMOOTH AND NONLINEAR CONVEX-CONCAVE MINIMAX PROBLEMS*

YUZIXUAN ZHU†, DEYI LIU†, AND QUOC TRAN-DINH‡

Abstract. In this paper, we develop new primal-dual algorithms to solve a class of nonsmooth and nonlinear convex-concave minimax problems, which covers many existing and brand-new models as special cases. Our approach relies on a combination of a generalized augmented Lagrangian function, Nesterov's accelerated scheme, and adaptive parameter updating strategies. Our algorithmic framework is single-loop and unifies two important settings: general convex-concave and convex-linear cases. Under mild assumptions, our algorithms achieve $\mathcal{O}(1/k)$ convergence rates through three different criteria: primal-dual gap, primal objective residual, and dual objective residual, where k is the iteration counter. Our rates are both *ergodic* (i.e., on a weighted averaging sequence) and *nonergodic* (i.e., on the last-iterate sequence). These convergence rates can be boosted up to $\mathcal{O}(1/k^2)$ if only one objective term is strongly convex (or, equivalently, its conjugate is L-smooth). To the best of our knowledge, this is the first algorithm achieving optimal rates on the primal last-iterate sequence for convex-linear minimax problems. As a byproduct, we specify our algorithms to solve a general convex cone constrained program with both ergodic and nonergodic rate guarantees. We test our algorithms and compare them with two recent methods on two numerical examples.

Key words. convex-concave minimax problem, primal-dual algorithm, optimal convergence rate, last-iterate convergence rate, Nesterov's accelerated scheme, convex cone constrained program

MSC codes. 90C25, 90C06, 90-08

DOI. 10.1137/21M1408683

1. Introduction.

The goal of this paper is to develop novel primal-dual algorithms to solve the nonsmooth and nonlinear convex-concave minimax problem

$$\min_{x\in\mathbb{R}^p}\max_{y\in\mathbb{R}^n}\Big\{\widetilde{\mathcal{L}}(x,y) := F(x) + \Phi(x,y) - H^*(y)\Big\}, \tag{SP}$$

where F, H, and Φ satisfy the following structures:

1. $F(x) := f(x) + h(x)$, 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;
2. $H : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is proper, closed, and convex, but not necessarily smooth, and $H^*(y) = \sup_s\{\langle y, s\rangle - H(s)\}$ is its Fenchel conjugate;
3. $\Phi : \mathbb{R}^p\times\mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, and convex in x and concave in y.

In this paper, we will focus on two settings: *general convex-concave* Φ and *convex-linear* Φ, i.e., $\Phi(x,y) = \langle g(x), y\rangle$ for some nonlinear function $g : \mathbb{R}^p \to \mathbb{R}^n$. In particular, if $g(x) = Kx$ for a given matrix K, then (SP) reduces to the well-known convex-concave minimax problem involving bilinear objective function.

*Received by the editors March 31, 2021; accepted for publication (in revised form) May 16, 2022; published electronically October 19, 2022.

https://doi.org/10.1137/21M1408683

Funding: This work was partially supported by the Office of Naval Research (ONR) through grant N00014-20-1-2088 and by the National Science Foundation (NSF) through NSF-RTG grant NSF DMS-2134107.

†Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill (UNC), Chapel Hill, NC 27599-3260 USA (zyzx@live.unc.edu, deyi@live.unc.edu).

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

Associated with (SP), we can define the primal-dual problem pair as follows:

$$\text{(P)}\quad \mathcal{P}^\star := \min_{x\in\mathbb{R}^p}\{\mathcal{P}(x) := F(x)+P(x)\},\quad \text{where}\quad P(x) := \max_{y\in\mathbb{R}^n}\{\Phi(x,y)-H^*(y)\},$$

$$\text{(D)}\quad \mathcal{D}^\star := \max_{y\in\mathbb{R}^n}\{\mathcal{D}(y) := D(y)-H^*(y)\},\quad \text{where}\quad D(y) := \min_{x\in\mathbb{R}^p}\{F(x)+\Phi(x,y)\}.$$

If $\Phi(x,y)=\langle g(x),y\rangle$, then $P(x)=H(g(x))$ in (P), and thus (P) also covers the nonlinear compositional convex optimization problem as a special case. In particular, if $H^*(y) := \delta_{\mathcal{K}^*}(y)$, the indicator of the dual cone $\mathcal{K}^*$ of a proper cone $\mathcal{K}$ in $\mathbb{R}^n$, and $\Phi(x,y)=\langle g(x),y\rangle$, then (P) reduces to the following general convex cone constrained program:

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

This problem covers several subclasses such as conic programs and convex programs with nonlinear convex constraints (e.g., quadratically constrained quadratic programs) [2]. Let us first review some representative applications and then discuss the limitations of existing works regarding (SP) and its primal-dual pair (P) and (D).

Representative applications. If $\Phi(x,y)=\langle Kx,y\rangle$, then (SP) already covers various applications in signal and image processing, compressive sensing, machine learning, and statistics; see, e.g., [2, 5, 8, 13, 19]. When Φ is convex-linear or generally convex-concave, it additionally covers many other key applications in different fields. For instance, the kernel matrix learning problem for support vector machines studied in [26, problem (20)] can be formulated into (SP), where Φ is quadratic in x (model parameters) and linear or concave in y (a kernel matrix). Another related problem is the maximum margin clustering application studied in [53], where the coupling objective is linear in y. Various robust optimization models relying on the well-known Wald's max-min formulation can be cast into (SP), where y characterizes a source of uncertainty; see, e.g., [3]. The generative adversarial networks (GANs) problem involving Wasserstein distances studied in [1] can also be formulated as a special case of (SP). This model is also related to optimal transport problems as shown in [16]. Other applications of (SP) in machine learning, (distributionally) robust optimization, game theory, and signal and image processing can be found, e.g., in [14, 15, 24, 38, 39, 42]. It is also worth noting that (SP) and its special case (CP) can serve as subproblems in several nonconvex-concave minimax and nonconvex optimization methods such as proximal-point, inner approximation, and penalty-based schemes; see, e.g., [4, 28, 47].

Limitation of existing work. Methods for solving (SP) and its primal problem (P) when Φ is bilinear are well developed; see, e.g., [2, 5, 9, 11, 13, 20, 21, 35, 36, 41, 43, 46, 48, 49]. However, when Φ is no longer bilinear, algorithms for solving (SP) remain limited; see, e.g., [18, 22, 27, 33, 44, 45, 58]. We find that existing works have the following limitations.

$\diamond$ *Model assumptions.* Gradient-based methods such as [27, 45, 57, 58] require $\nabla_x\Phi$ and $\nabla_y\Phi$ to be uniformly Lipschitz continuous on both x and y (see Assumption 2.4), which unfortunately excludes some important cases, e.g., the convex cone constrained problem (CP), where $\nabla_x\Phi(x,y)=g'(x)^\top y$, which is not uniformly Lipschitz continuous on x for all y (see Assumption 2.3). In addition, if $\Phi(x,y)=\langle g(x),y\rangle$ and H^* is not strongly convex or restricted strongly convex as in [12, 27, 44, 52], then $P(\cdot)$ in (P) can be nonsmooth, which creates several challenges for first-order optimization methods.

$\diamond$ *High per-iteration complexity.* Various methods, including [27, 33, 45, 54, 56, 57], require double loops even when $\Phi(x, y) = \langle g(x), y\rangle$, where the inner loop approximately solves a subproblem, e.g., a penalized or an augmented Lagrangian subproblem in x. These methods (including variants of the alternating minimization algorithm and the alternating direction method of multipliers (ADMM)) can be viewed as inexact first-order schemes to solve (P), where the complexity of each outer iteration is often high. In addition, related parameters such as the inner iteration number are often chosen based on some convergence bounds and may depend on a desired accuracy. This dependence requires sophisticated hyperparameter tuning strategy to achieve good performance, and it is often challenging to implement in practice. There exist very limited single-loop algorithms such as [18, 27, 29, 30] for general convex-concave minimax problems, and [55] for a special case of (CP).

$\diamond$ *Convergence guarantees.* Subgradient and mirror descent-based methods such as [22, 33] often have slow convergence rates compared to gradient and accelerated gradient-based methods [34]. Hitherto, existing works can only show the best known convergence rates on *ergodic* (or *averaging*) sequences, via a gap function (cf. (4)); see, e.g., [7, 18, 22, 30, 31, 32, 33, 45, 55]. It means that the convergence guarantee is based on an averaging or a weighted averaging sequence of all the past iterates. In practical implementation, however, researchers often report performance on the *nonergodic* (or the *last-iterate*) sequence, which may only have asymptotic convergence or suboptimal rate compared to the averaging one. As indicated in [17], the theoretical guarantee on the last-iterate sequence can be significantly slower than an averaging one. To achieve faster convergence rates on the last iterates, as shown in [48, 50], one needs to fundamentally redesign the underlying algorithm. Note that averaging sequences break desired structures of final solutions such as sparsity, low-rankness, or sharp-edged structures required in many applications, including imaging science.

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

Our approach. Problem (SP) is much more challenging to solve than its bilinear case, especially under Assumption 2.3, where $\nabla_x\Phi(\cdot, y)$ is not uniformly Lipschitz continuous in x for all y. Our approach relies on a novel combination of different techniques. First, we introduce a surrogate $\mathcal{L}$ of the Lagrange function $\widetilde{\mathcal{L}}$ in (SP) and a new potential function $\mathcal{L}_\rho$ (see subsection 2.4). The function $\mathcal{L}_\rho$ plays a key role in our convergence analysis. Second, we alternatively minimize $\mathcal{L}_\rho$ w.r.t. its auxiliary variable s and then the primal variable x. The subproblem in x is linearized to use the proximal operator of h and ∇f. Third, we also utilize Nesterov's accelerated momentum step with a new step-size rule to obtain optimal convergence rates. Finally, we exploit a homotopy strategy developed in [46, 48] to dynamically update the involved parameters in an explicit manner.

Our contribution. The contribution of this paper can be summarized as follows:

(a) We develop a novel unified single-loop primal-dual algorithmic framework, Algorithm 1, to solve (SP), (P), and (D) simultaneously, which covers six different variants. We introduce a new potential function $\mathcal{L}_\rho$ for (SP) to analyze the convergence of our algorithms. We establish key properties of ψ_ρ, a component of $\mathcal{L}_\rho$, which could be of independent interest, and can be used to develop other algorithms.

(b) We establish $\mathcal{O}\left(1/k\right)$ optimal convergence rates for different variants of Algorithm 1 in the general convex-concave case and the convex-linear case on the

primal-dual gap, where k is the iteration counter. Our sublinear convergence rates are achieved via both averaging sequences and the primal last-iterate sequence, which we call ergodic and semi-ergodic rates, respectively. In addition, we develop adaptive update rules for our algorithmic parameters in the ergodic case.

(c) When F is strongly convex (i.e., either f or h is strongly convex), by deriving new parameter update rules, we establish $\mathcal{O}\left(1/k^2\right)$ optimal convergence rates for different variants of Algorithm 1 on the primal-dual gap. Again, our convergence rates are achieved via both averaging and primal last-iterate sequences.

(d) As a byproduct, we also obtain the same convergence rates on the primal objective residual and the dual objective residual for both (P) and (D), respectively, in all variants. We also specify Algorithm 1 to solve the general convex cone constrained program (CP), where our optimal convergence rates, both ergodic and nonergodic, on the primal objective residual and primal feasibility violation are established.

Comparison. Problem (SP) can be cast into a special variational inequality problem (VIP) or a maximally monotone inclusion, where several methods can be applied to solve it; see, e.g., [2, 14, 15, 23, 25, 31, 32]. However, the following aspects make our methods stand out from existing works on the minimax setting (SP) and its primal-dual pair (P) and (D). First, the main assumption of the VIP approach is the uniformly Lipschitz continuity of the underlying monotone operator, which unfortunately does not hold for our second setting under Assumption 2.3, and in particular for (CP). Second, our algorithms are different from those in [2, 14, 15, 31, 32], where we focus on nonasymptotically sublinear convergence rates under mild assumptions, instead of asymptotic or linear rates as approaches in [2, 14, 15, 31, 32]. Third, many variants of Algorithm 1 are single-loop as opposed to double-loop ones as in [27, 33, 37, 45, 56, 57]. Note that single-loop algorithms are often easy to implement and extend. Fourth, we do not require $\nabla_y \Phi$ to be uniformly Lipschitz continuous in x for all y as in [27, 45, 56, 57], where the domain of y in our setting can be unbounded. Fifth, compared to other single-loop methods such as in [18, 29, 44], our rates are nonergodic on the primal sequence. To the best of our knowledge, this is the first work establishing such nonergodic rates for the convex-linear case of (SP). Sixth, 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 [2, 14, 15]. Seventh, our convergence guarantees are on three different standard criteria, and in a semi-ergodic sense. Even in the ergodic sense, our parameter updates as well as assumptions are adaptive and also different from those in [18, 29] (see Theorems 2, 3, and 4). Finally, our special setting (CP) remains more general than the one in [54, 55], which can cover conic programs. Our algorithm and its convergence rates stated in Theorem 11 are still new compared to [54, 55]. Our rates include both ergodic and nonergodic ones as opposed to the ergodic rates in [55].

This work is also different from [18, 48] in the following aspects. First, our setting (SP) is much more general than the one in [48] both in terms of model and structured assumptions. Second, [48, Algorithm 1] can be considered as a special case of the variant (33) when F is non–strongly convex. However, when F is strongly convex, [48, Algorithm 2] is really different from both (30) and (33). Except for this similarity, other results in this paper are new compared to [48]. Third, [18] only studies ergodic

convergence under Assumption 2.4, which is similar to Theorem 2. Nevertheless, [18] can exploit the gradient $\nabla_y\Phi$ to avoid the proximal operator of $-\Phi(x,\cdot)$. However, under Assumption 2.3, [18] requires a linesearch-type procedure (see [18, eqn. (4.9)]) to achieve convergence guarantees. Finally, our semi-ergodic convergence rates are new compared to [18].

Paper outline. The rest of this paper is organized as follows. Section 2 recalls some basic concepts used in this paper and introduces our new potential function. Section 3 develops our unified algorithmic framework, Algorithm 1, for solving (SP) and its primal and dual formulations. Section 4 analyzes the ergodic convergence rates of Algorithm 1, while section 5 is devoted to studying its semi-ergodic convergence rates. Section 6 specifies our method to (CP). Section 7 provides two numerical examples to verify our theoretical results. For clarity of presentation, all technical proofs are deferred to the appendices.

2. Background, assumptions, and new potential function.

This section recalls some necessary concepts and tools used in this paper and states our main assumptions.

2.1. Basic concepts.

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

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

2.2. Fundamental assumptions.

The algorithms developed in this paper rely on the following assumptions imposed on (SP) and its primal and dual forms (P) and (D).

Assumption 2.1. There exists $(x^\star, y^\star) \in \mathrm{dom}(F) \times \mathrm{dom}(H^*)$ of (SP) such that

$$\widetilde{\mathcal{L}}(x^\star, y) \leq \widetilde{\mathcal{L}}(x^\star, y^\star) \leq \widetilde{\mathcal{L}}(x, y^\star)\ \ \forall (x, y) \in \mathrm{dom}(F) \times \mathrm{dom}(H^*). \tag{1}$$

Moreover, $\mathrm{dom}(F) \times \mathrm{dom}(H^*) \subseteq \mathrm{dom}(\Phi)$ and $\widetilde{\mathcal{L}}(x^\star, y^\star)$ is finite.

Assumption 2.1 is standard in convex-concave minimax problems. It guarantees strong duality $\mathcal{P}^\star = \mathcal{D}^\star = \widetilde{\mathcal{L}}(x^\star, y^\star)$ and the existence of solutions for (P) and (D); see, e.g., [40].

Assumption 2.2. The function Φ is continuously differentiable; f, h, and H are proper, closed, and convex; and $F := f + h$. Moreover, f is L_f-smooth with $L_f \in [0, +\infty)$.

Together with Assumptions 2.1 and 2.2, we consider two settings corresponding to Assumptions 2.3 and 2.4 below. We treat them separately in our analysis.

Assumption 2.3. $\Phi(x,y) = \langle g(x), y\rangle$ is convex in x for any $y \in \mathrm{dom}(H^*)$ and linear in y for any $x \in \mathrm{dom}(F)$. In addition, for any $x, \hat{x} \in \mathrm{dom}(F)$ and $y \in \mathrm{dom}(H^*)$, it satisfies

$$\left\{\begin{array}{ll} \|\nabla_x\Phi(\hat{x},y) - \nabla_x\Phi(x,y)\| = \|[g'(\hat{x}) - g'(x)]^\top y\| & \leq L_{11}\|y\|\|x-\hat{x}\|, \\ \|\nabla_y\Phi(\hat{x},y) - \nabla_y\Phi(x,y)\| = \|g(\hat{x}) - g(x)\| & \leq L_{21}\|\hat{x}-x\|, \end{array}\right. \tag{2}$$

where $L_{11}, L_{21} \in [0,+\infty)$ are given Lipschitz constants.

Assumption 2.4. $\Phi(\cdot, y)$ is convex in x for any $y \in \mathrm{dom}(H^*)$ and $\Phi(x,\cdot)$ is concave in y for any $x \in \mathrm{dom}(F)$. In addition, for any $x, \hat{x} \in \mathrm{dom}(F)$ and $y, \hat{y} \in \mathrm{dom}(H^*)$, it satisfies

$$\left\{\begin{array}{l} \|\nabla_x\Phi(\hat{x},y) - \nabla_x\Phi(x,y)\| \leq L_{11}\|x-\hat{x}\|, \\ \|\nabla_y\Phi(\hat{x},\hat{y}) - \nabla_y\Phi(x,y)\| \leq L_{21}\|\hat{x}-x\| + L_{22}\|\hat{y}-y\|, \end{array}\right. \tag{3}$$

where $L_{11}, L_{21}, L_{22} \in [0,+\infty)$ are given Lipschitz constants.

Assumption 2.4 is widely used in convex-concave minimax problems; see, e.g. [18, 27, 29, 45, 56, 57]. Clearly, if $\Phi(x,y) = \langle g(x), y\rangle$ as in Assumption 2.3, then it satisfies the last condition of (3) with $L_{22} = 0$. However, the first condition of (2) in Assumption 2.3 is weaker than the first line of (3) in Assumption 2.4 if $\mathrm{dom}(H^*)$ is not bounded. Hence, we treat two settings corresponding to these two assumptions separately in this paper. Clearly, if $\Phi(x,y) = \langle Kx, y\rangle$ is bilinear, then it automatically satisfies both Assumptions 2.3 and 2.4.

2.3. Optimality condition and gap function. In view of Assumptions 2.1 and 2.2, there exists a saddle-point $(x^\star, y^\star) \in \mathrm{dom}(F)\times\mathrm{dom}(H^*)$ of (SP) that satisfies

$$0 \in \partial F(x^\star) + \nabla_x\Phi(x^\star, y^\star) \quad \text{and} \quad 0 \in \nabla_y\Phi(x^\star, y^\star) - \partial H^*(y^\star).$$

To characterize saddle-points of (SP), we define the following gap function (see [5, 14, 15, 33]):

$$\mathcal{G}_{\mathcal{Z}}(x,y) := \sup\big\{\widetilde{\mathcal{L}}(x,\hat{y}) - \widetilde{\mathcal{L}}(\hat{x},y) : \hat{x} \in \mathcal{X},\ \hat{y} \in \mathcal{Y}\big\}, \tag{4}$$

where $\mathcal{X} \subseteq \mathrm{dom}(F)$ and $\mathcal{Y} \subseteq \mathrm{dom}(H^*)$ are two nonempty and closed subsets such that $\mathcal{Z} := \mathcal{X}\times\mathcal{Y}$ contains a saddle-point $(x^\star, y^\star)$. It is clear that $\mathcal{G}_{\mathcal{Z}}(x,y) \geq 0$ for any $(x,y) \in \mathrm{dom}(F)\times\mathrm{dom}(H^*)$. If $(x^\star, y^\star)$ is a saddle-point of (SP), then $\mathcal{G}_{\mathcal{Z}}(x^\star, y^\star) = 0$.

2.4. Potential function and its key properties. One of the main steps to develop our algorithms is to build an appropriate potential function, which is formalized as follows. First, we lower bound H^* using its Fenchel conjugate, i.e., $H^*(y) \geq \langle s, y\rangle - H(s)$ for any $s \in \mathrm{dom}(H)$. Consequently, we can upper bound the Lagrange function $\widetilde{\mathcal{L}}$ of (SP) by

$$\mathcal{L}(x,s,y) := F(x) + H(s) + \Phi(x,y) - \langle s, y\rangle. \tag{5}$$

Clearly, for any $(x,s,y)\in\mathrm{dom}(F)\times\mathrm{dom}(H)\times\mathrm{dom}(H^*)$, we have

$$\widetilde{\mathcal{L}}(x,y)\le\mathcal{L}(x,s,y)\quad\text{and}\quad\widetilde{\mathcal{L}}(x,y)=\mathcal{L}(x,s,y)\ \text{iff}\ s\in\partial H^*(y). \tag{6}$$

As a result, for $s^\star:=\nabla_y\Phi(x^\star,y^\star)\in\partial H^*(y^\star)$, (1) implies that

$$\widetilde{\mathcal{L}}(x^\star,y)\le\mathcal{L}(x^\star,s^\star,y)\le\mathcal{L}(x^\star,s^\star,y^\star)=\widetilde{\mathcal{L}}(x^\star,y^\star)\le\widetilde{\mathcal{L}}(x,y^\star)\le\mathcal{L}(x,s,y^\star). \tag{7}$$

Next, let us introduce our *potential function* as follows:

$$\mathcal{L}_\rho(x,s,y):=F(x)+H(s)+\psi_\rho(x,s,y), \tag{8}$$

where $\rho>0$ is a given parameter and $\psi_\rho(\cdot)$ is defined as follows:

$$\psi_\rho(x,s,y):=\max_{u\in\mathbb{R}^n}\Big\{\Phi(x,u)-\langle s,u\rangle-\tfrac{1}{2\rho}\|u-y\|^2\Big\}. \tag{9}$$

We also denote by $u^*_\rho(x,s,y):=\mathrm{prox}_{-\rho\Phi(x,\cdot)}(y-\rho s)$ the unique solution of the maximization problem in (9). In particular, if $\Phi(x,y)=\langle g(x),y\rangle$, which is convex-linear, then $\psi_\rho(x,s,y)=\langle g(x)-s,y\rangle+\frac{\rho}{2}\|g(x)-s\|^2$ and $u^*_\rho(x,s,y)=y+\rho(g(x)-s)$, which are explicitly given. In fact, (8) can be viewed as a generalized augmented Lagrangian function of (SP); see [40].

We first prove the following lemma in Appendix A, which will be used for our analysis. However, we believe that this result by itself is of independent interest and can be used to develop other algorithms for solving (SP). We therefore state it in the main text.

LEMMA 1. *Let ψ_ρ and u^*_ρ be defined by* (9)*, and Assumption* 2.2 *and either Assumption* 2.3 *or Assumption* 2.4 *hold. Then ψ_ρ is convex in x if $u^*_\rho(x,s,y)\in\mathrm{dom}(H^*)$, convex in s for given $y\in\mathrm{dom}(H^*)$, and concave in y for given $(x,s)\in\mathrm{dom}(F)\times\mathrm{dom}(H)$. Moreover, $\nabla_x\psi_\rho(x,s,y)=\nabla_x\Phi(x,u^*_\rho(x,s,y))$ and $\nabla_s\psi_\rho(x,s,y)=-u^*_\rho(x,s,y)$.*

Let us define

$$\begin{aligned}\ell_\rho(\hat{x},\hat{s};x,s,y)&:=\psi_\rho(x,s,y)+\langle\nabla_x\psi_\rho(x,s,y),\hat{x}-x\rangle+\langle\nabla_s\psi_\rho(x,s,y),\hat{s}-s\rangle,\\ \Delta_\rho(\hat{x},\hat{s};x,s,y)&:=\psi_\rho(\hat{x},\hat{s},y)-\ell_\rho(\hat{x},\hat{s};x,s,y).\end{aligned} \tag{10}$$

Then, for $\hat{x},x\in\mathrm{dom}(F)$, $\hat{s},s\in\mathrm{dom}(H)$, $y,\hat{y}\in\mathrm{dom}(H^*)$, and $\rho,\hat{\rho}>0$, we have

$$\begin{cases}\Phi(x,\hat{y})-\langle s,\hat{y}\rangle\ \le\ \psi_\rho(x,s,y)+\frac{1}{2\rho}\|\hat{y}-y\|^2,\\ \psi_\rho(x,s,y)\qquad\ \le\ \psi_{\hat{\rho}}(x,s,y)+\frac{(\rho-\hat{\rho})}{2\rho\hat{\rho}}\|u^*_\rho(x,s,y)-y\|^2,\\ \ell_\rho(\hat{x},\hat{s};x,s,y)\ \le\ \Phi(\hat{x},u^*_\rho(x,s,y))-\langle\hat{s},u^*_\rho(x,s,y)\rangle-\frac{1}{2\rho}\|u^*_\rho(x,s,y)-y\|^2.\end{cases} \tag{11}$$

In particular, if $\Phi(x,y)=\langle g(x),y\rangle$, then $\psi_\rho(x,s,y)=\psi_{\hat{\rho}}(x,s,y)+\frac{(\rho-\hat{\rho})}{2\rho^2}\|u^*_\rho(x,s,y)-y\|^2$. Moreover, if $u^*_\rho(x,s,y)\in\mathrm{dom}(H^*)$, then for $\hat{x},x\in\mathrm{dom}(F)$, $\hat{s},s\in\mathrm{dom}(H)$, and $y,\hat{y}\in\mathrm{dom}(H^*)$, we also have

$$\begin{aligned}&\|u^*_\rho(\hat{x},\hat{s},\hat{y})-u^*_\rho(x,s,y)\|\\ &\qquad\le\tfrac{\rho}{1+\mu_y\rho}\big[L_{21}\|x-\hat{x}\|+\|s-\hat{s}\|\big]+\tfrac{1}{1+\rho\mu_y}\|\hat{y}-y\|,\\ &\tfrac{(1+\rho\mu_y)}{2\rho}\|u^*_\rho(\hat{x},\hat{s},y)-u^*_\rho(x,s,y)\|^2\\ &\qquad\le\Delta_\rho(\hat{x},\hat{s};x,s,y)\\ &\qquad\le\tfrac{\mathbf{L}_{11}}{2}\|\hat{x}-x\|^2+\tfrac{(1+\rho L_{22})}{2\rho}\|u^*_\rho(\hat{x},\hat{s},y)-u^*_\rho(x,s,y)\|^2,\end{aligned} \tag{12}$$

where μ_y is the strong concavity parameter of $\Phi(x,\cdot)$, $\mathbf{L}_{11}:=L_{11}\|u^*_\rho(x,s,y)\|$ and $L_{22}=0$ if Assumption 2.3 *holds, and $\mathbf{L}_{11}:=L_{11}$ if Assumption* 2.4 *holds.*

3. A unified primal-dual algorithmic framework. We now develop a new primal-dual algorithmic framework to solve (SP) and its primal-dual forms (P) and (D).

3.1. Primal and dual updates. Our main idea is to exploit the potential function $\mathcal{L}_\rho$ defined in (8) to measure the progress of an iterate sequence $\{(x^k, \hat{y}^k)\}$ generated by the proposed algorithm. Since this function not only involves x but also the dual variable y and the auxiliary variable s, we also need to update them in an alternating manner.

Primal steps. First, given $\hat{x}^k \in \mathrm{dom}(F)$ and $\hat{y}^k \in \mathrm{dom}(H^*)$, we minimize $\mathcal{L}_{\rho_k}(\hat{x}^k, s, \hat{y}^k)$ w.r.t. s to obtain s^{k+1}. By combining this step and (9) and exchanging the min-max, we get

$$\left\{\begin{array}{l} u^{k+1} := \mathrm{argmin}_u\big\{H^*(u) - \Phi(\hat{x}^k, u) + \frac{1}{2\rho_k}\|u - \hat{y}^k\|^2\big\} = \mathrm{prox}_{\rho_k(H^*(\cdot) - \Phi(\hat{x}^k,\cdot))}\left(\hat{y}^k\right), \\ s^{k+1} := \nabla_y \Phi(\hat{x}^k, u^{k+1}) - \frac{1}{\rho_k}(u^{k+1} - \hat{y}^k). \end{array}\right. \tag{13}$$

With s^{k+1} obtained from (13), we minimize $\mathcal{L}_{\rho_k}(x, s^{k+1}, \hat{y}^k)$ w.r.t. x to obtain x^{k+1}. However, minimizing this function directly is difficult. We instead linearize $f(\cdot) + \psi_{\rho_k}(\cdot, s^{k+1}, \hat{y}^k)$ at $\hat{x}^k$ and then minimize the surrogate of $\mathcal{L}_{\rho_k}$ as

$$\begin{aligned} x^{k+1} &:= \mathrm{argmin}_x \big\{h(x) + \langle \nabla f(\hat{x}^k) + \nabla_x \psi_{\rho_k}(\hat{x}^k, s^{k+1}, \hat{y}^k), x - \hat{x}^k\rangle + \tfrac{L_k}{2}\|x - \hat{x}^k\|^2\big\} \\ &= \mathrm{prox}_{h/L_k}\big(\hat{x}^k - \tfrac{1}{L_k}\big[\nabla f(\hat{x}^k) + \nabla_x \psi_{\rho_k}(\hat{x}^k, s^{k+1}, \hat{y}^k)\big]\big), \end{aligned} \tag{14}$$

where $L_k > 0$ is a given parameter, which will be determined later.

Dual step. As analyzed in Lemma 13, we can update $\hat{y}^k$ with a step-size $\eta_k \geq 0$ as

$$\hat{y}^{k+1} := \hat{y}^k + \eta_k\big(\nabla_y \psi_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) - (1 - \tau_k)\nabla_y \psi_{\rho_{k-1}}(x^k, s^k, \hat{y}^k)\big). \tag{15}$$

Here, we allow $\eta_k = 0$, leading to $\hat{y}^k$ being fixed at y^0 as $\hat{y}^k := y^0$ for all $k \geq 0$ (see (30)).

Dual averaging step. Given $\tilde{y}^k, \tau_k \in (0, 1]$, and u^{k+1} in (13), we update $\tilde{y}^{k+1}$ as

$$\tilde{y}^{k+1} := (1 - \tau_k)\tilde{y}^k + \tau_k u^{k+1}. \tag{16}$$

While the primal step is key to our algorithms, the dual step may not be required as in the variant (30) below. This step is only required in the semi-ergodic variants.

3.2. The full algorithm. To explicitly express our algorithm, we note that

$$\nabla_x \psi_{\rho_k}(\hat{x}^k, s^{k+1}, \hat{y}^k) = \nabla_x \Phi(\hat{x}^k, u^{k+1}) \quad \text{and} \quad \nabla_y \psi_\rho(x, s, \hat{y}^k) = \tfrac{1}{\rho}(u^*_\rho(x, s, \hat{y}^k) - \hat{y}^k),$$

where $u^*_\rho(x, s, \hat{y}^k) = \mathrm{prox}_{-\rho\Phi(x,\cdot)}(\hat{y}^k - \rho s)$. To update $\hat{x}^k$, we apply Nesterov's accelerated scheme [34] as $\hat{x}^{k+1} := x^{k+1} + \beta_{k+1}(x^{k+1} - x^k)$ for an appropriate parameter $\beta_{k+1} \geq 0$.

Finally, putting all the above steps together, we obtain a complete single-loop primal-dual first-order algorithmic framework for solving (SP) as specified in Algorithm 1.

4. Ergodic convergence rates. Let us first analyze the ergodic convergence rates of Algorithm 1 for both general convex-concave and convex-linear settings.

Algorithm 1 (Unified single-loop primal-dual first-order algorithmic framework).

1: **Initialization:** Choose an initial point $(x^0, y^0) \in \mathrm{dom}(F) \times \mathrm{dom}(H^*)$.
2: Set $\hat{x}^0 := x^0$, $\hat{y}^0 := y^0$, $\breve{u}^0 := y^0$, and $s^0 := 0$. Set $\tilde{y}^0 := y^0$ for Theorem 6 or 8.
3: Choose τ_0, L_0, ρ_0, and η_0 according to Theorem 2, 3, 4, 6, or 8.
4: **For** $k := 0$ **to** $k_{\max}$
5: Update τ_k, L_k, ρ_k, η_k, and β_k as in Theorem 2, 3, 4, 6, or 8 (consistent with step 3).
6: Compute $u^{k+1} := \mathrm{prox}_{\rho_k(H^*(\cdot)-\Phi(\hat{x}^k,\cdot))}\big(\hat{y}^k\big)$ and $s^{k+1} := \nabla_y\Phi(\hat{x}^k, u^{k+1}) - \frac{1}{\rho_k}(u^{k+1} - \hat{y}^k)$.
7: Compute $x^{k+1} := \mathrm{prox}_{h/L_k}\big(\hat{x}^k - \frac{1}{L_k}[\nabla f(\hat{x}^k) + \nabla_x\Phi(\hat{x}^k, u^{k+1})]\big)$.
8: Update $\hat{x}^{k+1} := x^{k+1} + \beta_{k+1}(x^{k+1} - x^k)$.
9: Compute $\breve{u}^{k+1} := \mathrm{prox}_{-\rho_k\Phi(x^{k+1},\cdot)}\big(\hat{y}^k - \rho_k s^{k+1}\big)$.
10: Update $\hat{y}^{k+1} := \hat{y}^k + \frac{\eta_k}{\rho_k}(\breve{u}^{k+1} - \hat{y}^k) - \frac{(1-\tau_k)\eta_k}{\rho_{k-1}}(\breve{\breve{u}}^k - \hat{y}^k)$.
11: Compute $\breve{\breve{u}}^{k+1} := \mathrm{prox}_{-\rho_k\Phi(x^{k+1},\cdot)}\big(\hat{y}^{k+1} - \rho_k s^{k+1}\big)$.
12: Update $\tilde{y}^{k+1} := (1-\tau_k)\tilde{y}^k + \tau_k u^{k+1}$ for the variants in Theorems 6 and 8.
EndFor

4.1. The general convex-concave case. Let us fix $\tau_k := 1$ and $\hat{x}^k := x^k$ (i.e., $\beta_k = 0$) in Algorithm 1. In this case, the main steps, steps 6 to 11, of Algorithm 1 reduce to

$$\left\{\begin{array}{ll} u^{k+1} &:= \mathrm{prox}_{\rho_k(H^*(\cdot)-\Phi(x^k,\cdot))}\left(\hat{y}^k\right),\\ s^{k+1} &:= \nabla_y\Phi(x^k, u^{k+1}) - \frac{1}{\rho_k}(u^{k+1} - \hat{y}^k),\\ x^{k+1} &:= \mathrm{prox}_{h/L_k}\big(x^k - \frac{1}{L_k}[\nabla f(x^k) + \nabla_x\Phi(x^k, u^{k+1})]\big),\\ \hat{y}^{k+1} &:= \hat{y}^k + \frac{\eta_k}{\rho_k}(\mathrm{prox}_{-\rho_k\Phi(x^{k+1},\cdot)}\left(\hat{y}^k - \rho_k s^{k+1}\right) - \hat{y}^k). \end{array}\right. \tag{17}$$

In fact, the first line of (17) requires computing the proximal operator of $H^*(\cdot) - \Phi(x^k, \cdot)$, which could be computationally expensive if Φ is nonlinear in y. To overcome this issue, we can also linearize $\mathcal{L}_{\rho_k}(\hat{x}^k, s, \hat{y}^k)$ around $\hat{s}^k$ to decompose the first and second lines of (17) into the two alternating steps:

$$u^{k+1} := \mathrm{prox}_{-\rho_k\Phi(\hat{x}^k,\cdot)}(\hat{y}^k - \rho_k\hat{s}^k) \quad \text{and} \quad s^{k+1} := \mathrm{prox}_{H/\hat{L}_k}\big(\hat{s}^k + \hat{L}_k^{-1}u^{k+1}\big).$$

Here, $\hat{s}^k$ and $\hat{L}_k$ can be updated similarly to $\hat{x}^k$ and L_k, respectively. This decomposition allows us to use the proximal operators of $-\Phi(x^k, \cdot)$ and H separately. The convergence analysis of this variant is very similar to that of (17). However, to avoid overloading the paper, we skip it here and only focus on (17). Compared to [18], Algorithm 1 uses the proximal operator of $-\Phi(x, \cdot)$ instead of the gradient $\nabla_y\Phi(x, \cdot)$ as in [18].

Our goal in this section is to establish convergence rates of Algorithm 1, including the variant (17), on the following ergodic (i.e., weighted averaging) sequence $\{(\bar{x}^k, \bar{y}^k)\}$:

$$\bar{x}^k := \frac{1}{\Sigma_k}\sum_{j=0}^{k}\eta_j x^{j+1}, \quad \bar{y}^k := \frac{1}{\Sigma_k}\sum_{j=0}^{k}\eta_j u^{j+1}, \quad \text{where} \quad \Sigma_k := \sum_{j=0}^{k}\eta_j. \tag{18}$$

Here, η_j for $j = 0, \ldots, k$ are given weights.

The following theorem states both $\mathcal{O}(1/k)$- and $\mathcal{O}\left(1/k^2\right)$-ergodic convergence rates of the variant (17) under Assumption 2.4, whose proof can be found in Appendix C.1.

THEOREM 2. *Suppose that Assumptions* 2.1, 2.2, *and* 2.4 *hold for* (SP). *Let* $\{(x^k, y^k)\}$ *be generated by the variant* (17) *of Algorithm* 1 *and let* $\{(\bar{x}^k, \bar{y}^k)\}$ *be defined by* (18).

- *If* $\mu_F = 0$ (*i.e., F is only convex*), *then we fix* $\rho_k := \rho_0 > 0$ *for all* $k \geq 0$.
- *If* $\mu_F > 0$ (*i.e., F is strongly convex*) *and* $L_{22} = 0$, *then we update* $\rho_{k+1} := \frac{2b_k\rho_k}{\nu+\sqrt{\nu^2+8L_{21}^2 b_k\rho_k}}$, *where* $\nu := L_{11} + L_f - \mu_f$, $b_k := \nu + \mu_F + 2L_{21}^2\rho_k$, *and* $\rho_0 \geq \frac{\mu_F+4\nu}{16L_{21}^2}$.

Let L_k *and* η_k *be updated, respectively, by*

$$L_k := L_{11} + L_f + L_{21}^2(2 + \rho_k L_{22})\rho_k \quad \text{and} \quad \eta_k := \frac{\rho_k}{2}. \tag{19}$$

Then, with $\mathcal{G}_{\mathcal{Z}}$ *defined by* (4), *for all* $k \geq 0$, *we have*

$$\mathcal{G}_{\mathcal{Z}}(\bar{x}^k, \bar{y}^k) \leq \frac{1}{2S_k} \cdot \sup_{(x,y)\in\mathcal{Z}} \left\{\rho_0(L_0 - \mu_f)\|x - x^0\|^2 + 2\|y - y^0\|^2\right\}, \tag{20}$$

where $S_k := \rho_0(k+1)$ *if* $\mu_F = 0$, *and* $S_k := (k+1)\left[\rho_0 + \frac{\mu_F}{16L_{21}^2}(k+2)\right]$ *if* $\mu_F > 0$.

When F is strongly convex, Theorem 2 achieves $\mathcal{O}\left(1/k^2\right)$ rate if $L_{22} = 0$. If $L_{22} > 0$, then the rate of the variant (17) stated in Theorem 2 could be slower than $\mathcal{O}\left(1/k^2\right)$. Hence, one needs to modify our update rules and adapt the analysis to obtain a rate of $\mathcal{O}\left(1/k^2\right)$.

4.2. The convex-linear case. Next, we analyze the $\mathcal{O}(1/k)$-ergodic convergence of Algorithm 1 when $\Phi(x, y) := \langle g(x), y\rangle$ (convex-linear) under Assumption 2.3. In this case, the main steps, steps 6 to 11, of Algorithm 1 reduce to

$$\begin{cases} u^{k+1} := \operatorname{prox}_{\rho_k H^*(\cdot)}\left(\hat{y}^k + \rho_k g(x^k)\right), \\ x^{k+1} := \operatorname{prox}_{h/L_k}\left(x^k - \frac{1}{L_k}[\nabla f(x^k) + g'(x^k)^\top u^{k+1}]\right), \\ \hat{y}^{k+1} := \hat{y}^k + \eta_k\left[g(x^{k+1}) - g(x^k) + \frac{1}{\rho_k}(u^{k+1} - \hat{y}^k)\right]. \end{cases} \tag{21}$$

The following theorem states the convergence of (21), whose proof is given in Appendix C.3.

THEOREM 3. *Suppose that Assumptions* 2.1, 2.2, *and* 2.3 *hold for* (SP) *and* $\dot{y} \in \partial H(\dot{s})$ *for a given* $\dot{s} \in \operatorname{ri}(\operatorname{dom}(H))$. *Let* $\{(x^k, y^k)\}$ *be generated by the variant* (21) *of Algorithm* 1 *using* $\rho_k := \frac{2}{L_k}$ *and* $\eta_k := \frac{1}{L_k}$ *for all* $k \geq 0$, *and* L_k *is adaptively updated by*

$$L_k := L_{11}[\|\dot{y}\| + \|\hat{y}^k - \dot{y}\|] + \sqrt{2L_{11}\|g(x^k) - \dot{s}\|} + L_f + 2L_{21}. \tag{22}$$

Let $\{(\bar{x}^k, \bar{y}^k)\}$ *be an ergodic sequence defined by* (18). *Then, for all* $k \geq 0$, *we have*

$$\mathcal{G}_{\mathcal{Z}}(\bar{x}^k, \bar{y}^k) \leq \frac{\bar{L}}{2(k+1)} \cdot \sup_{(x,y)\in\mathcal{Z}} \left\{\|x - x^0\|^2 + \|y - y^0\|^2\right\}, \tag{23}$$

where $\bar{L}$ *is a given constant explicitly defined as in* (52) *in Lemma* 15.

Alternatively, we consider the case when F in (SP) is strongly convex with $\mu_F := \mu_f + \mu_h > 0$. The following theorem establishes an $\mathcal{O}\left(1/k^2\right)$-ergodic convergence rate of the variant (21) of Algorithm 1, whose proof is given in Appendix C.4.

THEOREM 4. *Suppose that Assumptions* 2.1, 2.2, *and* 2.3 *hold for* (SP). *Suppose further that* F *in* (SP) *is* μ_F*-strongly convex with* $\mu_F := \mu_f + \mu_h > 0$ *and* $\dot{y} \in \partial H(\dot{s})$ *for a given* $\dot{s} \in \mathrm{ri}\,(\mathrm{dom}(H))$. *Given* $\rho_0 > 0$ *and* $\kappa > 0$, *we first choose the initial parameters as*

$$\eta_0 := \tfrac{\rho_0}{2} \quad \text{and} \quad L_0 := L_f + A_0 + \rho_0(B_0 + 2L_{21}^2), \tag{24}$$

where $A_0 := L_{11}[\|\dot{y}\| + \|\hat{y}^0 - \dot{y}\|]$ *and* $B_0 := L_{11}\|g(x^0) - \dot{s}\|$. *Let* $\{(x^k, y^k)\}$ *be generated by the variant* (21) *of Algorithm* 1 *using the following parameter update rules:*

$$\begin{cases} \eta_k := \frac{\rho_k}{2}, \quad L_k := L_f + A_k + \rho_k(B_k + 2L_{21}^2), \\ \rho_k := \frac{2\rho_{k-1}(L_{k-1}+\mu_h)}{(A_k+L_f-\mu_f)+\sqrt{(A_k+L_f-\mu_f)^2+4\rho_{k-1}(L_{k-1}+\mu_h)(B_k+2L_{21}^2)}}, \end{cases} \tag{25}$$

where $\xi_k := L_{11}[\|\dot{y}\| + \|\hat{y}^k - \dot{y}\|]$ *and* $\zeta_k := L_{11}\|g(x^k) - \dot{s}\|$; *and* A_k *and* B_k *are updated by*

$$A_k := \begin{cases} A_{k-1} & \text{if } \xi_k < A_{k-1}, \\ \max\{A_{k-1} + \kappa, \xi_k\} & \text{otherwise} \end{cases} \quad \text{and}$$
$$B_k := \begin{cases} B_{k-1} & \text{if } \zeta_k < B_{k-1}, \\ \max\{B_{k-1} + \kappa, \zeta_k\} & \text{otherwise.} \end{cases} \tag{26}$$

Let $\{(\bar{x}^k, \bar{y}^k)\}$ *be defined by* (18). *Then, for all* $k \geq \bar{k}_0$, *the following bound holds:*

$$\mathcal{G}_{\mathcal{Z}}(\bar{x}^k, \bar{y}^k) \leq \frac{2(\bar{k}_0 + 1)^2}{P_0(k - \bar{k}_0)^2} \cdot \sup_{(x,y)\in\mathcal{Z}} \Big\{\rho_0(L_0 - \mu_f)\|x - x^0\|^2 + 2\|y - y^0\|^2\Big\}, \tag{27}$$

where $\bar{k}_0$ *and* P_0 *are two positive constants explicitly defined in Lemma* 17 *of Appendix* C.2.

Note that the choice of κ in Theorem 4 will affect the value of $\bar{k}_0$. Let us roughly explain how $\bar{k}_0$ depends on κ, $\|x^0 - x^\star\|$, $\|y^0 - y^\star\|$, $\|g(x^\star) - g(x^0)\|$, L_{11}, L_{21}, and L_f. Let $\dot{s} = g(x^0)$ and $\hat{y}_0 = \dot{y}$. To simplify our explanation, we define $L_{\max} := \max\{L_{11}, L_{21}, L_f\}$ and $M_{\max} := \max\{\|x^0 - x^\star\|, \|y^0 - y^\star\|, \|g(x^\star) - g(x^0)\|\}$. We also choose $\eta_0 = \frac{\rho_0}{2} = \mathcal{O}(1/L_{\max})$. Then we have $L_0 = \mathcal{O}(L_{\max})$ from (24), $C_1 = \mathcal{O}(M_{\max})$ and $C_2 = \mathcal{O}(L_{\max}M_{\max} + \kappa)$ from (56), and $\bar{L} = \mathcal{O}(L_{\max}M_{\max})$ from (52). Therefore, from (61) we obtain $\bar{k}_0 = \mathcal{O}\big(L_{\max}M_{\max}\big(\frac{L_{\max}M_{\max}+\kappa}{\kappa}\big)^2\big)$. Suppose that we choose $\kappa := \mathcal{O}(L_{\max}M_{\max})$. Then we obtain the convergence rate guarantee of Algorithm 1 starting from $\bar{k}_0 := \mathcal{O}(L_{\max}M_{\max})$.

Unlike existing works, Assumption 2.3 associated with Φ is challenging to handle due to $\mathbf{L}_{11} = L_{11}\|y\|$ depending on y. If $\mathrm{dom}(H^*)$ is unbounded, which is usually the case in constrained convex optimization, then $\|y\|$ is not bounded. In Theorems 3 and 4, we have to use adaptive techniques to update L_k in order to overcome this challenge.

Finally, we prove the primal-dual convergence rates on (P) and (D) (see Appendix C.5).

COROLLARY 5. *Under the conditions and configuration of either Theorem* 2, 3, *or* 4, *the following statements hold. If* H *in* (P) *is* M_H*-Lipschitz continuous, then*

$$\mathcal{P}(\bar{x}^k) - \mathcal{P}^\star \leq \frac{1}{2S_k}\big[\rho_0(L_0 - \mu_f)\|x^0 - x^\star\|^2 + 2\big(\|y^0\| + M_H\big)^2\big]. \tag{28}$$

Alternatively, if F^* in (D) *is M_{F^*}-Lipschitz continuous, then*

$$\mathcal{D}^\star - \mathcal{D}(\bar{y}^k) \le \frac{1}{2S_k}\big[\rho_0(L_0-\mu_f)(\|x^0\|+M_{F^*})^2 + 2\|y^0-y^\star\|^2\big]. \tag{29}$$

Here, $S_k := \rho_0(k+1)$ if $\mu_F = 0$, and $S_k := (k+1)\big[\rho_0 + \frac{\mu_F}{16L_{21}^2}(k+2)\big]$ if $\mu_F > 0$ in Theorem 2, *$S_k := \frac{2(k+1)}{\bar{L}}$ in Theorem* 3, *and $S_k := \frac{P_0(k-\bar{k}_0)^2}{4(\bar{k}_0+1)^2}$ in Theorem* 4.

5. Semi-ergodic convergence rates. In this section, we analyze semi-ergodic convergence rates (i.e., the rates on the primal last-iterate sequence and the dual averaging sequence) of Algorithm 1 for two different variants detailed below.

5.1. The convex-linear case without dual update. We first consider a variant of Algorithm 1 with $\eta_k = 0$ (i.e., $\hat{y}^k = y^0$ for all $k \ge 0$ in (15)) for the convex-linear setting $\Phi(x,y) = \langle g(x), y\rangle$. Clearly, the main steps, steps 6 to 11, of Algorithm 1 reduce to

$$\begin{cases} u^{k+1} := \operatorname{prox}_{\rho_k H^*}\left(y^0 + \rho_k g(\hat{x}^k)\right), \\ x^{k+1} := \operatorname{prox}_{h/L_k}\big(\hat{x}^k - \frac{1}{L_k}[\nabla f(\hat{x}^k) + \nabla_x \Phi(\hat{x}^k, u^{k+1})]\big), \\ \hat{x}^{k+1} := x^{k+1} + \beta_{k+1}(x^{k+1} - x^k). \end{cases} \tag{30}$$

The variant (30) has a similar form to that of standard primal-dual methods in the bilinear case such as [5, 10, 11, 13, 51]. However, the first line is different from those due to fixed y^0.

The following theorem states the convergence of (30) in both convex-linear and strongly convex-linear cases, whose proof is deferred to Appendix D.2.

THEOREM 6. *Suppose that Assumptions* 2.1, 2.2, *and* 2.3 *hold for* (SP). *Let us fix $\dot{x} \in \operatorname{dom}(F)$ and $\dot{y} \in \partial H^*(g(\dot{x}))$. Assume that either H is M_H-Lipschitz continuous or $\|g(x) - g(\dot{x})\| \le B_g$ for all $x \in \operatorname{dom}(F) \cap \operatorname{dom}(g)$ such that $L_{11}B_g < +\infty$. Let $\{(x^k, \tilde{y}^k)\}$ be generated by the variant* (30) *of Algorithm* 1 *and* (16) *with $y^0 := \dot{y}$. Let L_k be updated by*

$$L_k := \begin{cases} L_{11}M_H + L_f + \rho_k L_{21}^2 & \text{if } H \text{ is } M_H\text{-Lipschitz continuous}, \\ L_{11}\|\dot{y}\| + L_f + \rho_k(L_{21}^2 + L_{11}B_g) & \text{otherwise}. \end{cases} \tag{31}$$

Let $\tau_0 := 1$, $\rho_k := \frac{\rho_{k-1}}{1-\tau_k}$, $\eta_k := \frac{\rho_k}{2}$, and τ_k and β_k be updated as follows:

- *If $\mu_F = 0$, then update $\tau_k := \frac{1}{k+1}$ and $\beta_{k+1} := \frac{(1-\tau_k)\tau_{k+1}}{\tau_k}$, and $\rho_0 > 0$ arbitrarily.*
- *If $\mu_F > 0$, then update $\tau_{k+1} := \frac{\tau_k\left(\sqrt{\tau_k^2+4}-\tau_k\right)}{2}$ and $\beta_{k+1} := \frac{(L_k+\mu_h)\tau_{k+1}(1-\tau_k)}{(L_{k+1}-\mu_f)\tau_k}$. Moreover, we choose $\frac{\mu_F\tau_1}{L_{21}^2} \le \rho_0 \le \frac{\mu_F}{L_{21}^2}$ if H is M_H-Lipschitz continuous, and $\frac{\mu_F\tau_1}{L_{11}B_g+L_{21}^2} \le \rho_0 \le \frac{\mu_F}{L_{11}B_g+L_{21}^2}$ otherwise.*

Then, with $\mathcal{G}_{\mathcal{Z}}$ defined by (4), *for all $k \ge 0$, we have*

$$\mathcal{G}_{\mathcal{Z}}(x^k, \tilde{y}^k) \le \frac{1}{2S_k} \cdot \sup_{(x,y)\in\mathcal{Z}} \Big\{\rho_0(L_0-\mu_f)\|x-x^0\|^2 + 2\|y-\dot{y}\|^2\Big\}, \tag{32}$$

where $S_k := \rho_0(k+1)$ if $\mu_F = 0$, and $S_k := \frac{\rho_0}{4}(k+2)^2$ if $\mu_F > 0$.

Theorem 6 establishes an $\mathcal{O}(1/k)$-semi-ergodic convergence rate of the variant (30) on the gap function $\mathcal{G}_{\mathcal{Z}}$, and an $\mathcal{O}\left(1/k^2\right)$-semi-ergodic rate when either f or h is strongly convex.

5.2. The convex-linear case with dual update. For $\Phi(x,y) = \langle g(x), y\rangle$, if we update $\hat{y}^k$ as in (15), then the main steps, steps 6 to 11, of Algorithm 1 reduce to

$$\left\{\begin{array}{ll} u^{k+1} &:= \mathrm{prox}_{\rho_k H^*}\left(\hat{y}^k + \rho_k g(\hat{x}^k)\right),\\ x^{k+1} &:= \mathrm{prox}_{h/L_k}\big(\hat{x}^k - \frac{1}{L_k}[\nabla f(\hat{x}^k) + g'(\hat{x}^k)^\top u^{k+1}]\big),\\ \hat{x}^{k+1} &:= x^{k+1} + \beta_{k+1}(x^{k+1} - x^k),\\ \Theta_{k+1} &:= g(x^{k+1}) - g(\hat{x}^k) + \frac{1}{\rho_k}(u^{k+1} - \hat{y}^k),\\ \hat{y}^{k+1} &:= \hat{y}^k + \eta_k[\Theta_{k+1} - (1-\tau_k)\Theta_k]. \end{array}\right. \tag{33}$$

This variant essentially has the same per-iteration complexity as (30), but requires one additional evaluation $g(x^{k+1})$. Nevertheless, due to the new dual update of $\hat{y}^k$, it is really different from existing methods. If Φ is bilinear, then it reduces to the scheme in [48]. However, if F is strongly convex, then (33) is new compared to [48] even when Φ is bilinear.

Remark 7. We note that (33) is also different from existing augmented Lagrangian-based methods, including ADMM in the following aspects. First, its primal step in x^{k+1} minimizes a surrogate of $\mathcal{L}_{\rho_k}(x, s^{k+1}, \hat{y}^k)$ by linearizing f and the augmented term ψ_ρ. This is similar to the preconditioned ADMM variant, e.g., in [5]. Second, it has Nesterov's accelerated step on $\hat{x}^k$ in the third line. Third, the dual updates $\hat{y}^k$ and $\tilde{y}^k$ are also different from existing methods in the literature. Finally, our convergence rates in Theorems 8 are achieved on the primal last-iterate x^k instead of an averaging one as in existing ADMM. These rates are also optimal under our assumptions (up to a constant factor).

Theorem 8 (see Appendix D.3 for its proof) proves semi-ergodic rates of Algorithm 1 on the primal last-iterate sequence $\{x^k\}$ and the dual averaging sequence $\{\tilde{y}^k\}$.

THEOREM 8. *Suppose that Assumptions* 2.1, 2.2, *and* 2.3 *hold for* (SP). *Let us fix* $\dot{x} \in \mathrm{dom}(F) \cap \mathrm{dom}(g)$ *and take* $\dot{y} \in \partial H^*(g(\dot{x}))$. *Let either* H *be* M_H*-Lipschitz continuous or* $\|g(x) - g(\dot{x})\| \le B_g$ *for all* $x \in \mathrm{dom}(F) \cap \mathrm{dom}(g)$ *such that* $L_{11}B_g < +\infty$. *Let* $\{(x^k, \tilde{y}^k)\}$ *be generated by the variant* (33) *of Algorithm* 1 *and* (16) *using* $y^0 := \dot{y}$ *and the update rules:*

$$\left\{\begin{array}{l} \rho_k := \frac{\rho_{k-1}}{1-\tau_k}, \quad \eta_k := \frac{\rho_k}{2} \quad (\textit{for a given } \rho_0 > 0),\\ L_k := \left\{\begin{array}{ll} L_{11}M_H + L_f + 3\rho_k L_{21}^2 & \textit{if } H \textit{ is } M_H\textit{-Lipschitz continuous},\\ L_{11}\|y^0\| + L_f + 3\rho_k(L_{11}B_g + L_{21}^2) & \textit{otherwise}. \end{array}\right. \end{array}\right. \tag{34}$$

Moreover, τ_k *and* β_{k+1} *are updated as follows:*

- *If* $\mu_F = 0$ (*i.e., F is only convex*), *then we update* $\tau_k := \frac{1}{k+1}$ *and* $\beta_{k+1} := \frac{(1-\tau_k)\tau_{k+1}}{\tau_k}$.
- *If* $\mu_F > 0$ (*i.e., F is strongly convex*), *then we update* $\tau_{k+1} := \frac{\tau_k}{2}\big(\sqrt{\tau_k^2 + 4} - \tau_k\big)$ *and* $\beta_{k+1} := \frac{(L_k + \mu_h)\tau_{k+1}(1-\tau_k)}{(L_{k+1} - \mu_f)\tau_k}$ *with* $\tau_0 := 1$. *In addition, we choose* $\frac{\mu_F\tau_1}{L_{21}^2} \le \rho_0 \le \frac{\mu_F}{L_{21}^2}$ *if* H *is* M_H*-Lipschitz continuous and* $\frac{\mu_F\tau_1}{L_{11}B_g + L_{21}^2} \le \rho_0 \le \frac{\mu_F}{L_{11}B_g + L_{21}^2}$ *otherwise.*

Then, with $\mathcal{G}_{\mathcal{Z}}$ *defined by* (4), *for all* $k \ge 1$, *the following bound holds:*

$$\mathcal{G}_{\mathcal{Z}}(x^k, \tilde{y}^k) \le \frac{1}{2S_k} \cdot \sup_{(x,y)\in\mathcal{Z}} \Big\{\rho_0(L_0 - \mu_f)\|x - x^0\|^2 + 2\|y - y^0\|^2\Big\}, \tag{35}$$

where $S_k := \rho_0 k$ *if* $\mu_F = 0$, *and* $S_k := \frac{\rho_0(k+1)^2}{4}$ *if* $\mu_F > 0$.

Remark 9. The $\mathcal{O}(1/k)$- and $\mathcal{O}(1/k^2)$-convergence rates stated in this paper are optimal (up to a constant factor) under the corresponding assumptions in the above theorems since these rates are optimal for the special bilinear case as shown in [48].

Remark 10. Similar to Corollary 5, we can derive primal-dual convergence rates for (P) and (D) of both the variants (30) and (33) based on the results of Theorems 6 and 8. However, we skip the detailed statements here to avoid repetition.

6. Application to convex cone constrained optimization. In this section, we specify Algorithm 1 and their convergence results to handle the special case (CP) of (SP). This problem is a general convex cone constrained program as in [18] and is more general than the setting studied in existing works such as [55]. By Assumption 2.3, since $\Phi(x, y) = \langle y,\ g(x)\rangle$ is convex in x for any $y \in \mathcal{K}^*$, g is $\mathcal{K}$-convex, i.e., for all $x, \hat{x} \in \mathrm{dom}(g)$ and $\lambda \in [0,1]$, it holds that $(1-\lambda)g(x) + \lambda g(\hat{x}) - g\left((1-\lambda)x + \lambda\hat{x}\right) \in \mathcal{K}$. To guarantee strong duality, we require the Slater condition on (CP): $\{x \in \mathrm{ri}\,(\mathrm{dom}(F)) : g(x) \in -\mathrm{int}(\mathcal{K})\} \neq \emptyset$.

To solve (CP), we apply Algorithm 1 and replace the update of u^{k+1} at step 6 by

$$u^{k+1} := \mathrm{proj}_{\mathcal{K}^*}\left(\hat{y}^k + \rho_k g(\hat{x}^k)\right), \tag{36}$$

where $\mathrm{proj}_{\mathcal{K}^*}$ is the projection onto the dual cone $\mathcal{K}^*$ (see also (51)). We will characterize the convergence of Algorithm 1 via the following combined primal-dual measurement:

$$\begin{aligned} \mathcal{E}(x) &:= \max\left\{c_\star |F(x) - F^\star|,\ \mathrm{dist}_{-\mathcal{K}}\left(g(x)\right)\right\}, \\ &\text{where } \mathrm{dist}_{-\mathcal{K}}\left(g(x)\right) := \inf_{s \in -\mathcal{K}} \|g(x) - s\|, \end{aligned} \tag{37}$$

and $c_\star := \max\{1, \|y^\star\|\}$.

The following theorem proves the convergence of the proposed variant of Algorithm 1 for solving (CP), whose proof can be found in Appendix E.

THEOREM 11. *Suppose that Assumptions* 2.1, 2.2, *and* 2.3 *and the Slater condition hold for* (CP). *Let* $\{x^k\}$ *be generated by the variant of Algorithm* 1 *using* (36) *for solving* (CP). *Let* $\mathcal{E}(x)$ *be defined by* (37) *and* $\Delta_0 := \frac{\rho_0 L_0}{2}\|x^0 - x^\star\|^2 + \left(\|y^0\| + \|y^\star\| + 1\right)^2$. *Then the following hold:*

(a) *Under the conditions of Theorem* 3, *we have* $\mathcal{E}(\bar{x}^k) \leq \frac{\bar{L}\Delta_0}{2(k+1)}$ *in an ergodic sense.*

(b) *Under the conditions of Theorem* 4, *we have* $\mathcal{E}(\bar{x}^k) \leq \frac{4(\bar{k}_0+1)^2\Delta_0}{P_0(k-\bar{k}_0)^2}$ *in an ergodic sense.*

Alternatively, if $\|g(x) - g(\dot{x})\| \leq B_g$ *for all* $x \in \mathrm{dom}(F) \cap \mathrm{dom}(g)$ *and* S_k *is given as in Theorem* 6 *or* 8, *then the following statements hold:*

(c) *Under the conditions of Theorem* 6, *we have* $\mathcal{E}(x^k) \leq \frac{\Delta_0}{4S_k}$ *in the last iterate* x^k.

(d) *Under the conditions of Theorem* 8, *we have* $\mathcal{E}(x^k) \leq \frac{\Delta_0}{4S_k}$ *in the last iterate* x^k.

In the last two cases (c) and (d) of Theorem 11, we do not have the Lipschitz continuity of H since $H(\cdot) = \delta_{-\mathcal{K}}(\cdot)$. Hence, we need the assumption that $\|g(x) - g(\dot{x})\| \leq B_g$ for all $x \in \mathrm{dom}(F) \cap \mathrm{dom}(g)$, when g is not affine. The convergence bounds of Theorem 11 already combine both the primal objective residual $|F(x) - F^\star|$ (scaled by a factor $c_\star := \max\{1, \|y^\star\|\}$) and the primal feasibility violation $\mathrm{dist}_{-\mathcal{K}}\left(g(x)\right)$. Moreover, their convergence rates are optimal. The statements (a) and (b) cover [55] as special cases.

7. Numerical experiments. In this section, we test and compare different variants of Algorithm 1 on two numerical examples in subsections 7.1 and 7.2, respectively. Our experiments are implemented in MATLAB R2018b, running on a laptop with 2.8 GHz Quad-Core Intel Core i7 and 16Gb RAM using Microsoft Windows.

7.1. Quadratically constrained quadratic programming. To test our algorithms on an unbounded dual domain problem, we consider the following quadratically constrained quadratic program (QCQP):

$$\min_{x\in\mathbb{R}^p}\Big\{f(x) := \tfrac{1}{2}x^\top A_0 x + b_0^\top x \;\; \text{s.t.} \;\; \|x\| \le D, \\ \tfrac{1}{2}x^\top A_i x + b_i^\top x + c_i \le 0,\; i = 1,\dots,n\Big\}, \tag{38}$$

where $A_i \in \mathbb{R}^{p\times p}$ is symmetric and positive semidefinite, $b_i \in \mathbb{R}^p$, and $c_i \in \mathbb{R}$ for all i. In addition, A_0 is positive definite, and $D := 10$. This problem is a special case of (CP), and hence of (SP), where $h := \delta_{\mathcal{X}}$ is the indicator function of $\mathcal{X} := \{x \in \mathbb{R}^p : \|x\| \le D\}$, $H^*(y)$ is the indicator of $\mathbb{R}^n_+$, and $\Phi(x,y) := \sum_{i=1}^n y_i g_i(x)$ for $g_i(x) := \frac{1}{2}x^\top A_i x + b_i^\top x + c_i$.

We first denote $M_i := \|A_i\|D + \|b_i\|$ and $N_i := \frac{1}{2}\|A_i\|D^2 + \|b_i\|D + |c_i|$. Then Assumption 2.3 is satisfied with $L_{11} := \sqrt{n}\max\{\|A_i\| : 1 \le i \le n\}$ and $L_{21} := \|M\|$. Moreover, for a given $\dot{x}$ with $g(\dot{x}) = 0$, we have $\|g(x) - g(\dot{x})\| \le B_g := \|N\|$ for all $x \in \mathcal{X}$.

In this experiment, we solve (38) using six variants of Algorithm 1: `Alg.1(v1)` (Theorem 3), `Alg.1(v2)` (Theorem 4), `Alg.1(v3)` (Theorem 6 for the merely convex case), `Alg.1(v3s)` (Theorem 6 for the strongly convex case), `Alg.1(v4)` (Theorem 8 for the merely convex case), and `Alg.1(v4s)` (Theorem 8 for the strongly convex case). We compare our schemes with the Accelerated Primal-Dual (`APD`) algorithm (the strongly convex variant) in [18] and the `Mirror Descent` method in [33]. Note that `Mirror Descent` is double-loop where the inner loop approximately computes the prox-mapping. The input data is generated randomly using the standard Gaussian distribution in MATLAB to make sure that (38) is feasible. We generate five test cases, where p varies from 100 to 5000 variables, and n is from 10 to 100 constraints as shown in Table 1.

To obtain a fair comparison, we tune the hyperparameters of these algorithms. More specifically, for `APD` and `Mirror-Descent`, we tune their primal-dual step-sizes in the range of $[1\times 10^{-6}, 1]$. For our algorithms, we only tune ρ_0 in the same range, while updating other parameters based on our theoretical results in Theorems 3, 4, 6, and 8, respectively. We use both the relative primal-dual (or duality) gap $(\mathcal{P}(x^k) - \mathcal{D}(y^k))/|\mathcal{P}^\star|$ and the CPU time (in seconds) to measure algorithm's performance.

Our numerical results on the five tests are summarized in Table 1 after 10^5 iterations. Note that, from the theoretical convergence guarantees, the gap is computed based on averaging sequences, i.e., $\mathcal{P}(\bar{x}^k) - \mathcal{D}(\bar{y}^k)$ for `Alg.1(v1)`, `Alg.1(v2)`, `APD`, and `Mirror-Descent`. For `Alg.1(v3)`, `Alg.1(v3s)`, `Alg.1(v4)`, and `Alg.1(v4s)`, it is computed based on the primal last-iterate and the dual averaging sequence, i.e., $\mathcal{P}(x^k) - \mathcal{D}(\bar{y}^k)$.

We observe from Table 1 that `Alg.1(v4s)` outperforms other algorithms in terms of primal-dual gap in many cases. `Alg.1(v2)` is slightly better than `APD` when the strong convexity is exploited. The CPU time of our variants is comparable with `APD` and `Mirror-Descent`. Since `Alg.1(v1)`, `Alg.1(v3)`, and `Alg.1(v4)` do not utilize the strong convexity of f in (38), their performance is worse than the strongly convex variants: `Alg.1(v2)`, `(v3s)`, `(v4s)`, `APD`, and `Mirror-descent`. For the merely convex

TABLE 1
The numerical results of the eight algorithms on the five tests of the QCQP (38) *after* 10^5 *iterations.*

Prob. Size (p,n)	$(100,10)$		$(1000,10)$		$(1000,100)$		$(5000,10)$		$(5000,100)$	
Algorithms	Rel. gap	Time[s]	Rel. gap	Time[s]	Rel. gap	Time[s]	Rel. gap	Time[s]	Rel. gap	Time[s]
`Alg.1(v1)`	3.9e-4	2.8e0	1.1e-2	2.7e1	7.2e-3	2.8e2	2.6e-2	1.0e2	2.6e-2	1.0e3
`Alg.1(v2)`	5.3e-6	2.8e0	1.8e-5	2.8e1	1.2e-4	2.8e2	4.1e-6	1.0e2	8.4e-6	1.0e3
`Alg.1(v3)`	4.5e-4	2.6e0	4.5e-5	2.7e1	9.0e-3	2.6e2	2.7e-4	1.2e2	1.7e-4	9.9e2
`Alg.1(v3s)`	5.2e-6	2.7e0	1.6e-6	2.8e1	5.1e-6	2.7e2	3.2e-6	1.1e2	5.5e-6	1.0e3
`Alg.1(v4)`	2.9e-4	2.7e0	3.0e-5	2.7e1	6.2e-3	2.8e2	1.8e-5	1.0e2	1.2e-4	1.0e3
`Alg.1(v4s)`	3.4e-6	2.7e0	1.2e-6	2.8e1	9.0e-6	2.8e2	1.6e-6	1.0e2	3.6e-6	1.0e3
`APD`	2.3e-7	2.7e0	6.6e-5	2.8e1	3.8e-5	2.8e2	2.9e-4	1.0e2	1.6e-4	1.0e3
`Mirror-Descent`	4.7e-4	5.3e0	2.2e-3	4.0e1	1.1e-3	5.5e2	4.6e-3	2.1e2	4.5e-3	1.4e3

case, the semi-ergodic variants `Alg.1(v3)` and `Alg.1(v4)` are still better than the ergodic one, `Alg.1(v1)`, in most cases.

7.2. Convex-concave minimax game. We consider a convex-concave minimax game between two players, where Player 1 chooses her strategy $x \in \Delta_p := \{x \in \mathbb{R}^p_+ : \sum_{j=1}^p x_j = 1\}$ to minimize a cost function $F(x)$, and simultaneously Player 2 chooses her strategy $y \in \Delta_n := \{y \in \mathbb{R}^n_+ : \sum_{i=1}^n y_i = 1\}$ to minimize a cost function H^*. In addition, Player 1 has to pay $\Phi(x,y)$ loss to Player 2. By concrete choices of Φ as in [7, section 4.3], we can model this problem into the following minimax problem with convex-concave term:

$$\min_{x\in\Delta_p}\max_{y\in\Delta_n}\left\{\widetilde{\mathcal{L}}(x,y) := \frac{1}{N}\sum_{j=1}^{N}\log(1+\exp(a_j^\top x)) + \langle g(x), l(y)\rangle\right\}. \tag{39}$$

Clearly, (39) can be cast into our model (SP) with $f(x) := \frac{1}{N}\sum_{j=1}^N \log(1+e^{a_j^\top x})$, $h(x)$ and $H^*(y)$ being the indicator functions of Δ_p and Δ_n, and $\widetilde{\Phi}(x,y) := \langle g(x), l(y)\rangle$. By choosing different $g(x)$ and $l(y)$, (39) can cover both convex-concave and convex-linear cases.

Convex-linear case. Let $g_i(x) := \frac{b_i}{1+x_i}$ and $l_i(y) := y_i$ for $i = 1,\dots,n$. Then (39) becomes a convex-linear problem. It is easy to check that Assumption 2.3 is satisfied with $L_{11} := 2\|b\|_\infty$ and $L_{21} := \|b\|_\infty$. Since f in (39) is not strongly convex, we solve (39) using two variants of Algorithm 1: `Alg.1(v1)` and `Alg.1(v3)`, both with $\mathcal{O}(1/k)$ convergence rates on the primal-dual gap. We compare `Alg.1(v1)` and `Alg.1(v3)` with `APD` and `Mirror-Descent`. The hyperparameters of these algorithms are tuned as in subsection 7.1.

Convex-concave case. Let $g_i(x) := \frac{b_i}{1+x_i}$ and $l_i(y) := -\frac{1}{2}y_i^2$ for $i = 1,\dots,n$. Then (39) becomes a general convex-concave problem. It is easy to check that Assumption 2.4 is satisfied with $L_{11} := L_{21} := L_{22} := \|b\|_\infty$. Since this is a general convex-concave problem, we solve (39) using the variant of Algorithm 1 in Theorem 2 named `Alg.1(v5)` with $\mathcal{O}(1/k)$ convergence rate on the primal-dual gap. We compare `Alg.1(v5)` with `APD` and `Mirror-Descent`. The hyperparameters of these algorithms are tuned as in subsection 7.1.

For input data, we take the `real-sim` dataset from `LIBSVM` [6] to form vector a_i and generate b_i by using a standard uniform distribution. To fully test the performance of five algorithms, we generate 30 problem instances by randomly splitting the `real-sim` dataset into 30 equal blocks ($N = 2411$ samples per block), respectively. The performance of the five algorithms on 30 problem instances is depicted in Figure 1. Here, for `Alg.1(v1)`, `Alg.1(v5)`, `APD`, and `Mirror-Descent`, the primal-dual gap is

computed based on the primal and dual averaging sequences, i.e., $\mathcal{P}(\bar{x}^k) - \mathcal{D}(\bar{y}^k)$, while for `Alg.1(v3)`, this gap is computed through the primal last-iterate and the dual averaging sequence, i.e., $\mathcal{P}(x^k) - \mathcal{D}(\bar{y}^k)$. Next, we compute the statistic mean over all 30 instances and highlight it in a thick curve of Figure 1, while the deviation range of the duality gap is plotted in a shaded area.

From Figure 1 (left), we observe that `Alg.1(v3)` converges faster than `Alg.1(v1)`, `APD`, and `Mirror-Descent`. However, it exhibits the most oscillation behavior as shown through both the mean curve and the shaded deviation area. In fact, this is normal behavior since it uses the last-iterate sequence, which does not have a monotone decrease on the duality gap, and thus is less smooth than other curves, which use an averaging sequence. Alternatively, Figure 1 (right) shows the performance of `Alg.1(v5)` (stated in Theorem 2) and its competitors. Clearly, `Alg.1(v5)` outperforms both `APD` and `Mirror-Descent`.

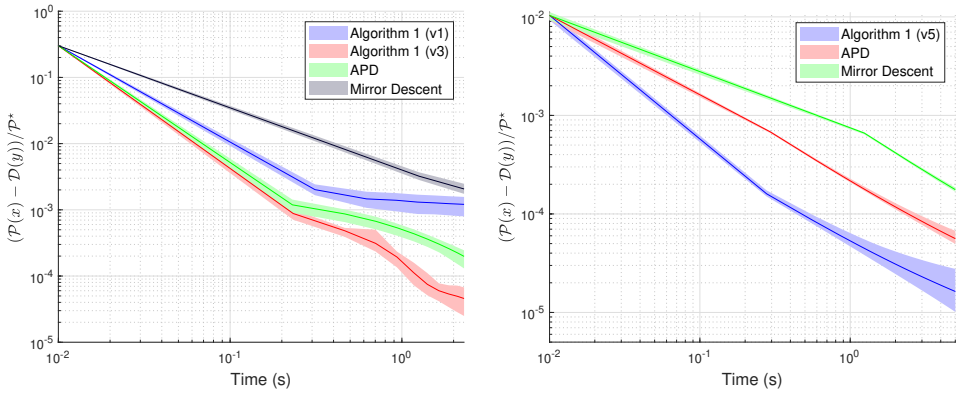

FIG. 1. *The average performance of different algorithms on* 30 *problem instances of* (39) *using the* `real-sim` *dataset. Left: The convex-linear case. Right: The convex-concave case.*

Appendix A. Proof of Lemma 1: Properties of ψ_ρ. Since $\xi(x,s,y,u) := \Phi(x,u) - \langle s,u\rangle - \frac{1}{2\rho}\|u-y\|^2$ is linear in s and convex in x, $\psi_\rho(x,s,y) = \max_u \xi(x,s,y,u)$ is convex in both x and s for any $y \in \mathrm{dom}(H^*)$ if $u^*_\rho(x,s,y) \in \mathrm{dom}(H^*)$. Moreover, since $\xi(x,s,y,u)$ is jointly strongly concave in (y,u), $\psi_\rho(x,s,y) = \max_u \xi(x,s,y,u)$ is concave in y for any $(x,s) \in \mathrm{dom}(F)\times\mathrm{dom}(H)$. The proof of (11) is similar to [46, Lemma A.1(b)].

Now, from the optimality condition of (9), we have $\nabla_y\Phi(x,u^*_\rho) = \frac{1}{\rho}(u^*_\rho - y) + s$ and $\nabla_y\Phi(\hat{x},\hat{u}^*_\rho) = \frac{1}{\rho}(\hat{u}^*_\rho - \hat{y}) + \hat{s}$, where we abbreviate $u^*_\rho := u^*_\rho(x,s,y)$ and $\hat{u}^*_\rho := u^*_\rho(\hat{x},\hat{s},\hat{y})$, respectively. Using μ_y-concavity of $\Phi(x,\cdot)$, we can easily derive that

$$\begin{aligned}\tfrac{1}{\rho}\|\hat{u}^*_\rho - u^*_\rho\|^2 \,+\, \langle \hat{s}-s, \hat{u}^*_\rho - u^*_\rho\rangle - \tfrac{1}{\rho}\langle \hat{y}-y, \hat{u}^*_\rho - u^*_\rho\rangle &= \langle \nabla_y\Phi(\hat{x},\hat{u}^*_\rho) - \nabla_y\Phi(x,u^*_\rho), \hat{u}^*_\rho - u^*_\rho\rangle\\ &\le -\mu_y\|\hat{u}^*_\rho - u^*_\rho\|^2 + \langle \nabla_y\Phi(\hat{x},u^*_\rho) - \nabla_y\Phi(x,u^*_\rho), \hat{u}^*_\rho - u^*_\rho\rangle.\end{aligned}$$

Furthermore, by the Cauchy–Schwarz inequality and the L_{21}-smoothness of $\Phi(\cdot,u^*_\rho)$, we can derive from the last expression that

$$\begin{aligned}\big(\tfrac{1}{\rho} + \mu_y\big)\|\hat{u}^*_\rho - u^*_\rho\|^2 &\le \langle \nabla_y\Phi(\hat{x},u^*_\rho) - \nabla_y\Phi(x,u^*_\rho), \hat{u}^*_\rho - u^*_\rho\rangle - \langle \hat{s}-s - \tfrac{1}{\rho}(\hat{y}-y), \hat{u}^*_\rho - u^*_\rho\rangle\\ &\le \big[L_{21}\|\hat{x}-x\| + \|\hat{s}-s\| + \tfrac{1}{\rho}\|\hat{y}-y\|\big]\|\hat{u}^*_\rho - u^*_\rho\|,\end{aligned}$$

which proves the first estimate of (12).

Next, let us redefine $\hat{u}^*_\rho := u^*_\rho(\hat{x},\hat{s},y)$ and use a shorthand $\Delta := \Delta(\hat{x},\hat{s};x,s,y)$. Then, by the optimality condition of (9), we have $\hat{s} + \frac{1}{\rho}(\hat{u}^*_\rho - y) = \nabla_y\Phi(\hat{x},\hat{u}^*_\rho)$. Using this expression and $\psi_\rho(x,s,y) = \Phi(x,u^*_\rho) - \langle s,u^*_\rho\rangle - \frac{1}{2\rho}\|u^*_\rho - y\|^2$ from (9), we can easily show that

$$\Delta = \Phi(\hat{x},\hat{u}^*_\rho) + \langle\nabla_y\Phi(\hat{x},\hat{u}^*_\rho), u^*_\rho - \hat{u}^*_\rho\rangle - \Phi(x,u^*_\rho) - \langle\nabla_x\Phi(x,u^*_\rho),\hat{x}-x\rangle + \tfrac{1}{2\rho}\|\hat{u}^*_\rho - u^*_\rho\|^2.$$

Suppose that $u^*_\rho \in \mathrm{dom}(H^*)$. Then, by μ_y-concavity of $\Phi(\hat{x},\cdot)$ and convexity of $\Phi(\cdot,u^*_\rho)$, the last expression leads to

$$\Delta \geq \Phi(\hat{x},u^*_\rho) - \Phi(x,u^*_\rho) - \langle\nabla_x\Phi(x,u^*_\rho),\hat{x}-x\rangle + \left(\tfrac{1}{2\rho} + \tfrac{\mu_y}{2}\right)\|\hat{u}^*_\rho - u^*_\rho\|^2 \geq \tfrac{(1+\rho\mu_y)}{2\rho}\|\hat{u}^*_\rho - u^*_\rho\|^2.$$

Alternatively, by L_{22}-smoothness of $\Phi(\hat{x},\cdot)$ and $\mathbf{L}_{11}$-smoothness of $\Phi(\cdot,u^*_\rho)$, we also have

$$\begin{aligned}\Delta &\leq \Phi(\hat{x},u^*_\rho) - \Phi(x,u^*_\rho) - \langle\nabla_x\Phi(x,u^*_\rho),\hat{x}-x\rangle + \left(\tfrac{1}{2\rho} + \tfrac{L_{22}}{2}\right)\|\hat{u}^*_\rho - u^*_\rho\|^2\\ &\leq \tfrac{\mathbf{L}_{11}}{2}\|\hat{x}-x\|^2 + \tfrac{(1+\rho L_{22})}{2\rho}\|\hat{u}^*_\rho - u^*_\rho\|^2.\end{aligned}$$

Combining both inequalities obtained above, we get the second estimate of (12). ∎

Appendix B. Key estimates for convergence analysis. This appendix provides different general bounds for our convergence analysis in what follows.

LEMMA 12. *Suppose that Assumptions* 2.1, 2.2, *and either* 2.3 *or* 2.4 *hold. Let* u^{k+1}, s^{k+1}, *and* x^{k+1} *be updated by* (13) *and* (14), *respectively. Let* $\mathcal{L}$, $\mathcal{L}_\rho$, *and* ψ_ρ *be given by* (5), (8), *and* (9), *respectively. Then, for any* $(x,s) \in \mathrm{dom}(F)\times\mathrm{dom}(H)$, *we have*

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},s^{k+1},\hat{y}^k) \leq\ & F(x) + H(s) + \hat{\ell}_{\rho_k}(x,s;\hat{y}^k) + L_k\langle x^{k+1} - \hat{x}^k,\ x - x^{k+1}\rangle\\ &+ \tfrac{1}{2}\big(\mathbf{L}^k_{11} + L_f + \tfrac{\rho_k(1+\rho_k L_{22})L^2_{21}}{(1+\rho_k\mu_y)^2}\big)\|x^{k+1} - \hat{x}^k\|^2\\ &- \tfrac{\mu_H}{2}\|s^{k+1} - s\|^2 - \tfrac{\mu_h}{2}\|x^{k+1} - x\|^2 - \tfrac{\mu_f}{2}\|\hat{x}^k - x\|^2,\end{aligned} \tag{40}$$

where $\hat{\ell}_{\rho_k}(x,s;\hat{y}^k) := \psi_{\rho_k}(\hat{x}^k,s^{k+1},\hat{y}^k) + \langle\nabla_x\psi_{\rho_k}(\hat{x}^k,s^{k+1},\hat{y}^k), x - \hat{x}^k\rangle + \langle\nabla_s\psi_{\rho_k}(\hat{x}^k, s^{k+1},\hat{y}^k), s - s^{k+1}\rangle$, *and* $\mathbf{L}^k_{11} := L_{11}\|u^{k+1}\|$ *under Assumption* 2.3 *and* $\mathbf{L}^k_{11} := L_{11}$ *under Assumption* 2.4.

Proof. First, the optimality condition of (14) can be written as

$$\begin{aligned}\nabla h(x^{k+1}) + \nabla f(\hat{x}^k) + \nabla_x\psi_{\rho_k}(\hat{x}^k,s^{k+1},\hat{y}^k) + L_k(x^{k+1} - \hat{x}^k) = 0,\\ \nabla h(x^{k+1}) \in \partial h(x^{k+1}).\end{aligned} \tag{41}$$

Next, by convexity of h and f, and L_f-smoothness of f, for any $x \in \mathrm{dom}(F)$, we have

$$\begin{cases} h(x^{k+1}) \leq h(x) + \langle\nabla h(x^{k+1}),\ x^{k+1} - x\rangle - \frac{\mu_h}{2}\|x^{k+1} - x\|^2,\\ f(x^{k+1}) \leq f(x) + \langle\nabla f(\hat{x}^k),\ x^{k+1} - x\rangle + \frac{L_f}{2}\|x^{k+1} - \hat{x}^k\|^2 - \frac{\mu_f}{2}\|\hat{x}^k - x\|^2.\end{cases}$$

Combining these two inequalities and then using (41) and $F := f + h$, we can derive

$$\begin{aligned}F(x^{k+1}) \overset{(41)}{\leq}\ & F(x) - \langle\nabla_x\psi_{\rho_k}(\hat{x}^k,s^{k+1},\hat{y}^k),\ x^{k+1} - x\rangle + \tfrac{L_f}{2}\|x^{k+1} - \hat{x}^k\|^2\\ &+ L_k\langle x^{k+1} - \hat{x}^k,\ x - x^{k+1}\rangle - \tfrac{\mu_h}{2}\|x^{k+1} - x\|^2 - \tfrac{\mu_f}{2}\|\hat{x}^k - x\|^2.\end{aligned} \tag{42}$$

Similarly, using the optimality condition $-\nabla_s\psi_{\rho_k}(\hat{x}^k, s^{k+1}, \hat{y}^k) \in \partial H(s^{k+1})$ of (13) and the μ_H-convexity of H, we have

$$H(s^{k+1}) \le H(s) + \langle \nabla_s\psi_{\rho_k}(\hat{x}^k, s^{k+1}, \hat{y}^k),\ s - s^{k+1}\rangle - \tfrac{\mu_H}{2}\|s^{k+1} - s\|^2. \tag{43}$$

In addition, from $u^{k+1} \in \partial H(s^{k+1})$ (or equivalently $s^{k+1} \in \partial H^*(u^{k+1})$), we have $u^{k+1} \in \mathrm{dom}(H^*)$. Using (12) in Lemma 1, we get

$$\begin{aligned}\psi_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) \le\ & \psi_{\rho_k}(\hat{x}^k, s^{k+1}, \hat{y}^k) + \langle \nabla_x\psi_{\rho_k}(\hat{x}^k, s^{k+1}, \hat{y}^k), x^{k+1} - \hat{x}^k\rangle \\ &+ \tfrac{\mathbf{L}^k_{11}}{2}\|x^{k+1} - \hat{x}^k\|^2 + \tfrac{(1+\rho_k L_{22})}{2\rho_k}\|u^*_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) - u^{k+1}\|^2.\end{aligned} \tag{44}$$

By (12) again, we have $\|u^*_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) - u^{k+1}\| \le \frac{\rho_k L_{21}}{1+\rho_k\mu_y}\|x^{k+1} - \hat{x}^k\|$. Now, combining (8), (42), (43), (44), and the last inequality, and using $\hat{\ell}_{\rho_k}(\cdot; \hat{y}^k)$, we can easily derive

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) \le\ & F(x) + H(s) + \hat{\ell}_{\rho_k}(x, s; \hat{y}^k) + L_k\langle x^{k+1} - \hat{x}^k,\ x - x^{k+1}\rangle \\ &+ \tfrac{1}{2}\left(\mathbf{L}^k_{11} + L_f + \tfrac{\rho_k(1+\rho_k L_{22})L^2_{21}}{(1+\rho_k\mu_y)^2}\right)\|x^{k+1} - \hat{x}^k\|^2 \\ &- \tfrac{\mu_H}{2}\|s^{k+1} - s\|^2 - \tfrac{\mu_h}{2}\|x^{k+1} - x\|^2 - \tfrac{\mu_f}{2}\|\hat{x}^k - x\|^2,\end{aligned}$$

which proves (40). ∎

Next, using Lemma 12, we can prove the following estimate for accelerated methods.

LEMMA 13. *Suppose that Assumptions* 2.1, 2.2, *and either* 2.3 *or* 2.4 *hold. Let* s^{k+1}, x^{k+1}, *and* $\tilde{y}^{k+1}$ *be computed by* (13), (14), *and* (16), *respectively. Let* $\mathcal{L}$ *and* $\mathcal{L}_\rho$ *be given by* (5) *and* (8), *respectively. Then, for any* $(x, s) \in \mathrm{dom}(F) \times \mathrm{dom}(H)$, *we have*

$$\begin{aligned}&\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) - \mathcal{L}(x, s, \tilde{y}^{k+1}) \le (1-\tau_k)\big[\mathcal{L}_{\rho_{k-1}}(x^k, s^k, \hat{y}^k) - \mathcal{L}(x, s, \tilde{y}^k)\big] \\ &+ \tfrac{\tau_k^2(L_k - \mu_f)}{2}\|\tfrac{1}{\tau_k}[\hat{x}^k - (1-\tau_k)x^k] - x\|^2 - \tfrac{\tau_k^2(L_k+\mu_h)}{2}\|\tfrac{1}{\tau_k}[x^{k+1} - (1-\tau_k)x^k] - x\|^2 \\ &- \tfrac{(L_k - \mathbf{L}^k_{11} - L_f - \rho_k(1+\rho_k L_{22})L^2_{21})}{2}\|x^{k+1} - \hat{x}^k\|^2 - \tfrac{1}{2\rho_k}\|(u^{k+1} - \hat{y}^k) - (1-\tau_k)(\bar{u}^k - \hat{y}^k)\|^2 \\ &- \tfrac{(1-\tau_k)(\tau_k - \Delta\rho_k)}{2\rho_k}\|\bar{u}^k - \hat{y}^k\|^2 - \tfrac{\mu_F\tau_k(1-\tau_k)}{2}\|x^k - x\|^2,\end{aligned} \tag{45}$$

where $\bar{u}^k := u^*_{\rho_k}(x^k, s^k, \hat{y}^k)$; $u^{k+1} := u^*_{\rho_k}(\hat{x}^k, s^{k+1}, \hat{y}^k)$; $\mu_F := \mu_f + \mu_h$; $\mathbf{L}^k_{11}$ *is given in Lemma* 12; *and* $\Delta\rho_k := \frac{\rho_k - \rho_{k-1}}{\rho_k}$ *if* $\Phi(x, y) := \langle g(x), y\rangle$, *and* $\Delta\rho_k := \frac{\rho_k - \rho_{k-1}}{\rho_{k-1}}$ *otherwise.*

Proof. Without loss of generality, assume that $\mu_y = 0$. Let $\mathbf{L}^k_{11}$, $\bar{u}^k$, and u^{k+1} be as given in Lemma 13, and let $B_k := \mathbf{L}^k_{11} + L_f + \rho_k(1 + \rho_k L_{22})L^2_{21}$. First, we have

$$\begin{cases}(1-\tau_k)\|x^{k+1} - x^k\|^2 + \tau_k\|x^{k+1} - x\|^2 = \|x^{k+1} - (1-\tau_k)x^k - \tau_k x\|^2 + \tau_k(1-\tau_k)\|x^k - x\|^2, \\ (1-\tau_k)\|\hat{x}^k - x^k\|^2 + \tau_k\|\hat{x}^k - x\|^2 = \|\hat{x}^k - (1-\tau_k)x^k - \tau_k x\|^2 + \tau_k(1-\tau_k)\|x^k - x\|^2.\end{cases}$$

Plugging $(x, s) := (x^k, s^k)$ into (40) of Lemma 12, and using (12) once again, we obtain

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},\ s^{k+1}, \hat{y}^k) \le\ & \mathcal{L}_{\rho_k}(x^k, s^k, \hat{y}^k) + L_k\langle x^{k+1} - \hat{x}^k,\ x^k - x^{k+1}\rangle \\ &+ \tfrac{B_k}{2}\|x^{k+1} - \hat{x}^k\|^2 - \tfrac{\mu_f}{2}\|\hat{x}^k - x^k\|^2 - \tfrac{\mu_h}{2}\|x^{k+1} - x^k\|^2 - \tfrac{1}{2\rho_k}\|u^{k+1} - \bar{u}^k\|^2.\end{aligned}$$

Next, multiplying the above estimate by $1-\tau_k \in [0,1)$, and (40) by $\tau_k \in (0,1]$, and then summing up the results, and using the first two elementary expressions, we get

$$\begin{aligned}
&\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k)\\
&\leq (1-\tau_k)\mathcal{L}_{\rho_k}(x^k, s^k, \hat{y}^k) + \tau_k\big[F(x) + H(s) + \hat{\ell}_{\rho_k}(x, s; \hat{y}^k)\big]\\
&\quad + L_k\langle x^{k+1} - \hat{x}^k, (1-\tau_k)x^k + \tau_k x - x^{k+1}\rangle - \tfrac{(1-\tau_k)}{2\rho_k}\|u^{k+1} - \bar{u}^k\|^2\\
&\quad - \tfrac{\mu_h\tau_k^2}{2}\|\tfrac{1}{\tau_k}[x^{k+1} - (1-\tau_k)x^k] - x\|^2 + \tfrac{B_k}{2}\|x^{k+1} - \hat{x}^k\|^2\\
&\quad - \tfrac{(\mu_f+\mu_h)\tau_k(1-\tau_k)}{2}\|x^k - x\|^2 - \tfrac{\mu_f\tau_k^2}{2}\|\tfrac{1}{\tau_k}[\hat{x}^k - (1-\tau_k)x^k] - x\|^2.
\end{aligned} \tag{46}$$

Using the relation $2\langle u, v\rangle = \|u+v\|^2 - \|u\|^2 - \|v\|^2$, we further have

$$\begin{aligned}
\mathcal{T}_1 &:= L_k\langle x^{k+1} - \hat{x}^k, (1-\tau_k)x^k + \tau_k x - x^{k+1}\rangle + \tfrac{B_k}{2}\|x^{k+1} - \hat{x}^k\|^2\\
&= \tfrac{L_k\tau_k^2}{2}\big[\|\tfrac{1}{\tau_k}[\hat{x}^k - (1-\tau_k)x^k] - x\|^2 - \|\tfrac{1}{\tau_k}[x^{k+1} - (1-\tau_k)x^k] - x\|^2\big]\\
&\quad - \tfrac{(L_k - B_k)}{2}\|x^{k+1} - \hat{x}^k\|^2.
\end{aligned}$$

Now, let $\Delta\rho_k := \frac{\rho_k - \rho_{k-1}}{\rho_k}$ if $\Phi(x,y) := \langle g(x), y\rangle$, and $\Delta\rho_k := \frac{\rho_k - \rho_{k-1}}{\rho_{k-1}}$ otherwise. Using (11) and (8), we can easily get

$$\begin{cases}
\mathcal{L}_{\rho_k}(x^k, s^k, \hat{y}^k) \leq \mathcal{L}_{\rho_{k-1}}(x^k, s^k, \hat{y}^k) + \frac{\Delta\rho_k}{2\rho_k}\|\bar{u}^k - \hat{y}^k\|^2,\\
F(x) + H(s) + \hat{\ell}_{\rho_k}(x, s; \hat{y}^k) \leq \mathcal{L}(x, s, u^{k+1}) - \frac{1}{2\rho_k}\|u^{k+1} - \hat{y}^k\|^2.
\end{cases} \tag{47}$$

Substituting $\mathcal{T}_1$ and (47) into (46), we can further derive

$$\begin{aligned}
&\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k)\\
&\leq (1-\tau_k)\mathcal{L}_{\rho_{k-1}}(x^k, s^k, \hat{y}^k) + \tau_k\mathcal{L}(x, s, u^{k+1}) - \tfrac{(\mu_f+\mu_h)\tau_k(1-\tau_k)}{2}\|x^k - x\|^2\\
&\quad + \tfrac{\tau_k^2}{2}(L_k - \mu_f)\|\tfrac{1}{\tau_k}[\hat{x}^k - (1-\tau_k)x^k] - x\|^2 - \tfrac{(L_k - B_k)}{2}\|x^{k+1} - \hat{x}^k\|^2\\
&\quad - \tfrac{\tau_k^2}{2}(L_k + \mu_h)\|\tfrac{1}{\tau_k}[x^{k+1} - (1-\tau_k)x^k] - x\|^2 - \tfrac{(1-\tau_k)}{2\rho_k}\|u^{k+1} - \bar{u}^k\|^2\\
&\quad - \tfrac{\tau_k}{2\rho_k}\|u^{k+1} - \hat{y}^k\|^2 + \tfrac{(1-\tau_k)\Delta\rho_k}{2\rho_k}\|\bar{u}^k - \hat{y}^k\|^2.
\end{aligned} \tag{48}$$

The last three terms of (48) can be processed as follows:

$$\begin{aligned}
\mathcal{T}_2 &:= (1-\tau_k)\|u^{k+1} - \bar{u}^k\|^2 + \tau_k\|u^{k+1} - \hat{y}^k\|^2 - (1-\tau_k)\Delta\rho_k\|\bar{u}^k - \hat{y}^k\|^2\\
&= \|(u^{k+1} - \hat{y}^k) - (1-\tau_k)(\bar{u}^k - \hat{y}^k)\|^2 + (1-\tau_k)(\tau_k - \Delta\rho_k)\|\bar{u}^k - \hat{y}^k\|^2.
\end{aligned}$$

Moreover, since $\tilde{y}^{k+1} := (1-\tau_k)\tilde{y}^k + \tau_k u^{k+1}$ due to (16), by concavity of $\mathcal{L}(x,s,\cdot)$ w.r.t. y, we have $\tau_k\mathcal{L}(x,s,u^{k+1}) \leq \mathcal{L}(x,s,\tilde{y}^{k+1}) - (1-\tau_k)\mathcal{L}(x,s,\tilde{y}^k)$. Substituting this expression and $\mathcal{T}_2$ into (48), we obtain (45). ∎

For the convenience of our analysis in what follows, we need the following additional result.

COROLLARY 14. *Under the conditions and configuration of Lemma* 12, *if* $\{(x^k, s^k, \hat{y}^k)\}$ *is updated by the variant* (17) *of Algorithm* 1 *with* $\rho_k > \eta_k > 0$, *then*

$$\begin{aligned}
&\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y)\\
&\leq \mathcal{L}(x, s, u^{k+1}) + \tfrac{(L_k - \mu_f)}{2}\|x^k - x\|^2 - \tfrac{(L_k + \mu_h)}{2}\|x^{k+1} - x\|^2\\
&\quad - \tfrac{1}{2}\Big[L_k - \mathbf{L}_{11}^k - L_f - \rho_k(1 + \rho_k L_{22})L_{21}^2 - \tfrac{\eta_k\rho_k L_{21}^2}{\rho_k - \eta_k}\Big]\|x^{k+1} - \hat{x}^k\|^2\\
&\quad + \tfrac{1}{2\eta_k}\|y - \hat{y}^k\|^2 - \tfrac{1}{2\eta_k}\|y - \hat{y}^{k+1}\|^2.
\end{aligned} \tag{49}$$

Proof. Using the second line of (47) and $\hat{x}^k = x^k$, we can derive from (40) that

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) \le \mathcal{L}(x, s, u^{k+1}) + \tfrac{(L_k-\mu_f)}{2}\|x^k - x\|^2 - \tfrac{(L_k+\mu_h)}{2}\|x^{k+1} - x\|^2 \\ - \tfrac{\left[L_k - \mathbf{L}_{11}^k - \rho_k(1+\rho_k L_{22})L_{21}^2\right]}{2}\|x^{k+1} - \hat{x}^k\|^2 - \tfrac{1}{2\rho_k}\|u^{k+1} - \hat{y}^k\|^2,\end{aligned}$$

where $u^{k+1} := u^*_{\rho_k}(\hat{x}^k, s^{k+1}, \hat{y}^k)$. By concavity of $\psi_\rho(x, s, \cdot)$ w.r.t. y and $\hat{y}^{k+1} := \hat{y}^k + \eta_k\nabla_y\psi_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) = \hat{y}^k + \frac{\eta_k}{\rho_k}(u^*_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) - \hat{y}^k)$ from (17), we have

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y) &\le \mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) + \langle\nabla_y\psi_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k), y - \hat{y}^k\rangle \\ &= \mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) + \tfrac{1}{2\eta_k}\|y - \hat{y}^k\|^2 - \tfrac{1}{2\eta_k}\|y - \hat{y}^{k+1}\|^2 \\ &\quad + \tfrac{\eta_k}{2\rho_k^2}\|u^*_\rho(x^{k+1}, s^{k+1}, \hat{y}^k) - \hat{y}^k\|^2.\end{aligned}$$

Combining the last two estimates, we obtain

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y) \le \mathcal{L}(x, s, u^{k+1}) + \tfrac{(L_k-\mu_f)}{2}\|x^k - x\|^2 - \tfrac{(L_k+\mu_h)}{2}\|x^{k+1} - x\|^2 \\ + \tfrac{1}{2\eta_k}\|y - \hat{y}^k\|^2 - \tfrac{1}{2\eta_k}\|y - \hat{y}^{k+1}\|^2 - \tfrac{(L_k - \mathbf{L}_{11}^k - \rho_k(1+\rho_k L_{22})L_{21}^2)}{2}\|x^{k+1} - \hat{x}^k\|^2 \\ - \tfrac{1}{2\rho_k}\|u^{k+1} - \hat{y}^k\|^2 + \tfrac{\eta_k}{2\rho_k^2}\|u^*_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) - \hat{y}^k\|^2.\end{aligned}$$

Using (12), $\rho_k > \eta_k > 0$, and an elementary inequality $\frac{\rho_k}{\eta_k}\|w\|^2 \le \|z\|^2 + \frac{\eta_k}{\rho_k - \eta_k}\|w - z\|^2$, we can easily show that

$$\begin{aligned}\tfrac{\eta_k}{\rho_k}\|u^*_\rho(x^{k+1}, s^{k+1}, \hat{y}^k) - \hat{y}^k\|^2 - \|u^{k+1} - \hat{y}^k\|^2 &\le \tfrac{\eta_k}{\rho_k - \eta_k}\|u^*_\rho(x^{k+1}, s^{k+1}, \hat{y}^k) - u^{k+1}\|^2 \\ &\overset{(12)}{\le} \tfrac{\eta_k\rho_k^2 L_{21}^2}{\rho_k - \eta_k}\|x^{k+1} - x^k\|^2.\end{aligned}$$

Substituting this estimate into the above inequality, we obtain (49). ∎

Appendix C. Ergodic convergence of Algorithm 1. This appendix provides the full proofs of Theorems 2, 3, and 4.

C.1. Proof of Theorem 2. First, by (19), we have $L_k - \mathbf{L}_{11}^k - L_f - \rho_k(1 + \rho_k L_{22})L_{21}^2 - \frac{\eta_k\rho_k L_{21}^2}{\rho_k - \eta_k} = 0$. Moreover, the update of ρ_k in Theorem 2 for both cases implies that $\rho_j(L_j + \mu_h) = \rho_{j+1}(L_{j+1} - \mu_f)$. Using these facts and $\rho_j = 2\eta_j$ into (49), we have

$$\begin{aligned}\eta_j\big[\mathcal{L}(x^{j+1}, s^{j+1}, y) - \mathcal{L}(x, s, u^{j+1})\big] \le \tfrac{\rho_j(L_j-\mu_f)}{4}\|x^j - x\|^2 + \tfrac{1}{2}\|\hat{y}^j - y\|^2 \\ - \big[\tfrac{\rho_{j+1}(L_{j+1}-\mu_f)}{4}\|x^{j+1} - x\|^2 + \tfrac{1}{2}\|\hat{y}^{j+1} - y\|^2\big].\end{aligned}$$

Summing up this inequality from $j = 0$ to $j = k$ and noting that $\hat{y}^0 = y^0$, we obtain

$$\textstyle\sum_{j=0}^k \eta_j\big[\mathcal{L}(x^{j+1}, s^{j+1}, y) - \mathcal{L}(x, s, u^{j+1})\big] \le \tfrac{\rho_0(L_0-\mu_f)}{4}\|x^0 - x\|^2 + \tfrac{1}{2}\|y^0 - y\|^2.$$

Dividing this estimate by $\sum_{j=0}^k \eta_j$, and using convexity of $\mathcal{L}$ w.r.t. x and s, concavity of $\mathcal{L}$ in y, $\{(\bar{x}^k, \bar{y}^k)\}$ from (18), and $\bar{s}^k := \big(\sum_{j=0}^k \eta_j\big)^{-1}\sum_{j=0}^k \eta_j s^{j+1}$, we get

$$\mathcal{L}(\bar{x}^k, \bar{s}^k, y) - \mathcal{L}(x, s, \bar{y}^k) \le \tfrac{1}{4\sum_{j=0}^k \eta_j}\left[\rho_0(L_0 - \mu_f)\|x^0 - x\|^2 + 2\|y^0 - y\|^2\right]. \tag{50}$$

Case 1: If $\mu_F = 0$, then using $\eta_j = \frac{\rho_j}{2} = \frac{\rho_0}{2}$, we have $S_k := 2\sum_{j=0}^k \eta_j = \rho_0(k+1)$.

Case 2: If $\mu_F > 0$ and $L_{22} = 0$, then applying again the update rule of ρ_k from (19), we can show that $\rho_{k+1} \ge \rho_k + \frac{\mu_F}{8L_{21}^2}$, provided that $\rho_0 \ge \frac{\mu_F + 4\nu}{16L_{21}^2}$, where

$\nu := L_{11} + L_f - \mu_f \geq 0$. By induction, we get $\rho_k \geq \rho_0 + \frac{\mu_F}{8L_{21}^2}(k+1)$. Therefore, if we define $S_k := (k+1)\big[\rho_0 + \frac{\mu_F}{16L_{21}^2}(k+2)\big]$, then we have $2\sum_{j=0}^k \eta_j = \sum_{j=0}^k \rho_j \geq \rho_0(k+1) + \frac{\mu_F}{16L_{21}^2}(k+1)(k+2) = S_k$.

Finally, by (6), we have $\widetilde{\mathcal{L}}(\bar{x}^k, y) \leq \mathcal{L}(\bar{x}^k, \bar{s}^k, y)$ and $\widetilde{\mathcal{L}}(x, \bar{y}^k) = \mathcal{L}(x, \breve{s}^k, \bar{y}^k)$ for $\breve{s}^k \in \partial H^*(\bar{y}^k)$. Hence, $\widetilde{\mathcal{L}}(\bar{x}^k, y) - \widetilde{\mathcal{L}}(x, \bar{y}^k) \leq \mathcal{L}(\bar{x}^k, \bar{s}^k, y) - \mathcal{L}(x, \breve{s}^k, \bar{y}^k)$. Substituting $s := \breve{s}^k$ and this inequality into (50) and using S_k as defined in Cases 1 and 2, we obtain $\widetilde{\mathcal{L}}(\bar{x}^k, y) - \widetilde{\mathcal{L}}(x, \bar{y}^k) \leq \frac{1}{2S_k}\big[\rho_0(L_0 - \mu_f)\|x^0 - x\|^2 + 2\|y^0 - y\|^2\big]$. Taking the supremum on both sides of this inequality over $(x, y) \in \mathcal{Z}$ and using $\mathcal{G}_{\mathcal{Z}}$ from (4), we get (20). ∎

C.2. Technical lemmas for Theorems 3 and 4. Since we consider $\Phi(x, y) = \langle g(x), y\rangle$ as a convex-linear function in Theorems 3 and 4, (13) becomes

$$\begin{aligned} u^{k+1} &:= \arg\min_u \Big\{H^*(u) - \langle g(\hat{x}^k), u\rangle + \tfrac{1}{2\rho_k}\|u - \hat{y}^k\|^2\Big\} \\ &= \operatorname{prox}_{\rho_k H^*}(\hat{y}^k + \rho_k g(\hat{x}^k)). \end{aligned} \tag{51}$$

We will need the following lemma to prove Theorem 3.

LEMMA 15. *Let $\{(x^k, \hat{y}^k)\}$ be generated by the variant (17) of Algorithm 1, where L_k, ρ_k, and η_k satisfy (22). Then we have $L_k - \mathbf{L}_{11}^k - L_f - \rho_k L_{21}^2 - \frac{\rho_k\eta_k L_{21}^2}{\rho_k - \eta_k} \geq 0$ for all $k \geq 0$. In addition, we can upper bound L_k by $0 < L_k \leq \bar{L}$, where*

$$\begin{aligned} \bar{L} \;:=\; & L_{11}[\|\dot{y}\| + \|x^0 - x^\star\| + \|y^0 - y^\star\| + \|y^\star - \dot{y}\|] + L_f + 2L_{21} \\ & + \sqrt{2L_{11}[L_{21}\|x^0 - x^\star\| + L_{21}\|y^0 - y^\star\| + \|g(x^\star) - \dot{s}\|]}. \end{aligned} \tag{52}$$

Proof. Since $\dot{y} \in \partial H(\dot{s})$ for some $\dot{s} \in \operatorname{ri}(\operatorname{dom}(H))$, we have $\dot{y} - \operatorname{prox}_{\rho_k H^*}(\dot{y} + \rho_k \dot{s}) = 0$. Since $\hat{x}^k = x^k$, using $u^{k+1} = \operatorname{prox}_{\rho_k H^*}(\hat{y}^k + \rho_k g(x^k))$ from (51), we can deduce that

$$\begin{aligned} \|u^{k+1}\| &= \|\operatorname{prox}_{\rho_k H^*}\big(\hat{y}^k + \rho_k g(x^k)\big) - \operatorname{prox}_{\rho_k H^*}(\dot{y} + \rho_k \dot{s}) + \dot{y}\| \\ &\leq \|\dot{y}\| + \|\hat{y}^k - \dot{y}\| + \rho_k\|g(x^k) - \dot{s}\|. \end{aligned} \tag{53}$$

By the update rule of L_k, η_k, and ρ_k as (22), and using (53), for $k \geq 0$, we have

$$\begin{cases} L_{11}\rho_k\|g(x^k) - \dot{s}\| \;=\; \frac{2}{L_k}L_{11}\|g(x^k) - \dot{s}\| \;\leq\; \sqrt{2L_{11}\|g(x^k) - \dot{s}\|}, \\ 2\rho_k L_{21}^2 \qquad\qquad\quad = \frac{4L_{21}^2}{L_k} \qquad\qquad\qquad\quad \leq 2L_{21}. \end{cases} \tag{54}$$

Therefore, in view of (53) and (54), we can bound

$$\begin{aligned} L_k - \mathbf{L}_{11}^k - L_f - \rho_k L_{21}^2 - \tfrac{\rho_k\eta_k L_{21}^2}{\rho_k - \eta_k} &\geq L_k - L_{11}[\|\dot{y}\| + \|\hat{y}^k - \dot{y}\|] - \sqrt{2L_{11}\|g(x^k) - \dot{s}\|} \\ &\quad - L_f - 2L_{21} \\ &= 0. \end{aligned}$$

Substituting this condition, $\mu_f = \mu_h = 0$, and $\eta_k := \frac{1}{L_k}$ into (49) of Corollary 14, we obtain for any $(x, s, y) \in \operatorname{dom}(F) \times \operatorname{dom}(H) \times \operatorname{dom}(H^*)$ that

$$\begin{aligned} \mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y) - \mathcal{L}(x, s, u^{k+1}) \;\leq\; & \tfrac{L_k}{2}\big[\|x^k - x\|^2 - \|x^{k+1} - x\|^2\big] \\ & + \tfrac{L_k}{2}\big[\|\hat{y}^k - y\|^2 - \|\hat{y}^{k+1} - y\|^2\big]. \end{aligned} \tag{55}$$

By (7), we have $\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y^\star) - \mathcal{L}(x^\star, s^\star, u^{k+1}) \geq 0$. Hence, (55) implies that

$$\|x^{k+1} - x^\star\|^2 + \|\hat{y}^{k+1} - y^\star\|^2 \leq \|x^k - x^\star\|^2 + \|\hat{y}^k - y^\star\|^2.$$

This means that $\{\|x^k - x^\star\|^2 + \|\hat{y}^k - y^\star\|^2\}$ is a nonincreasing sequence. Therefore, by induction, we get $\|\hat{y}^k - y^\star\|^2 \leq \|x^0 - x^\star\|^2 + \|y^0 - y^\star\|^2$ and $\|x^k - x^\star\|^2 \leq \|x^0 - x^\star\|^2 + \|y^0 - y^\star\|^2$.

Finally, using these bounds, $\hat{x}^k = x^k$, and $\|g(x^k) - g(x^\star)\| \le L_{21}\|x^k - x^\star\|$, we have

$$\begin{aligned} L_k &:= L_{11}[\|\dot{y}\| + \|\hat{y}^k - \dot{y}\|] + \sqrt{2L_{11}\|g(x^k) - \dot{s}\|} + L_f + 2L_{21} \\ &\le L_{11}[\|\dot{y}\| + \|x^0 - x^\star\| + \|y^0 - y^\star\| + \|y^\star - \dot{y}\|] + L_f + 2L_{21} \\ &\quad + \sqrt{2L_{11}[L_{21}\|x^0 - x^\star\| + L_{21}\|y^0 - y^\star\| + \|g(x^\star) - \dot{s}\|]} =: \bar{L}, \end{aligned}$$

which proves that $L_k \le \bar{L}$. ∎

We also need the following two additional lemmas to prove Theorem 4.

LEMMA 16. *Let η_k, ρ_k, L_k be updated by (25). We define the following constants:*

$$\begin{cases} C_1 := \sqrt{\eta_0(L_0 - \mu_f)\|x^0 - x^\star\|^2 + \|\hat{y}^0 - y^\star\|^2}, \\ C_2 := L_{11}\left(\|\dot{y}\| + C_1 + \|y^\star - \dot{y}\|\right) + \kappa, \\ \hat{L} := L_f + C_2 + L_{11}C_1 + \rho_0\left(L_{11}\|g(x^\star) - \dot{s}\| + 2L_{21}^2 + \kappa\right). \end{cases} \tag{56}$$

Then $\rho_k \ge \frac{\rho_0 L_0}{\hat{L}}$. Moreover, A_k and B_k updated by (26) are, respectively, upper bounded by

$$A_k \le C_2 \quad \text{and} \quad B_k \le \tfrac{\hat{L}L_{11}C_1}{\rho_0 L_0} + L_{11}\|g(x^\star) - \dot{s}\| + \kappa.$$

Proof. First, from (53) we have $\|u^{k+1}\| \le \|\dot{y}\| + \|\hat{y}^k - \dot{y}\| + \rho_k\|g(x^k) - \dot{s}\|$. By the update rule (26), we also have $A_k \ge L_{11}[\|\dot{y}\| + \|\hat{y}^k - \dot{y}\|]$ and $B_k \ge L_{11}\|g(x^k) - \dot{s}\|$. Combining these two statements, we can show that

$$\begin{aligned} L_k - L_{11}^k - L_f - \rho_k L_{21}^2 - \tfrac{\rho_k\eta_k L_{21}^2}{\rho_k - \eta_k} &\overset{(25)}{=} A_k + \rho_k B_k + L_f + 2\rho_k L_{21}^2 - [L_{11}\|u^{k+1}\| + L_f + 2\rho_k L_{21}^2] \\ &\ge A_k - L_{11}\left(\|\dot{y}\| + \|\hat{y}^k - \dot{y}\|\right) + \rho_k\left(B_k - L_{11}\|g(x^k) - \dot{s}\|\right) \\ &\ge 0. \end{aligned}$$

Using this condition in (49), for any $(x, s, y) \in \mathrm{dom}(F) \times \mathrm{dom}(H) \times \mathrm{dom}(H^*)$, we have

$$\begin{aligned} \mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y) - \mathcal{L}(x, s, u^{k+1}) \le\ & \tfrac{(L_k - \mu_f)}{2}\|x^k - x\|^2 - \tfrac{(L_k + \mu_h)}{2}\|x^{k+1} - x\|^2 \\ & + \tfrac{1}{2\eta_k}\|\hat{y}^k - y\|^2 - \tfrac{1}{2\eta_k}\|\hat{y}^{k+1} - y\|^2. \end{aligned}$$

Multiplying both sides of the above inequality by ρ_k and using $\rho_k(L_k + \mu_h) = \rho_{k+1}(L_{k+1} - \mu_f)$ obtained from the update rule of ρ_k in (25), we get

$$\begin{aligned} & \rho_k\left[\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y) - \mathcal{L}(x, s, u^{k+1})\right] \\ \overset{(25)}{\le}\ & \tfrac{\rho_k(L_k - \mu_f)}{2}\|x^k - x\|^2 + \|\hat{y}^k - y\|^2 \\ & - \tfrac{\rho_{k+1}(L_{k+1} - \mu_f)}{2}\|x^{k+1} - x\|^2 - \|\hat{y}^{k+1} - y\|^2. \end{aligned} \tag{57}$$

By (7), we have $\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y^\star) - \mathcal{L}(x^\star, s^\star, u^{k+1}) \ge 0$. Therefore, the last inequality implies that

$$\rho_{k+1}(L_{k+1} - \mu_f)\|x^{k+1} - x^\star\|^2 + 2\|\hat{y}^{k+1} - y^\star\|^2 \le \rho_k(L_k - \mu_f)\|x^k - x^\star\|^2 + 2\|\hat{y}^k - y^\star\|^2.$$

As a consequence, $\{\rho_k(L_k - \mu_f)\|x^k - x^\star\|^2 + 2\|\hat{y}^k - y^\star\|^2\}$ is nonincreasing, and

$$\rho_k(L_k - \mu_f)\|x^k - x^\star\|^2 + 2\|\hat{y}^k - y^\star\|^2 \le C_1^2 := \rho_0(L_0 - \mu_f)\|x^0 - x^\star\|^2 + 2\|\hat{y}^0 - y^\star\|^2.$$

Moreover, using $L_k := L_f + A_k + \rho_k(B_k + 2L_{21}^2)$ from (25), one can easily show that $\rho_k(L_k - \mu_f) \geq 2\rho_k^2 L_{21}^2$. Therefore, we have

$$\rho_k L_{21}\|x^k - x^\star\| \leq C_1 \quad \text{and} \quad \|\hat{y}^k - y^\star\| \leq C_1. \tag{58}$$

From (26), by induction, we have

$$A_k \leq \max_{1\leq j\leq k}\{L_{11}[\|\dot{y}\| + \|\hat{y}^j - \dot{y}\|]\} + \kappa \quad \text{and} \quad B_k \leq \max_{1\leq j\leq k}\{L_{11}\|g(x^j) - \dot{s}\|\} + \kappa.$$

Therefore, A_k can be upper bounded as

$$\begin{aligned} A_k &\leq \max_{1\leq j\leq k}\{L_{11}\left[\|\dot{y}\| + \|\hat{y}^j - \dot{y}\|\right]\} + \kappa \\ &\overset{(58)}{\leq} C_2 := L_{11}\left[\|\dot{y}\| + C_1 + \|y^\star - \dot{s}\|\right] + \kappa. \end{aligned} \tag{59}$$

Similarly, B_k can be upper bounded as

$$\begin{aligned} B_k &\leq \max_{1\leq j\leq k}\{L_{11}\|g(x^j) - \dot{s}\|\} + \kappa \\ &\leq \max_{1\leq j\leq k}\left\{L_{11}\left[L_{21}\|x^j - x^\star\| + \|g(x^\star) - \dot{s}\|\right]\right\} + \kappa \\ &\overset{(58)}{\leq} L_{11}C_1 \max_{1\leq j\leq k}\{\tfrac{1}{\rho_k}\} + L_{11}\|g(x^\star) - \dot{s}\| + \kappa. \end{aligned} \tag{60}$$

Now, we show that ρ_k is lower bounded. Note that ρ_k updated by (25) satisfies $\rho_{k-1}(L_{k-1}+\mu_h) = \rho_k(L_k - \mu_f)$, leading to $\rho_{k-1}L_{k-1} = \rho_k L_k - \rho_k\mu_f - \rho_{k-1}\mu_h \leq \rho_k L_k$. Hence, by induction, we have $\rho_k L_k \geq \rho_0 L_0$. Assume ρ_t is the smallest value up to the iteration k, i.e., $\rho_t = \min_{0\leq j\leq k}\{\rho_j\} \leq \rho_k$ for some $0 \leq t \leq k$. Then

$$\begin{aligned} \rho_t L_t &= \rho_t(A_t + \rho_t B_t + L_f + 2\rho_t L_{21}^2) \\ &\overset{(59)(60)}{\leq} \rho_t\left[L_f + C_2 + L_{11}C_1 + \rho_t(L_{11}\|g(x^\star) - \dot{s}\| + \kappa + 2L_{21}^2)\right] \\ &\leq \rho_t\left[L_f + C_2 + L_{11}C_1 + \rho_0(L_{11}\|g(x^\star) - \dot{s}\| + \kappa + 2L_{21}^2)\right] =: \rho_t\hat{L}. \end{aligned}$$

Consequently, we have $\rho_t \geq \frac{\rho_0 L_0}{\hat{L}}$, and hence $\rho_k \geq \frac{\rho L_0}{\hat{L}}$. Using $\rho_k \geq \frac{\rho L_0}{\hat{L}}$ in (60), we get

$$B_k \leq \tfrac{L_{11}C_1\hat{L}}{\rho_0 L_0} + L_{11}\|g(x^\star) - \dot{s}\| + \kappa,$$

which completes the proof of Lemma 16. ∎

LEMMA 17. *Let η_k, ρ_k, L_k, A_k, and B_k be updated by* (25) *and* (26)*, respectively, and $\mathcal{J}_0 := \{k \geq 0 : A_{k+1} > A_k$ or $B_{k+1} > B_k\}$. Then, for all $k \in \mathcal{J}_0$, we have*

$$|\mathcal{J}_0| < \bar{k}_0 := \left\lceil \frac{C_2}{\kappa} \right\rceil \cdot \left\lceil \frac{C_1 L_{11}\bar{L} + \rho_0 L_0[L_{11}\|g(x^\star) - \dot{s}\| + \kappa]}{\kappa\rho_0 L_0} \right\rceil. \tag{61}$$

In addition, for any $k \geq 0$ such that $k \notin \mathcal{J}_0$, we have

$$\rho_{k+1} - \rho_k \geq P_0 := \frac{\rho_0 L_0 \mu_F}{\hat{L}(C_2 + L_f + \mu_h) + 2L_{11}C_1\hat{L} + \left[2L_{11}\|g(x^\star) - \dot{s}\| + 2\kappa + 4L_{21}^2\right]\rho_0 L_0} > 0, \tag{62}$$

where C_1, C_2, and $\hat{L}$ are defined as in (56)*, and $\bar{L}$ is defined by* (52)*.*

Proof. First, since $A_k \leq C_2$ and $B_k \leq \frac{L_{11}C_1\hat{L}}{\rho_0 L_0} + L_{11}\|g(x^\star) - \dot{s}\| + \kappa$ due to Lemma 16, combining these facts and (26), we can easily show that for any $k \in \mathcal{J}_0$, it holds that $k \leq \bar{k}_0$, where $\bar{k}_0$ is given in (61).

Next, from the update rules of A_k and B_k in (26) again, for all $k \notin \mathcal{J}_0$, we have $B_{k+1} = B_k$ and $A_{k+1} = A_k$. Hence, using $\rho_k(L_k + \mu_h) = \rho_{k+1}(L_{k+1} - \mu_f)$ derived from the update rule of ρ_k in (25), we have

$$\rho_k(A_k + L_f + \mu_h) + \rho_k^2(B_k + 2L_{21}^2) = \rho_{k+1}(A_k + L_f - \mu_f) + \rho_{k+1}^2(B_k + 2L_{21}^2).$$

Using $\rho_k^2 \geq \rho_{k+1}^2 + 2\rho_{k+1}(\rho_k - \rho_{k+1})$, the above equality can be relaxed to

$$\rho_k(A_k+L_f+\mu_h)+2(B_k+2L_{21}^2)\rho_{k+1}(\rho_k-\rho_{k+1})\leq\rho_{k+1}(A_k+L_f+\mu_h)-\rho_{k+1}(\mu_h+\mu_f),$$

which is equivalent to $\rho_{k+1}-\rho_k\geq\frac{(\mu_h+\mu_f)\rho_{k+1}}{A_k+L_f+\mu_h+(2B_k+4L_{21}^2)\rho_{k+1}}$. Using $\rho_k\geq\frac{\rho_0 L_0}{\hat{L}}$, $A_k\leq C_2$, and $B_k\leq\frac{L_{11}C_1\hat{L}}{\rho_0 L_0}+L_{11}\|g(x^\star)-\dot{s}\|+\kappa$, we have

$$\rho_{k+1}-\rho_k\geq P_0:=\frac{\rho_0 L_0\mu_F}{\hat{L}(C_2+L_f+\mu_h)+2L_{11}C_1\hat{L}+\left[2L_{11}\|g(x^\star)-\dot{s}\|+2\kappa+4L_{21}^2\right]\rho_0 L_0},$$

which is exactly (62). ∎

C.3. Proof of Theorem 3. First, from (55) and the first line of (11), we have

$$\mathcal{L}(x^{j+1},s^{j+1},y)-\mathcal{L}(x,s,u^{j+1})\leq\tfrac{L_j}{2}\left[\|x^j-x\|^2-\|x^{j+1}-x\|^2+\|\hat{y}^j-y\|^2-\|\hat{y}^{j+1}-y\|^2\right].$$

Multiplying the above inequality by $2\eta_j$ and noticing that $\eta_j=\frac{1}{L_j}$, we obtain

$$2\eta_j[\mathcal{L}(x^{j+1},s^{j+1},y)-\mathcal{L}(x,s,u^{j+1})]\leq\|x^j-x\|^2-\|x^{j+1}-x\|^2+\|\hat{y}^j-y\|^2-\|\hat{y}^{j+1}-y\|^2.$$

Summing up this inequality from $j:=0$ to $j:=k$, and using $\hat{y}^0=y^0$, we get

$$\sum_{j=0}^{k}\eta_j\left[\mathcal{L}(x^{j+1},s^{j+1},y)-\mathcal{L}(x,s,u^{j+1})\right]\leq\frac{1}{2}\left[\|x^0-x\|^2+\|y^0-y\|^2\right].$$

Dividing it by $\sum_{j=0}^k\eta_j$, and using the convexity of $\mathcal{L}$ in x and s, the concavity of $\mathcal{L}$ in y, $\{(\bar{x}^k,\bar{y}^k)\}$ defined by (18), and $\bar{s}^k:=\left(\sum_{j=0}^k\eta_j\right)^{-1}\sum_{j=0}^k\eta_j s^{j+1}$, we get

$$\mathcal{L}(\bar{x}^k,\bar{s}^k,y)-\mathcal{L}(x,s,\bar{y}^k)\leq\tfrac{1}{2\sum_{j=0}^k\eta_j}\left[\|x^0-x\|^2+\|y^0-y\|^2\right].$$

Moreover, (52) implies $\sum_{j=0}^k\eta_j=\sum_{j=0}^k\frac{1}{L_j}\geq\frac{(k+1)}{\bar{L}}$. Using this and $\rho_0 L_0=2$, we have

$$\mathcal{L}(\bar{x}^k,\bar{s}^k,y)-\mathcal{L}(x,s,\bar{y}^k)\leq\frac{\bar{L}}{4(k+1)}[\rho_0 L_0\|x^0-x\|^2+2\|y^0-y\|^2]. \tag{63}$$

By (6), we have $\widetilde{\mathcal{L}}(\bar{x}^k,y)\leq\mathcal{L}(\bar{x}^k,\bar{s}^k,y)$ and $\widetilde{\mathcal{L}}(x,\bar{y}^k)=\mathcal{L}(x,\breve{s}^k,\bar{y}^k)$ for $\breve{s}^k\in\partial H^*(\bar{y}^k)$. Hence, $\widetilde{\mathcal{L}}(\bar{x}^k,y)-\widetilde{\mathcal{L}}(x,\bar{y}^k)\leq\mathcal{L}(\bar{x}^k,\bar{s}^k,y)-\mathcal{L}(x,\breve{s}^k,\bar{y}^k)$. Substituting $s:=\breve{s}^k$ and this inequality into (63), we obtain $\widetilde{\mathcal{L}}(\bar{x}^k,y)-\widetilde{\mathcal{L}}(x,\bar{y}^k)\leq\frac{\bar{L}}{4(k+1)}\left[\rho_0 L_0\|x^0-x\|^2+2\|y^0-y\|^2\right]$. Taking the supremum on both sides of this estimate over $\mathcal{Z}$ and using $\mathcal{G}_{\mathcal{Z}}$ from (4), we prove (23). ∎

C.4. Proof of Theorem 4. First, from (57) and the first line of (11), we have

$$\begin{aligned}\eta_j\left[\mathcal{L}(x^{j+1},s^{j+1},y)-\mathcal{L}(x,s,u^{j+1})\right]\;\leq\;&\tfrac{\eta_j(L_j-\mu_f)}{2}\|x^j-x\|^2-\tfrac{\eta_{j+1}(L_{j+1}-\mu_f)}{2}\|x^{j+1}-x\|^2\\&+\;\tfrac{1}{2}\|\hat{y}^j-y\|^2-\tfrac{1}{2}\|\hat{y}^{j+1}-y\|^2.\end{aligned}$$

Summing up this inequality from $j:=0$ to $j:=k$, and noting that $\rho_0=2\eta_0$, we obtain

$$2\sum_{j=0}^{k}\eta_j[\mathcal{L}(x^{j+1},s^{j+1},y)-\mathcal{L}(x,s,u^{j+1})]\leq\rho_0(L_0-\mu_f)\|x^0-x\|^2+\|y^0-y\|^2:=\mathcal{R}_0^2(x,y),$$

Dividing this estimate by $\sum_{j=0}^k \eta_j$, and using the convexity of $\mathcal{L}$ in x and s, the concavity of $\mathcal{L}$ in y, $\{(\bar{x}^k,\bar{y}^k)\}$ defined by (18), and $\bar{s}^k := \big(\sum_{j=0}^k \eta_j\big)^{-1}\sum_{j=0}^k \eta_j s^{j+1}$, we get

$$\mathcal{L}(\bar{x}^k,\bar{s}^k,y)-\mathcal{L}(x,s,\bar{y}^k) \le \tfrac{1}{\sum_{j=0}^k \eta_j}\sum_{j=0}^k \eta_j\big[\mathcal{L}(x^{j+1},s^{j+1},y)-\mathcal{L}(x,s,u^{j+1})\big] \le \tfrac{\mathcal{R}_0^2(x,y)}{2\sum_{j=0}^k \rho_j}.$$

Here, we have used $\eta_k = \frac{\rho_k}{2}$.

Now, let us lower bound $\sum_{j=0}^k \rho_j$ as follows. Suppose that we have run our algorithm for k iterations. From Lemma 17, we have $|\mathcal{J}_0| \le \bar{k}_0$. Therefore, there exists an interval $[s,t] \subseteq [0,k]$ such that $[s,t]\cap\mathcal{J}_0 = \emptyset$ and $t-s \ge \frac{k-\bar{k}_0}{\bar{k}_0+1}$. Using (62), we have $\sum_{j=0}^k \rho_j \ge \sum_{j=s}^t \rho_j \overset{(62)}{\ge} \sum_{j=s}^t P_0(j-s) \ge \frac{P_0}{2}\big(\frac{k-\bar{k}_0}{\bar{k}_0+1}\big)^2$. Hence, it follows that

$$\mathcal{L}(\bar{x}^k,\bar{s}^k,y)-\mathcal{L}(x,s,\bar{y}^k) \le \frac{2(\bar{k}_0+1)^2}{P_0(k-\bar{k}_0)^2}\cdot\mathcal{R}_0^2(x,y).$$

Using the same arguments as in the proof of Theorem 3, we can prove (27). We omit repeating this derivation here. ∎

C.5. Proof of Corollary 5. It is sufficient to prove (28) and (29) for Theorem 2. The results for Theorems 3 and 4 are proven similarly. Let S_k be defined as in Corollary 5.

Let us take $\bar{u}_*^k$ such that $\nabla_y\Phi(\bar{x}^k,\bar{u}_*^k) \in \partial H^*(\bar{u}_*^k)$. Using this fact, we can derive that

$$\begin{aligned}\mathcal{P}(\bar{x}^k)-\mathcal{P}^\star &\overset{(\mathrm{P})}{=} F(\bar{x}^k)+\max_y\big\{\Phi(\bar{x}^k,y)-H^*(y)\big\}-\mathcal{P}^\star = F(\bar{x}^k)+\Phi(\bar{x}^k,\bar{u}_*^k)-H^*(\bar{u}_*^k)-\mathcal{P}^\star\\ &\overset{(7)}{\le} F(\bar{x}^k)+\Phi(\bar{x}^k,\bar{u}_*^k)-H^*(\bar{u}_*^k)-\widetilde{\mathcal{L}}(x^\star,\bar{y}^k) = \widetilde{\mathcal{L}}(\bar{x}^k,\bar{u}_*^k)-\widetilde{\mathcal{L}}(x^\star,\bar{y}^k)\\ &\overset{(50)}{\le} \tfrac{1}{S_k}\big[\rho_0(L_0-\mu_f)\|x^0-x^\star\|^2+2\|\bar{u}_*^k-y^0\|^2\big].\end{aligned}$$

If H is M_H-Lipschitz continuous, then since $\bar{u}_*^k \in \partial H(\nabla_y\Phi(\bar{x}^k,\bar{u}_*^k))$, we have $\|\bar{u}_*^k\| \le M_H$. This condition leads to $\|y^0-\bar{u}_*^k\|^2 \le (\|y^0\|+\|\bar{u}_*^k\|)^2 = (\|y^0\|+M_H)^2$. Using this bound in the above estimate, we obtain (28).

Alternatively, let $-\nabla_x\Phi(\bar{x}_*^k,\bar{y}^k) \in \partial F(\bar{x}_*^k)$ (or $\bar{x}_*^k \in \partial F(-\nabla_x\Phi(\bar{x}_*^k,\bar{y}^k))$). Then

$$\begin{aligned}\mathcal{D}^\star-\mathcal{D}(\bar{y}^k) &\overset{(\mathrm{D})}{=} \mathcal{D}^\star+H^*(\bar{y}^k)-\min_x\big\{\Phi(x,\bar{y}^k)+F(x)\big\} = \mathcal{D}^\star+H^*(\bar{y}^k)-\Phi(\bar{x}_*^k,\bar{y}^k)-F(\bar{x}_*^k)\\ &\overset{(7)}{\le} \widetilde{\mathcal{L}}(\bar{x}^k,y^\star)-\big[F(\bar{x}_*^k)+\Phi(\bar{x}_*^k,\bar{y}^k)-H^*(\bar{y}^k)\big] = \widetilde{\mathcal{L}}(\bar{x}^k,y^\star)-\widetilde{\mathcal{L}}(\bar{x}_*^k,\bar{y}^k)\\ &\overset{(50)}{\le} \tfrac{1}{S_k}\big[\rho_0(L_0-\mu_f)\|\bar{x}_*^k-x^0\|^2+2\|y^0-y^\star\|^2\big].\end{aligned}$$

If F^* is M_{F^*}-Lipschitz continuous, then since $\bar{x}_*^k \in \partial F^*(-\nabla_x\Phi(\bar{x}_*^k,\bar{y}^k))$, we have $\|\bar{x}_*^k\| \le M_{F^*}$. This condition leads to $\|x^0-\bar{x}_*^k\|^2 \le (\|x^0\|+\|\bar{x}_*^k\|)^2 = (\|x^0\|+M_{F^*})^2$. Using this bound in the above estimate, we obtain (29). ∎

Appendix D. Semi-ergodic convergence of Algorithm 1. This appendix provides the full proof of Theorems 6 and 8.

D.1. Technical lemmas. In order to prove Theorems 6 and 8, we need the following results.

LEMMA 18. *Let $\dot{y}\in\partial H(g(\dot{x}))$ for a given $\dot{x}\in\mathrm{dom}(F)$ and either H is M_H-Lipschitz continuous or $\|g(x)-g(\dot{x})\|\le B_g$ for all $x\in\mathrm{dom}(F)\cap\mathrm{dom}(g)$. Then the following hold:*

(a) *If $u^{k+1}:=\mathrm{prox}_{\rho_k H^*}\big(y^0+\rho_k g(\hat{x}^k)\big)$ as in (30), then $\|u^{k+1}\|\le M_H$ if H is M_H-Lipschitz continuous, and $\|u^{k+1}\|\le\|\dot{y}\|+\|y^0-\dot{y}\|+\rho_k B_g$ otherwise.*

(b) *If $u^{k+1}:=\mathrm{prox}_{\rho_k H^*}\big(\hat{y}^k+\rho_k g(\hat{x}^k)\big)$ as in (33) and $y^0:=\dot{y}$, then $\|u^{k+1}\|\le M_H$ if H is M_H-Lipschitz continuous, and $\|u^{k+1}\|\le\|y^0\|+3\rho_k B_g$ otherwise.*

Proof. (a) If H is M_H-Lipschitz continuous, then $\mathrm{dom}(H^*)$ is bounded by M_H. Since $u^{k+1}\in\mathrm{dom}(H^*)$ due to (30), we get $\|u^{k+1}\|\le M_H$. Otherwise, since $\dot{y}\in\partial H(g(\dot{x}))$, we have $\dot{y}=\mathrm{prox}_{\rho_k H^*}(\dot{y}+\rho_k g(\dot{x}))$. Hence, we have $\|u^{k+1}\|=\|\mathrm{prox}_{\rho_k H^*}(\dot{y}+\rho_k g(\dot{x}))-\mathrm{prox}_{\rho_k H^*}\big(y^0+\rho_k g(\hat{x}^k)\big)-\dot{y}\|\le\|\dot{y}\|+\|y^0-\dot{y}\|+\rho_k\|g(\hat{x}^k)-g(\dot{x})\|\le\|\dot{y}\|+\|y^0-\dot{y}\|+\rho_k B_g$.

(b) From (33), we have $\hat{y}^{k+1}=\hat{y}^k+\eta_k[\Theta_{k+1}-(1-\tau_k)\Theta_k]\overset{(34)}{=}\hat{y}^k+\eta_k\Theta_{k+1}-\eta_{k-1}\Theta_k$. By induction, we obtain

$$\hat{y}^{k+1}-\eta_k\Theta_{k+1}=\hat{y}^k-\eta_{k-1}\Theta_k=\hat{y}^1-\eta_0\Theta_1=y^0-(1-\tau_0)\eta_0\Theta_0=y^0\quad\forall k\ge 0. \tag{64}$$

Next, from $g(\dot{x})\in\partial H^*(\dot{y})$, we get $g(\dot{x})=\mathrm{prox}_{H/\rho_k}(g(\dot{x})+\dot{y}/\rho_k)$. By (13), we have

$$\begin{aligned}\|\Theta_{k+1}\| &= \|g(x^{k+1})-s^{k+1}\|=\|g(x^{k+1})-\mathrm{prox}_{H/\rho_k}(g(\hat{x}^k)+\tfrac{\hat{y}^k}{\rho_k})\|\\ &\le \|g(x^{k+1})-g(\dot{x})+\mathrm{prox}_{H/\rho_k}(g(\dot{x})+\tfrac{\dot{y}}{\rho_k})-\mathrm{prox}_{H/\rho_k}(g(\hat{x}^k)+\tfrac{\hat{y}^k}{\rho_k})\|\\ &\le \|g(x^{k+1})-g(\dot{x})\|+\|g(\hat{x}^k)-g(\dot{x})\|+\tfrac{1}{\rho_k}\|\hat{y}^k-\dot{y}\|\end{aligned}$$

Combining this estimate and (64) and noting that $\eta_k=\frac{\rho_k}{2}$ and $y^0:=\dot{y}$, we obtain

$$\|\hat{y}^{k+1}-y^0\| \le \eta_k[\|g(x^{k+1})-g(\dot{x})\|+\|g(\hat{x}^k)-g(\dot{x})\|]+\tfrac{1}{2}\|\hat{y}^k-y^0\|\le\rho_k B_g+\tfrac{1}{2}\|\hat{y}^k-y^0\|.$$

Since $0<\rho_k\le\rho_{k+1}$, by induction, we can prove that $\|\hat{y}^k-y^0\|\le 2\rho_k B_g$. Using this bound, we can show that $\|u^{k+1}\|=\|\mathrm{prox}_{\rho_k H^*}(\dot{y}+\rho_k g(\dot{x}))-\mathrm{prox}_{\rho_k H^*}\big(\hat{y}^k+\rho_k g(\hat{x}^k)\big)-\dot{y}\|\le\|\dot{y}\|+\|\hat{y}^k-y^0\|+\rho_k\|g(\hat{x}^k)-g(\dot{x})\|\le\|y^0\|+3\rho_k B_g$. ∎

LEMMA 19. *Given two constants μ_f and μ_h such that $\mu_f+\mu_h>0$, let $\{\tau_k\}\subset(0,1]$ and $\{L_k\}\subset(0,+\infty)$ be two sequences such that $\tau_0:=1$ and $\tau_k^2=(1-\tau_k)\tau_{k-1}^2$. Let $\{x^k\}$ be a given sequence in $\mathbb{R}^p$. We define $\hat{x}^k:=x^k+\beta_k(x^k-x^{k-1})$ with $\beta_k:=\frac{(L_{k-1}+\mu_h)\tau_k(1-\tau_{k-1})}{(L_k-\mu_f)\tau_{k-1}}$. Then if $\mu_f+\mu_h\le L_k-L_{k-1}\le\frac{\mu_f+\mu_h}{\tau_k}$, then for any $x\in\mathbb{R}^p$, we have*

$$\begin{aligned}&\tau_k^2(L_k-\mu_f)\|\tfrac{1}{\tau_k}[\hat{x}^k-(1-\tau_k)x^k]-x\|^2-(\mu_f+\mu_h)\tau_k(1-\tau_k)\|x^k-x\|^2\\ &\le (1-\tau_k)\tau_{k-1}^2(L_{k-1}+\mu_h)\|\tfrac{1}{\tau_{k-1}}[x^k-(1-\tau_{k-1})x^{k-1}]-x\|^2.\end{aligned} \tag{65}$$

Proof. Let $\tilde{x}^k:=\frac{1}{\tau_k}[\hat{x}^k-(1-\tau_k)x^k]$ and $\breve{x}^k:=\frac{1}{\tau_{k-1}}[x^k-(1-\tau_{k-1})x^{k-1}]$. Combining these expressions and $\hat{x}^k=x^k+\beta_k(x^k-x^{k-1})$, we can easily show that $\tilde{x}^k=(1-t_k)x^k+t_k\breve{x}^k$ with $t_k:=\frac{\tau_{k-1}\beta_k}{\tau_k(1-\tau_{k-1})}$. Assuming that $t_k\in[0,1]$. By convexity of $\|\cdot-x\|^2$, we have

$$\|\tilde{x}^k-x\|^2\le t_k\|\breve{x}^k-x\|^2+(1-t_k)\|x^k-x\|^2. \tag{66}$$

On the other hand, using $\tau_k^2=(1-\tau_k)\tau_{k-1}^2$, (65) is equivalent to

$$\|\tilde{x}^k-x\|^2\le\tfrac{L_{k-1}+\mu_h}{L_k-\mu_f}\|\breve{x}^k-x\|^2+\tfrac{(1-\tau_k)(\mu_f+\mu_h)}{(L_k-\mu_f)\tau_k}\|x^k-x\|^2. \tag{67}$$

Let us choose β_k such that $\frac{L_{k-1}+\mu_h}{L_k-\mu_f}=t_k=\frac{\tau_{k-1}\beta_k}{\tau_k(1-\tau_{k-1})}$, leading to $\beta_k=\frac{(L_{k-1}+\mu_h)\tau_k(1-\tau_{k-1})}{(L_{k+1}-\mu_f)\tau_{k-1}}$. Then the condition $\mu_f+\mu_h\le L_k-L_{k-1}$ guarantees that $t_k\in[0,1]$. To guarantee $1-t_k\le\frac{(1-\tau_k)(\mu_f+\mu_h)}{(L_k-\mu_f)\tau_k}$, we need $L_k-L_{k-1}\le\frac{\mu_f+\mu_h}{\tau_k}$. These two conditions show that (66) implies (67). Consequently, (65) holds. ∎

D.2. Proof of Theorem 6. Since $L_{22}=0$ and $\Delta\rho_k := \frac{\rho_k-\rho_{k-1}}{\rho_k}$ due to $\Phi(x,y)=\langle g(x),y\rangle$, and $\hat{y}^k := \dot{y}$ is fixed, for $(x,s)\in\mathrm{dom}(F)\times\mathrm{dom}(H)$, we obtain from (45) that

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1},s^{k+1},\dot{y})-\mathcal{L}(x,s,\tilde{y}^{k+1}) &\le (1-\tau_k)\big[\mathcal{L}_{\rho_{k-1}}(x^k,s^k,\dot{y})-\mathcal{L}(x,s,\tilde{y}^k)\big]\\ &+ \tfrac{\tau_k^2(L_k-\mu_f)}{2}\|\tfrac{1}{\tau_k}[\hat{x}^k-(1-\tau_k)x^k]-x\|^2-\tfrac{\tau_k^2(L_k+\mu_h)}{2}\|\tfrac{1}{\tau_k}[x^{k+1}-(1-\tau_k)x^k]-x\|^2\\ &- \tfrac{(L_k-\mathbf{L}_{11}^k-L_f-\rho_k L_{21}^2)}{2}\|x^{k+1}-\hat{x}^k\|^2-\tfrac{(1-\tau_k)[\rho_{k-1}-\rho_k(1-\tau_k)]}{2\rho_k^2}\|\bar{u}^k-\hat{y}^k\|^2\\ &- \tfrac{(\mu_f+\mu_h)\tau_k(1-\tau_k)}{2}\|x^k-x\|^2.\end{aligned}\tag{68}$$

Since $\mathbf{L}_{11}^k = L_{11}\|u^{k+1}\|$, using Lemma 18(a) and $y^0 := \dot{y}$, we obtain

$$\mathbf{L}_{11}^k \le \begin{cases} L_{11}M_H, & H \text{ is } M_H\text{-Lipschitz continuous},\\ L_{11}(\|\dot{y}\|+\rho_k B_g) & \text{otherwise.}\end{cases}$$

Hence, if we choose L_k as in (31), then $L_k - L_f - \mathbf{L}_{11}^k - \rho_k L_{21}^2 \ge 0$.

Now, let us consider the case where $\|g(x)-g(\dot{x})\|\le B_g$ for all $x\in\mathrm{dom}(F)\cap\mathrm{dom}(g)$, but H is not necessarily Lipschitz continuous. The case when H is M_H-Lipschitz continuous is proven similarly. We divide the proof into two cases as follows.

Case 1 ($\mu_F := \mu_f+\mu_h = 0$). Since $\beta_{k+1} := \frac{(1-\tau_k)\tau_{k+1}}{\tau_k}$, if we define $\tilde{x}^k := \frac{1}{\tau_k}[\hat{x}^k-(1-\tau_k)x^k]$, then we can easily show that $\tilde{x}^{k+1} = \frac{1}{\tau_k}[x^{k+1}-(1-\tau_k)x^k]$. Moreover, since $\rho_k = \frac{\rho_{k-1}}{1-\tau_k}$ and $\tau_k = \frac{1}{k+1}$, we have $\tau_k^2 L_k \le (1-\tau_k)\tau_{k-1}^2 L_{k-1}$.

Case 2 ($\mu_F := \mu_f+\mu_h > 0$). From (31) and $\rho_k = \frac{\rho_{k-1}}{1-\tau_k} = \frac{\rho_0}{\tau_k^2}$, we get $L_k - L_{k-1} = \tau_k\rho_k(L_{21}^2 + L_{11}B_g)$. The condition $\mu_f+\mu_h \le L_k - L_{k-1} \le \frac{\mu_f+\mu_h}{\tau_k}$ in Lemma 19 becomes $\mu_f+\mu_h \le \tau_k\rho_k(L_{21}^2+L_{11}B_g) \le \frac{\mu_f+\mu_h}{\tau_k}$. This condition holds if $\frac{\mu_F\tau_1}{L_{21}^2+L_{11}B_g} \le \rho_0 \le \frac{\mu_F}{L_{21}^2+L_{11}B_g}$.

In both cases, using $\tilde{x}^k := \frac{1}{\tau_k}[\hat{x}^k-(1-\tau_k)x^k]$, $L_k - L_f - \mathbf{L}_{11}^k - \rho_k L_{21}^2 \ge 0$, and $\rho_k(1-\tau_k)-\rho_{k-1}=0$, we can deduce from (68) that

$$\begin{aligned}&\mathcal{L}_{\rho_k}(x^{k+1},s^{k+1},\dot{y})-\mathcal{L}(x,s,\tilde{y}^{k+1})+\tfrac{\tau_k^2(L_k+\mu_h)}{2}\|\tilde{x}^{k+1}-x\|^2\\ &\quad\le (1-\tau_k)\big[\mathcal{L}_{\rho_{k-1}}(x^k,s^k,\dot{y})-\mathcal{L}(x,s,\tilde{y}^k)\big]+\tfrac{\tau_{k-1}^2(L_{k-1}+\mu_h)}{2}\|\tilde{x}^k-x\|^2.\end{aligned}$$

By induction, and using $\tilde{x}^0 := x^0$ and $\tau_0 := 1$, we obtain from the last estimate that

$$\mathcal{L}_{\rho_{k-1}}(x^k,s^k,\dot{y})-\mathcal{L}(x,s,\tilde{y}^k)+\tfrac{\tau_{k-1}^2(L_{k-1}+\mu_h)}{2}\|\tilde{x}^k-x\|^2 \le \tfrac{\omega_k(L_0-\mu_f)}{2}\|x^0-x\|^2,$$

where $\omega_k := \prod_{i=1}^{k-1}(1-\tau_i)$. Using the first line of (11), we have $\mathcal{L}(x^k,s^k,y)\le\mathcal{L}_{\rho_{k-1}}(x^k,s^k,\dot{y})+\frac{1}{2\rho_{k-1}}\|y-\dot{y}\|^2$. Substituting this estimate into the last one, for $s\in\partial H^*(y)$, we get

$$\widetilde{\mathcal{L}}(x^k,y)-\widetilde{\mathcal{L}}(x,\tilde{y}^k) \overset{(6)}{\le} \mathcal{L}(x^k,s^k,y)-\mathcal{L}(x,s,\tilde{y}^k) \le \tfrac{\omega_k(L_0-\mu_f)}{2}\|x^0-x\|^2+\tfrac{1}{2\rho_{k-1}}\|y-\dot{y}\|^2.\tag{69}$$

For Case 1 with $\mu_f+\mu_h=0$, using $\tau_k=\frac{1}{k+1}$, it is obvious to show that $\omega_k := \frac{1}{k}$ and $\rho_{k-1} := \rho_0 k$. Plugging these values into (69) and taking the supremum both sides of (69) over $(x,y)\in\mathcal{Z}$, we obtain (32) with $S_k := \rho_0 k$.

For Case 2 with $\mu_f+\mu_h>0$, the condition $\tau_k = (1-\tau_k)\tau_{k-1}^2$ from Lemma 19 leads to $\tau_k := \frac{\tau_{k-1}}{2}\big[(\tau_{k-1}^2+4)^{1/2}-\tau_{k-1}\big]$ with $\tau_0 := 1$. Hence, we can show that $\omega_k = \frac{\tau_{k-1}^2}{\tau_0^2} = \tau_{k-1}^2 \le \frac{4}{(k+1)^2}$ and $\rho_{k-1} = \frac{\rho_0}{\tau_{k-1}^2} \ge \frac{\rho_0(k+1)^2}{4}$. Substituting them into (69) and taking the supremum both sides of (69) over $(x,y)\in\mathcal{Z}$, we obtain (32) with $S_k := \frac{\rho_0}{4}(k+1)^2$. ∎

D.3. Proof of Theorem 8. Let us denote $v^k := (g(x^{k+1}) - s^{k+1}) - (1-\tau_k)(g(x^k) - s^k)$. Then, from (33), we have $\hat{y}^{k+1} = \hat{y}^k + \eta_k v^k$. Since $\Phi(x,y) = \langle g(x), y\rangle$, for any $y \in \mathrm{dom}(H^*)$, we have

$$\begin{aligned}\mathcal{T}_3 &:= \mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y) - (1-\tau_k)\mathcal{L}_{\rho_{k-1}}(x^k, s^k, y)\\ &= \mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) - (1-\tau_k)\mathcal{L}_{\rho_{k-1}}(x^k, s^k, \hat{y}^k) + \langle v^k, y - \hat{y}^k\rangle\\ &= \mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, \hat{y}^k) - (1-\tau_k)\mathcal{L}_{\rho_{k-1}}(x^k, s^k, \hat{y}^k) + \tfrac{1}{2\eta_k}\|y - \hat{y}^k\|\\ &\quad - \tfrac{1}{2\eta_k}\|y - \hat{y}^{k+1}\|^2 + \tfrac{\eta_k}{2}\|v^k\|^2.\end{aligned}$$

Note that, for $\Phi(x,y) = \langle g(x), y\rangle$, we have $u^{k+1} = u^*_{\rho_k}(\hat{x}^k, s^{k+1}, \hat{y}^k) = \hat{y}^k + \rho_k(g(\hat{x}^k) - s^{k+1})$ and $\bar{u}^k := u^*_{\rho_k}(x^k, s^k, \hat{y}^k) = \hat{y}^k + \rho_k(g(x^k) - s^k)$. Since $\rho_k > \eta_k > 0$, utilizing an elementary inequality $\eta_k\|v\|^2 \le \rho_k\|z\|^2 + \frac{\rho_k\eta_k}{\rho_k - \eta_k}\|v - z\|^2$ and the second line of (2), we can derive that

$$\begin{aligned}\tfrac{\eta_k}{2}\|v^k\|^2 &\le \tfrac{\rho_k}{2}\|g(\hat{x}^k) - s^{k+1} - (1-\tau_k)(g(x^k) - s^k)\|^2 + \tfrac{\rho_k\eta_k}{2(\rho_k - \eta_k)}\|g(x^{k+1}) - g(\hat{x}^k)\|^2\\ &\overset{(2)}{\le} \tfrac{1}{2\rho_k}\|(u^{k+1} - \hat{y}^k) - (1-\tau_k)(\bar{u}^k - \hat{y}^k)\|^2 + \tfrac{\rho_k\eta_k L_{21}^2}{2(\rho_k - \eta_k)}\|x^{k+1} - \hat{x}^k\|^2.\end{aligned}$$

Substituting these expressions into (45) and using $\Delta\rho_k = \frac{\rho_k - \rho_{k-1}}{\rho_k}$ and $L_{22} = 0$, we obtain

$$\begin{aligned}\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y) - \mathcal{L}(x, s, \tilde{y}^{k+1}) &\le (1-\tau_k)\big[\mathcal{L}_{\rho_{k-1}}(x^k, s^k, y) - \mathcal{L}(x, s, \tilde{y}^k)\big]\\ &+ \tfrac{\tau_k^2(L_k - \mu_f)}{2}\|\tfrac{1}{\tau_k}[\hat{x}^k - (1-\tau_k)x^k] - x\|^2 - \tfrac{\tau_k^2(L_k + \mu_h)}{2}\|\tfrac{1}{\tau_k}[x^{k+1} - (1-\tau_k)x^k] - x\|^2\\ &- \tfrac{1}{2}\Big[L_k - \mathbf{L}_{11}^k - L_f - \rho_k L_{21}^2 - \tfrac{L_{21}^2\rho_k\eta_k}{\rho_k - \eta_k}\Big]\|x^{k+1} - \hat{x}^k\|^2 + \tfrac{1}{2\eta_k}\big[\|y - \hat{y}^k\|^2 - \|y - \hat{y}^{k+1}\|^2\big]\\ &- \tfrac{(1-\tau_k)[\rho_{k-1} - \rho_k(1-\tau_k)]}{2\rho_k^2}\|\bar{u}^k - \hat{y}^k\|^2 - \tfrac{(\mu_f + \mu_h)\tau_k(1-\tau_k)}{2}\|x^k - x\|^2.\end{aligned} \tag{70}$$

Since L_k is chosen the same as in Theorem 6, we have $L_k - \mathbf{L}_{11}^k - L_f - \frac{\rho_k\eta_k L_{21}^2}{\rho_k - \eta_k} \ge 0$. Using this fact, $(1-\tau_k)\rho_k = \rho_{k-1}$, and $\eta_k = \frac{\rho_k}{2}$, (70) can be simplified as follows:

$$\begin{aligned}\rho_k[\mathcal{L}_{\rho_k}(x^{k+1}, s^{k+1}, y) - \mathcal{L}(x, s, \tilde{y}^{k+1})] &\le \rho_{k-1}[\mathcal{L}_{\rho_{k-1}}(x^k, s^k, y) - \mathcal{L}(x, s, \tilde{y}^k)]\\ &+ \tfrac{\rho_k\tau_k^2(L_k - \mu_f)}{2}\|\tfrac{1}{\tau_k}[\hat{x}^k - (1-\tau_k)x^k] - x\|^2\\ &- \tfrac{\rho_k\tau_k^2(L_k + \mu_h)}{2}\|\tfrac{1}{\tau_k}[x^{k+1} - (1-\tau_k)x^k] - x\|^2\\ &- \tfrac{\rho_k\tau_k(1-\tau_k)(\mu_f + \mu_h)}{2}\|x^k - x\|^2 + \|\hat{y}^k - y\|^2 - \|\hat{y}^{k+1} - y\|^2.\end{aligned}$$

With this estimate, the proof of the remaining part of Theorem 8 follows an argument very similar to that in the proof of Theorem 6 above. Therefore, we omit repeating it here. ∎

Appendix E. Proof of Theorem 11. We only prove statement (a) corresponding to Theorem 3. Statements (b), (c), and (d) can be proven similarly but using the results of Theorems 4, 6, and 8, respectively.

Since $H(\cdot) = \delta_{-\mathcal{K}}(\cdot)$ and $\Phi(x,y) = \langle g(x), y\rangle$, for any $r > 0$, we have

$$F(\bar{x}^k) - F^\star + r\|g(\bar{x}^k) - \bar{s}^k\| \le \max\big\{\mathcal{L}(\bar{x}^k, \bar{s}^k, y) - \mathcal{L}(x^\star, s^\star, \bar{y}^k) \,:\, \|y\| \le r\big\}. \tag{71}$$

On the other hand, by the saddle-point relation (7), we have $F(\bar{x}^k) + \langle y^\star,\, g(\bar{x}^k) - \bar{s}^k\rangle = \mathcal{L}(\bar{x}^k, \bar{s}^k, y^\star) \ge \mathcal{L}(x^\star, s^\star, y^\star) = F^\star$. By the Cauchy–Schwarz inequality, this leads to

$$F(\bar{x}^k) - F^\star \ge -\langle y^\star,\, g(\bar{x}^k) - \bar{s}^k\rangle \ge -\|y^\star\|\|g(\bar{x}^k) - \bar{s}^k\|. \tag{72}$$

Substituting (71) into (72), choosing $r := \|y^\star\|+1$, and noting that $\bar{s}^k \in -\mathcal{K} = \mathrm{dom}(H)$ due to (13), we can show that

$$\mathrm{dist}_{-\mathcal{K}}\big(g(\bar{x}^k)\big) := \inf_{s\in-\mathcal{K}} \|g(\bar{x}^k)-s\| \le \|g(\bar{x}^k)-\bar{s}^k\| \le \max_{\|y\|\le r} \big\{\mathcal{L}(\bar{x}^k,\bar{s}^k,y)-\mathcal{L}(x^\star,s^\star,\bar{y}^k)\big\}.$$

Therefore, we can easily derive from this inequality and (71) that

$$|F(\bar{x}^k)-F^\star| \le \max\{1,\|y^\star\|\}\max\big\{\mathcal{L}(\bar{x}^k,\bar{s}^k,y)-\mathcal{L}(x^\star,s^\star,\bar{y}^k) \,:\, \|y\|\le r\big\}.$$

Combining the last two estimates, and using (37), we eventually get

$$\mathcal{E}(\bar{x}^k) \le \max\big\{\mathcal{L}(\bar{x}^k,\bar{s}^k,y)-\mathcal{L}(x^\star,s^\star,\bar{y}^k) \,:\, \|y\|\le r\big\}.$$

Utilizing this bound in (63), we arrive at the conclusion in statement (a). ∎

Acknowledgment. The authors would like to sincerely thank two anonymous reviewers for their constructive comments that helped to improve the paper.

REFERENCES

[1] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, in Proceedings of the 34th International Conference on Machine Learning, PMLR, 2017, pp. 214–223.

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

[3] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*, Princeton University Press, Princeton, NJ, 2009.

[4] D. Boob, Q. Deng, and G. Lan, *Proximal Point Methods for Optimization with Nonconvex Functional Constraints*, preprint, https://arxiv.org/abs/1908.02734, 2019.

[5] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vis., 40 (2011), pp. 120–145.

[6] C.-C. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines*, ACM Trans. Intell. Systems Technol., 2 (2011), 27.

[7] Y. Chen, G. Lan, and Y. Ouyang, *Accelerated schemes for a class of variational inequalities*, Math. Program., 165 (2017), pp. 113–149.

[8] P. Combettes and J.-C. Pesquet, *Signal Recovery by Proximal Forward-backward Splitting*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer-Verlag, 2011, pp. 185–212.

[9] P. L. Combettes and J.-C. Pesquet, *Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators*, Set-Valued Var. Anal., 20 (2012), pp. 307–330.

[10] L. Condat, *A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms*, J. Optim. Theory Appl., 158 (2013), pp. 460–479.

[11] D. Davis, *Convergence rate analysis of primal-dual splitting schemes*, SIAM J. Optim., 25 (2015), pp. 1912–1943, https://doi.org/10.1137/151003076.

[12] S. S. Du and W. Hu, *Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity*, in Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 196–205.

[13] J. E. Esser, *Primal-dual Algorithm for Convex Models and Applications to Image Restoration, Registration and Nonlocal Inpainting*, Ph.D. Thesis, University of California, Los Angeles, 2010.

[14] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vol. 1, Springer-Verlag, 2003.

[15] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vol. 2, Springer-Verlag, 2003.

[16] F. Farnia and D. Tse, *A convex duality framework for GANs*, in Proceedings of the 32nd Conference on Neural Information Processing Systems, Curran Associates, 2018, pp. 5248–5258.

[17] N. Golowich, S. Pattathil, C. Daskalakis, and A. Ozdaglar, *Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems*, in Proceedings of 33rd Conference on Learning Theory, PMLR, 2020, pp. 1758–1784.

[18] E. Y. Hamedani and N. S. Aybat, *A primal-dual algorithm with line search for general convex-concave saddle point problems*, SIAM J. Optim., 31 (2021), pp. 1299–1329, https://doi.org/10.1137/18M1213488.
[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.
[20] B. He and X. Yuan, *Convergence analysis of primal-dual algorithms for saddle-point problem: From contraction perspective*, SIAM J. Imaging Sci., 5 (2012), pp. 119–149, https://doi.org/10.1137/100814494.
[21] 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 (2016), pp. 29–56, https://doi.org/10.1137/14096757X.
[22] A. Juditsky and A. Nemirovski, *First order methods for nonsmooth convex large-scale optimization* I*: General purpose methods*, Optim. Mach. Learning, 30 (2011), pp. 121–148.
[23] A. Juditsky, A. Nemirovski, and C. Tauvel, *Solving variational inequalities with stochastic mirror-prox algorithm*, Stoch. Syst., 1 (2011), pp. 17–58.
[24] S.-J. Kim and S. Boyd, *A minimax theorem with applications to machine learning, signal processing, and finance*, SIAM J. Optim., 19 (2008), pp. 1344–1367, https://doi.org/10.1137/060677586.
[25] O. Kolossoski and R. D. C. Monteiro, *An accelerated non-Euclidean hybrid proximal extragradient-type algorithm for convex–concave saddle-point problems*, Optim. Methods Softw., 32 (2017), pp. 1244–1272.
[26] 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 (2004), pp. 27–72.
[27] T. Lin, C. Jin, and M. I. Jordan, *Near-optimal algorithms for minimax optimization*, in Proceedings of the 33rd Annual Conference on Learning Theory, PMLR, 2020, pp. 2738–2779.
[28] T. Lin, C. Jin, and M. I. Jordan, *On gradient descent ascent for nonconvex-concave minimax problems*, in Proceedings of the 37th International Conference on Machine Learning, PMLR, 2020, pp. 6083–6093.
[29] A. Mokhtari, A. Ozdaglar, and S. Pattathil, *A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach*, in Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 1497–1507.
[30] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil, *Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems*, SIAM J. Optim., 30 (2020), pp. 3230–3251, https://doi.org/10.1137/19M127375X.
[31] R. D. C. Monteiro and B. F. Svaiter, *On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean*, SIAM J. Optim., 20 (2010), pp. 2755–2787, https://doi.org/10.1137/090753127.
[32] R. D. C. Monteiro and B. F. Svaiter, *Complexity of variants of Tseng's modified F-B splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems*, SIAM J. Optim., 21 (2011), pp. 1688–1720, https://doi.org/10.1137/100801652.
[33] A. Nemirovski, *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 (2004), pp. 229–251, https://doi.org/10.1137/S1052623403425629.
[34] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Kluwer Academic, 2004.
[35] Y. Nesterov, *Excessive gap technique in nonsmooth convex minimization*, SIAM J. Optim., 16 (2005), pp. 235–249, https://doi.org/10.1137/S1052623403422285.
[36] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.
[37] Y. Nesterov, *Dual extrapolation and its applications to solving variational inequalities and related problems*, Math. Program., 109 (2007), pp. 319–344.
[38] G. Peyré and M. Cuturi, *Computational Optimal Transport*, Found. Trends Mach. Learn., 11 (2019), pp. 355–607.
[39] H. Rahimian and S. Mehrotra, *Distributionally Robust Optimization: A Review*, preprint, https://arxiv.org/abs/1908.05659, 2019.
[40] R. Rockafellar and R. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer, 2004.
[41] S. Sabach and M. Teboulle, *Faster Lagrangian-based methods in convex optimization*, SIAM

J. Optim., 32 (2022), pp. 204–227, https://doi.org/10.1137/20M1375358.
[42] A. Shapiro and S. Ahmed, *On a class of minimax stochastic programs*, SIAM J. Optim., 14 (2004), pp. 1237–1249, https://doi.org/10.1137/S1052623403434012.
[43] R. Shefi and M. Teboulle, *Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization*, SIAM J. Optim., 24 (2014), pp. 269–297, https://doi.org/10.1137/130910774.
[44] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh, *Efficient algorithms for smooth minimax optimization*, in Proceedings of the 33rd Conference on Neural Information Processing Systems, Curran Associates, 2019, pp. 12659–12670.
[45] K. H. Le Thi, R. Zhao, and W. B. Haskell, *An Inexact Primal-Dual Smoothing Framework for Large-Scale Non-bilinear Saddle Point Problems*, preprint, https://arxiv.org/abs/1711.03669, 2017.
[46] Q. Tran-Dinh, O. Fercoq, and V. Cevher, *A smooth primal-dual optimization framework for nonsmooth composite convex minimization*, SIAM J. Optim., 28 (2018), pp. 96–134, https://doi.org/10.1137/16M1093094.
[47] 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. Automat. Control, 57 (2012), pp. 1377–1390.
[48] Q. Tran-Dinh and Y. Zhu, *Non-stationary first-order primal-dual algorithms with faster convergence rates*, SIAM J. Optim., 30 (2020), pp. 2866–2896, https://doi.org/10.1137/19M1293855.
[49] P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization*, unpublished manuscript, 2008; available from https://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf.
[50] T. Valkonen, *Inertial, corrected, primal-dual proximal splitting*, SIAM J. Optim., 30 (2020), pp. 1391–1420, https://doi.org/10.1137/18M1182851.
[51] C. B. Vu, *A splitting algorithm for dual monotone inclusions involving co-coercive operators*, Adv. Comput. Math., 38 (2013), pp. 667–681.
[52] Y. Wang and J. Li, *Improved algorithms for convex-concave minimax optimization*, in Proceedings of the 34th Conference on Neural Information Processing Systems, Curran Associates, 2020, pp. 4800–4810.
[53] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, *Maximum margin clustering*, in Advances in Neural Information Processing Systems 18, MIT Press, 2005, pp. 1537–1544.
[54] Y. Xu, *Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming*, Math. Program., 185 (2021), pp. 199–244.
[55] Y. Xu, *First-order methods for constrained convex programming based on linearized augmented Lagrangian function*, INFORMS J. Optim., 3 (2021), pp. 89–117.
[56] Y. Yan, Y. Xu, Q. Lin, W. Liu, and T. Yang, *Sharp Analysis of Epoch Stochastic Gradient Descent Ascent Methods for Min-Max Optimization*, preprint, https://arxiv.org/abs/2002.05309, 2020.
[57] J. Yang, S. Zhang, N. Kiyavash, and N. He, *A catalyst framework for minimax optimization*, in Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates, 2020, pp. 5667–5678.
[58] R. Zhao, *Accelerated stochastic algorithms for convex-concave saddle-point problems*, Math. Oper. Res., 47 (2022), pp. 1443–1473.