Solving bilevel optimization via sequential minimax optimization*

Zhaosong Lu † Sanyou Mei ‡

May 12, 2024 (Revised: November 5, 2025)

Abstract

In this paper we propose a sequential minimax optimization (SMO) method for solving a class of constrained bilevel optimization problems in which the lower-level part is a possibly nonsmooth convex optimization problem, while the upper-level part is a possibly nonconvex optimization problem. Specifically, SMO applies a first-order method to solve a sequence of minimax subproblems, which are obtained by employing a hybrid of modified augmented Lagrangian and penalty schemes on the bilevel optimization problems. Under suitable assumptions, we establish an *operation complexity* of $\mathcal{O}(\varepsilon^{-7}\log\varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^{-6}\log\varepsilon^{-1})$, measured in terms of fundamental operations, for SMO in finding an ε-KKT (Karush-Kuhn-Tucker) solution of the bilevel optimization problems with merely convex and strongly convex lower-level objective functions, respectively. The latter result improves the previous best-known operation complexity by a factor of ε^{-1}. Preliminary numerical results demonstrate significantly superior computational performance compared to the recently developed first-order penalty method.

Keywords: bilevel optimization, minimax optimization, first-order methods, operation complexity

Mathematics Subject Classification: 90C26, 90C30, 90C47, 90C99, 65K05

1 Introduction

Bilevel optimization (BO) is a two-level hierarchical optimization, which is typically in the form of

$$\begin{aligned} f^* = \min \quad & f(x,y) \\ \text{s.t.} \quad & y \in \operatorname*{argmin}_z\{\tilde{f}(x,z)|\tilde{g}(x,z)\le 0\}.^1 \end{aligned} \tag{1}$$

BO has widely been used in many areas, including adversarial training [51, 52, 64], continual learning [43], hyperparameter tuning [3, 19, 54], image reconstruction [10], meta-learning [4, 30, 57], neural architecture search [17, 41], reinforcement learning [25, 33], and Stackelberg games [67]. More applications about it can be found in [2, 9, 13, 14, 15, 61] and the references therein. Theoretical properties including optimality conditions of (1) have been extensively studied in the literature (e.g., see [15, 16, 49, 66, 75]).

Numerous methods have been developed for solving some special cases of (1). For example, constraint-based methods [23, 60], deterministic gradient-based methods [18, 19, 21, 26, 50, 56, 57], and stochastic gradient-based methods [7, 22, 25, 27, 28, 31, 32, 36, 37, 71] were proposed for solving (1) with $\tilde{g}\equiv 0$, f, $\tilde{f}$ being smooth, and $\tilde{f}$ being *strongly convex* with respect to y. For a similar case as this but with $\tilde{f}$ being *convex* with respect to y, a zeroth-order method was recently proposed in [6], and also numerical methods were developed in [38, 63, 40] by solving (1) as a single or sequential smooth constrained optimization problems. Besides, when all the functions in (1) are smooth and $\tilde{f}$, $\tilde{g}$ are *convex* with respect to y, gradient-type methods were proposed by solving a mathematical program with equilibrium constraints resulting from replacing the lower-level optimization problem of (1) by its first-order optimality conditions (e.g., see [1, 48, 55]). Furthermore, a single-loop gradient method was recently introduced in [74] based on

*This work was partially supported by the Air Force Office of Scientific Research under Award FA9550-24-1-0343, the Office of Naval Research under Award N00014-24-1-2702, and the National Science Foundation under Awards 2211491 and 2435911. It was primarily conducted during Sanyou Mei's Ph.D. studies at the University of Minnesota.

†Department of Industrial and Systems Engineering, University of Minnesota, USA (email: zhaosong@umn.edu).

‡Department of Industrial Engineering and Decision Analytics, the Hong Kong University of Science and Technology, Hong Kong, China (email: symei@ust.hk).

[1]For ease of reading, throughout this paper the tilde symbol is particularly used for the functions related to the lower-level optimization problem. Besides, "argmin" denotes the set of optimal solutions of the associated problem.

a novel reformulation of the bilevel optimization problem as a single-level smooth optimization problem using Moreau envelope of the Lagrangian function of the lower-level problem. Recently, difference-of-convex (DC) algorithms were developed in [76] for solving (1) with f being a DC function, and $\tilde{f}$, $\tilde{g}$ being convex functions. Lately, a practically efficient multi-stage gradient descent and ascent algorithm was developed in [68] for (1) with $\tilde{g} \equiv 0$, f being convex and Lipschitz continuous, and $\tilde{f}$ being strongly convex and Lipschitz smooth via solving the aforementioned minimax reformulation of (1). In addition, penalty methods were proposed in [29, 46, 58] for solving (1). Notably, the paper [46] demonstrates for the first time that BO can be approximately solved as minimax optimization. Specifically, it reformulates BO as minimax optimization by a novel double penalty scheme and proposes a first-order method with complexity guarantees for BO via solving a single minimax problem. In addition, a novel single-loop Hessian-free algorithm based on a doubly regularized gap function was proposed in [73] for solving (1). More discussion on algorithmic development for BO can be found in [2, 9, 15, 42, 62, 66]) and the references therein.

In this paper, we consider problem (1) under similar assumptions as in [46]. Specifically, we assume that problem (1) has at least one optimal solution and satisfies the following assumptions.

Assumption 1. *(i) $f(x,y) = f_1(x,y) + f_2(x)$ and $\tilde{f}(x,y) = \tilde{f}_1(x,y) + \tilde{f}_2(y)$ are respectively L_f- and $L_{\tilde{f}}$-Lipschitz continuous on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} := \operatorname{dom} f_2$ and $\mathcal{Y} := \operatorname{dom} \tilde{f}_2$, where $\tilde{f}_1(x,\cdot)$ is σ-strongly-convex for any given $x \in \mathcal{X}$ for some $\sigma \geq 0$,[2] f_1, $\tilde{f}_1$ are respectively $L_{\nabla f_1}$- and $L_{\nabla \tilde{f}_1}$-smooth on $\mathcal{X} \times \mathcal{Y}$, and $f_2 : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $\tilde{f}_2 : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ are proper closed convex functions.*

(ii) The proximal operators associated with f_2 and $\tilde{f}_2$ can be exactly evaluated.

(iii) $\tilde{g} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^l$ is $L_{\tilde{g}}$-Lipschitz continuous and $L_{\nabla \tilde{g}}$-smooth on $\mathcal{X} \times \mathcal{Y}$, and $\tilde{g}_i(x,\cdot)$ is convex for all $x \in \mathcal{X}$ and $i = 1, 2, \ldots, l$.

(iv) The sets $\mathcal{X}$ and $\mathcal{Y}$ (namely, $\operatorname{dom} f_2$ and $\operatorname{dom} \tilde{f}_2$) are compact.

Due to the sophisticated structure described in Assumption 1, existing methods except the first-order penalty method [46] are generally not applicable to problem (1). In particular, instead of solving (1) directly, the latter method applies a first-order method [44] to solve an approximate counterpart of (1) given by a single minimax problem

$$\min_{x,y} \max_{z} f(x,y) + \rho\big(\tilde{f}(x,y) + \mu \left\|[\tilde{g}(x,y)]_+\right\|^2 - \tilde{f}(x,z) - \mu \left\|[\tilde{g}(x,z)]_+\right\|^2\big) \tag{2}$$

with a suitable choice of penalty parameters $\rho, \mu > 0$. Notice that the minimax problem (2) can be obtained from (1) by performing two steps: (i) apply the classical quadratic penalty scheme to the lower-level problem of (2) and approximate (1) by a simpler BO problem, $\min_{x,y}\{f(x,y) | y \in \operatorname{argmin}_z \tilde{f}(x,z) + \mu \left\|[\tilde{g}(x,z)]_+\right\|^2\}$, which can be viewed as

$$\min_{x,y} \big\{ f(x,y) \big| \tilde{f}(x,y) + \mu \left\|[\tilde{g}(x,y)]_+\right\|^2 \leq \min_z \tilde{f}(x,z) + \mu \left\|[\tilde{g}(x,z)]_+\right\|^2 \big\}, \tag{3}$$

and (ii) apply a penalty method to (3) to obtain the minimax problem (2). While this method enjoys complexity guarantees for finding an approximate KKT (Karush-Kuhn-Tucker) solution of (1), it may suffer from practical inefficiency issues. Specifically, the penalty parameters are pre-chosen to achieve a desired operational complexity and may be overly large in practice. Additionally, the classical quadratic penalty scheme is used to obtain the minimax problem (2), and its associated penalty parameter could be much larger than the one associated with an augmented Lagrangian scheme.

To address the aforementioned issues, in this paper we propose a novel sequential minimax optimization (SMO) method to solve problem (1), which substantially outperforms the first-order penalty method [46] as observed in our numerical experiment. Specifically, instead of using the classical quadratic penalty scheme for the lower-level problem of (1), we propose a new augmented Lagrangian scheme by replacing the quadratic penalty function $\tilde{f}(x,y) + \mu \left\|[\tilde{g}(x,y)]_+\right\|^2$ with a *modified* augmented Lagrangian function $\tilde{f}(x,z) + \frac{1}{2\rho\mu}\left(\left\|[\lambda + \mu\tilde{g}(x,z)]_+\right\|^2 - \|\lambda\|^2\right)$ for $\lambda \in \mathbb{R}^l_+$ and $\mu > 0$.[3] Performing such a replacement in (2)

[2] $\tilde{f}_1(x,\cdot)$ is either merely convex for any given $x \in \mathcal{X}$ if $\sigma = 0$ or strongly convex with parameter σ if $\sigma > 0$.

[3] The standard augmented Lagrangian function associated with the lower-level problem of (1) is $\tilde{f}(x,z) + \frac{1}{2\mu}\left(\left\|[\lambda + \mu\tilde{g}(x,z)]_+\right\|^2 - \|\lambda\|^2\right)$. Therefore, $\tilde{f}(x,z) + \frac{1}{2\rho\mu}\left(\left\|[\lambda + \mu\tilde{g}(x,z)]_+\right\|^2 - \|\lambda\|^2\right)$ can be viewed as a modified augmented Lagrangian function. Its advantage over the standard augmented Lagrangian function will be discussed in Section 2.

results in a new minimax problem

$$\min_{x,y}\max_{z} f(x,y)+\rho\Big(\tilde{f}(x,y)+\frac{1}{2\rho\mu}\|[\lambda+\mu\tilde{g}(x,y)]_+\|^2-\tilde{f}(x,z)-\frac{1}{2\rho\mu}\|[\lambda+\mu\tilde{g}(x,z)]_+\|^2\Big). \tag{4}$$

Our SMO method solves a sequence of minimax subproblems in the form of (4). Specifically, let $\{\rho_k\}$, $\{\mu_k\}$, (x^0,y^0,z^0,λ^0) be given. At each iteration $k\geq 0$, SMO finds an approximate solution $(x^{k+1},y^{k+1},z^{k+1})$ of (4) with $\lambda=\lambda^k$, $\rho=\rho_k$ and $\mu=\mu_k$, starting at (x^k,y^k,z^k), and then updates λ^{k+1} according to $\lambda^{k+1}=[\lambda^k+\mu_k\tilde{g}(x^{k+1},z^{k+1})]_+$. The resulting SMO enjoys the following notable features.

- It uses only the first-order information of the problem. Specifically, its fundamental operations consist only of gradient evaluations of $\tilde{g}$ and the smooth component of f and $\tilde{f}$ and also proximal operator evaluations of the nonsmooth component of f and $\tilde{f}$ (see Algorithm 1).
- It has theoretical guarantees on operation complexity, which is measured by the aforementioned fundamental operations, for finding an ε-KKT solution of (1). Specifically, it enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-7}\log\varepsilon^{-1})$ when the lower objective function $\tilde{f}_1(x,\cdot)$ is merely convex (see Theorem 1). Moreover, it enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-6}\log\varepsilon^{-1})$ when $\tilde{f}_1(x,\cdot)$ is strongly convex, which *improves* the previous best-known operation complexity [46, Theorem 5] by a factor of ε^{-1} (see Theorem 2).
- It demonstrates significantly superior computational performance compared to the first-order penalty method [46] (see Section 3).

The rest of this paper is organized as follows. In Subsection 1.1, we introduce some notation and terminology. In Section 2, we propose a sequential minimax optimization method for solving (1) and study its complexity. Preliminary numerical results and the proofs of the main results are presented in Sections 3 and 4, respectively.

1.1 Notation and terminology

The following notation will be used throughout this paper. Let $\mathbb{R}^n$ denote the Euclidean space of dimension n and $\mathbb{R}^n_+$ denote the nonnegative orthant in $\mathbb{R}^n$. The standard inner product, l_1-norm and Euclidean norm are denoted by $\langle\cdot,\cdot\rangle$, $\|\cdot\|_1$ and $\|\cdot\|$, respectively. For any $v\in\mathbb{R}^n$, let v_+ denote the nonnegative part of v, that is, $(v_+)_i=\max\{v_i,0\}$ for all i. For any two vectors u and v, $(u;v)$ denotes the vector resulting from stacking v under u. Given a point x and a closed set S in $\mathbb{R}^n$, let $\mathrm{dist}(x,S)=\min_{x'\in S}\|x'-x\|$ and $\mathscr{I}_S$ denote the indicator function associated with S.

A function or mapping ϕ is said to be L_ϕ-*Lipschitz continuous* on a set S if $\|\phi(x)-\phi(x')\|\leq L_\phi\|x-x'\|$ for all $x,x'\in S$. In addition, it is said to be $L_{\nabla\phi}$-*smooth* on S if $\|\nabla\phi(x)-\nabla\phi(x')\|\leq L_{\nabla\phi}\|x-x'\|$ for all $x,x'\in S$. For a closed convex function $p:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$, associated with p is denoted by prox_p, that is,

$$\mathrm{prox}_p(x)=\operatorname*{argmin}_{x'\in\mathbb{R}^n}\left\{\frac{1}{2}\|x'-x\|^2+p(x')\right\}\quad\forall x\in\mathbb{R}^n.$$

Given that evaluation of $\mathrm{prox}_{\gamma p}(x)$ is often as cheap as $\mathrm{prox}_p(x)$, we count the evaluation of $\mathrm{prox}_{\gamma p}(x)$ as one evaluation of proximal operator of p for any $\gamma>0$ and $x\in\mathbb{R}^n$.

For a lower semicontinuous function $\phi:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$, its *domain* is the set $\mathrm{dom}\,\phi:=\{x|\phi(x)<\infty\}$. The *upper subderivative* of ϕ at $x\in\mathrm{dom}\,\phi$ in a direction $d\in\mathbb{R}^n$ is defined by

$$\phi'(x;d)=\limsup_{x'\xrightarrow{\phi}x,\,t\downarrow 0}\inf_{d'\to d}\frac{\phi(x'+td')-\phi(x')}{t},$$

where $t\downarrow 0$ means both $t>0$ and $t\to 0$, and $x'\xrightarrow{\phi}x$ means both $x'\to x$ and $\phi(x')\to\phi(x)$. The *subdifferential* of ϕ at $x\in\mathrm{dom}\,\phi$ is the set

$$\partial\phi(x)=\{s\in\mathbb{R}^n\big|s^Td\leq\phi'(x;d)\ \ \forall d\in\mathbb{R}^n\}.$$

We use $\partial_{x_i}\phi(x)$ to denote the subdifferential with respect to x_i. In addition, for an upper semicontinuous function ϕ, its subdifferential is defined as $\partial\phi=-\partial(-\phi)$. If ϕ is locally Lipschitz continuous, the above definition of subdifferential coincides with the Clarke subdifferential. Besides, if ϕ is convex, it coincides with the ordinary subdifferential for convex functions. Also, if ϕ is continuously differentiable at x , we

simply have $\partial\phi(x) = \{\nabla\phi(x)\}$, where $\nabla\phi(x)$ is the gradient of ϕ at x. In addition, it is not hard to verify that $\partial(\phi_1 + \phi_2)(x) = \nabla\phi_1(x) + \partial\phi_2(x)$ if ϕ_1 is continuously differentiable at x and ϕ_2 is lower or upper semicontinuous at x. See [8, 69] for more details.

Finally, we introduce an (approximate) primal-dual stationary point (e.g., see [11, 12, 34]) for a general minimax problem

$$\min_x \max_y \Psi(x, y), \tag{5}$$

where $\Psi(\cdot, y) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a lower semicontinuous function, and $\Psi(x, \cdot) : \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$ is an upper semicontinuous function.

Definition 1. *A point (x, y) is said to be a primal-dual stationary point of the minimax problem (5) if*

$$0 \in \partial_x \Psi(x, y), \quad 0 \in \partial_y \Psi(x, y).$$

In addition, for any $\epsilon > 0$, a point (x_ϵ, y_ϵ) is said to be an ϵ-primal-dual stationary point of the minimax problem (5) if

$$\mathrm{dist}\,(0, \partial_x \Psi(x_\epsilon, y_\epsilon)) \le \epsilon, \quad \mathrm{dist}\,(0, \partial_y \Psi(x_\epsilon, y_\epsilon)) \le \epsilon.$$

2 A sequential minimax optimization method for problem (1)

As discussed in Section 1, the first-order penalty method [46] may suffer practical inefficiency issues due to possibly overly large penalty parameters. To address these issues, in this section we propose a sequential minimax optimization (SMO) method for finding an approximate KKT solution of (1), which substantially outperforms the first-order penalty method [46] as observed in our numerical experiment.

To motivate our SMO method, we first apply a *modified* augmented Lagrangian (AL) scheme to migrate the constraint $\tilde{g}(x, y) \le 0$ of the lower-level problem of (1) to its objective function and obtain an approximation to (1) given by

$$\begin{array}{ll} \min & f(x, y) \\ \text{s.t.} & y \in \operatorname{argmin}_z \{\tilde{f}(x, z) + \frac{1}{2\rho\mu}\left(\|[\lambda + \mu\tilde{g}(x, z)]_+\|^2 - \|\lambda\|^2\right)\} \end{array} \tag{6}$$

for some $\rho, \mu > 0$ and $\lambda \in \mathbb{R}^l_+$. By applying a penalty scheme, problem (6) can be approximated by

$$\min_{x,y} f(x, y) + \rho\left(\tilde{f}(x, y) + \frac{1}{2\rho\mu}\left(\|[\lambda + \mu\tilde{g}(x, y)]_+\|^2 - \|\lambda\|^2\right) - \min_z \left\{\tilde{f}(x, z) + \frac{1}{2\rho\mu}\left(\|[\lambda + \mu\tilde{g}(x, z)]_+\|^2 - \|\lambda\|^2\right)\right\}\right),$$

which is equivalent to the minimax problem

$$\min_{x,y} \max_z \mathcal{L}(x, y, z, \lambda; \rho, \mu), \tag{7}$$

where

$$\mathcal{L}(x, y, z, \lambda; \rho, \mu) := f(x, y) + \rho\tilde{f}(x, y) + \frac{1}{2\mu}\|[\lambda + \mu\tilde{g}(x, y)]_+\|^2 - \rho\tilde{f}(x, z) - \frac{1}{2\mu}\|[\lambda + \mu\tilde{g}(x, z)]_+\|^2. \tag{8}$$

As observed above, the function $\tilde{f}(x, z) + \frac{1}{2\rho\mu}\left(\|[\lambda + \mu\tilde{g}(x, z)]_+\|^2 - \|\lambda\|^2\right)$, which serves as a modified AL function for the lower-level problem of (1), induces the minimax problem (7), where the penalty parameters ρ and μ are separately associated with the lower-level objective and constraint functions, respectively. This separation plays a crucial role in designing a practically efficient SMO method that achieves the desired operation complexity. In contrast, the standard AL function $\tilde{f}(x, z) + \frac{1}{2\mu}\left(\|[\lambda + \mu\tilde{g}(x, z)]_+\|^2 - \|\lambda\|^2\right)$ for the lower-level problem of (1) leads to a minimax problem that lacks this property and over-penalizes the constraints. As a result, the corresponding first-order method based on such a minimax problem cannot guarantee finding an ε-KKT solution of problem (1), since the condition $|\tilde{f}(x^k, y^k) - \tilde{f}^*(x^k)| \le \epsilon$ may fail to hold for the generated solution sequence when k is sufficiently large, which is required by the definition of an ε-KKT point (see Definition 2). Furthermore, under Assumption 1, one can observe that $\mathcal{L}$ possesses the following desirable structure.

- For any given $\rho, \mu > 0$ and $\lambda \in \mathbb{R}^l_+$, $\mathcal{L}$ is the sum of smooth function $h(x, y, z)$ with Lipschitz continuous gradient and possibly nonsmooth function $p(x, y) - q(z)$ with exactly computable proximal operator, where

$$h(x, y, z) = f_1(x, y) + \rho\tilde{f}_1(x, y) + \frac{1}{2\mu}\|[\lambda + \mu\tilde{g}(x, y)]_+\|^2 - \rho\tilde{f}_1(x, z) - \frac{1}{2\mu}\|[\lambda + \mu\tilde{g}(x, z)]_+\|^2,$$
$$p(x, y) = f_2(x) + \rho\tilde{f}_2(y), \quad q(z) = \rho\tilde{f}_2(z).$$

- $\mathcal{L}$ is nonconvex in (x, y) but $\rho\sigma$-strongly-concave in z.

Thanks to the above nice structure of $\mathcal{L}$, an approximate primal-dual stationary point of problem (7) can be suitably found by Algorithm 5 (see Appendix B). Additionally, recall from the above discussion that the minimax problem provides an approximation to the bilevel optimization problem (1). Based on these observations, we propose solving problem (1) by iteratively solving a sequence of minimax subproblems in the form of (7), similar to the classical AL method for constrained nonlinear optimization.

Specifically, let $\{(\rho_k, \mu_k)\}$ be a sequence of penalty parameters. Given the current iterate $(x^k, y^k, z^k, \lambda^k)$, our method calls Algorithm 2 or 3 (see Appendix A), depending on $\sigma = 0$ or $\sigma > 0$, to obtain an approximate solution $y^k_{\rm init}$ of $\min_z \widetilde{L}(x^k, z, \lambda^k; \rho_k, \mu_k)$, where

$$\widetilde{\mathcal{L}}(x, z, \lambda; \rho, \mu) := \tilde{f}(x, z) + \frac{1}{2\rho\mu}\|[\lambda + \mu\tilde{g}(x, z)]_+\|^2. \tag{9}$$

It then uses $(x^k, y^k, y^k_{\rm init})$ as the initial point and calls Algorithm 5 (see Appendix B) with properly chosen parameters to find an approximate primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of $\min_{x,y}\max_z \mathcal{L}(x, y, z, \lambda^k; \rho_k, \mu_k)$. Subsequently, it updates λ^{k+1} in a standard manner.

Let $\tilde{g}_{\rm hi} = \max\{\|\tilde{g}(x, y)\| | (x, y) \in \mathcal{X} \times \mathcal{Y}\}$. We now present our method for solving problem (1) below.

Algorithm 1 A sequential minimax optimization (SMO) method for (1)

Input: $\varepsilon, \tau \in (0, 1)$, $\epsilon_0 \in (\tau\varepsilon, 1]$, $x^0 \in \mathcal{X}$, $z^0 \in \mathcal{Y}$, $\epsilon_k = \epsilon_0\tau^k$, $\rho_k = \epsilon_k^{-1}$, $\mu_k = \epsilon_k^{-3}$, and $\lambda^0 \in \mathbb{R}^l_+$.

1: **for** $k = 0, 1 \ldots$ **do**

2: Call Algorithm 2 (see Appendix A) if $\sigma = 0$ or Algorithm 3 (see Appendix A) if $\sigma > 0$ with $\Psi(\cdot) \leftarrow \widetilde{\mathcal{L}}(x^k, \cdot, \lambda^k; \rho_k, \mu_k)$, $\tilde{\epsilon} \leftarrow \epsilon_k$, $\sigma_\phi \leftarrow \sigma$, $L_{\nabla\phi} \leftarrow \widetilde{L}_k$, $\tilde{x}^0 \leftarrow y^k$ to find an approximate solution $y^k_{\rm init}$ of $\min_z \widetilde{L}(x^k, z, \lambda^k; \rho_k, \mu_k)$ such that

$$\widetilde{\mathcal{L}}(x^k, y^k_{\rm init}, \lambda^k; \rho_k, \mu_k) - \min_z \widetilde{\mathcal{L}}(x^k, z, \lambda^k; \rho_k, \mu_k) \le \epsilon_k, \tag{10}$$

where $\widetilde{\mathcal{L}}$ is given in (9) and

$$\widetilde{L}_k = L_{\nabla\tilde{f}_1} + \rho_k^{-1}(\mu_k L^2_{\tilde{g}} + \mu_k\tilde{g}_{\rm hi}L_{\nabla\tilde{g}} + \|\lambda^k\|L_{\nabla\tilde{g}}). \tag{11}$$

3: Call Algorithm 5 (see Appendix B) with $\epsilon \leftarrow \epsilon_k$, $\hat{x}^0 \leftarrow (x^k, y^k_{\rm init})$, $\hat{y}^0 \leftarrow z^k$, $L_{\nabla h} \leftarrow L_k$, and $\hat{\epsilon}_0 \leftarrow \epsilon_k/(2\sqrt{\mu_k})$ if $\sigma = 0$ and $\hat{\epsilon}_0 \leftarrow \epsilon_k/2$ if $\sigma > 0$ to find an ϵ_k-primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of

$$\min_{x,y}\max_z \mathcal{L}(x, y, z, \lambda^k; \rho_k, \mu_k), \tag{12}$$

where

$$L_k = L_{\nabla f_1} + 2\rho_k L_{\nabla\tilde{f}_1} + 2\mu_k L^2_{\tilde{g}} + 2\mu_k\tilde{g}_{\rm hi}L_{\nabla\tilde{g}} + 2\|\lambda^k\|L_{\nabla\tilde{g}}. \tag{13}$$

4: Set $\lambda^{k+1} = [\lambda^k + \mu_k\tilde{g}(x^{k+1}, z^{k+1})]_+$.

5: If $\epsilon_k \le \varepsilon$, terminate the algorithm and output (x^{k+1}, y^{k+1}).

6: **end for**

We next provide some remarks regarding the well-definedness of Algorithm 1.

Remark 1. *Notice that* $\widetilde{L}(x^k, y, \lambda^k; \rho_k, \mu_k) = \phi(y) + \tilde{f}_2(y)$ *with* $\phi(y) = \tilde{f}_1(x^k, y) + \|[\lambda^k + \mu_k\tilde{g}(x^k, y)]_+\|^2/(2\rho_k\mu_k)$. *By Assumption 1 and* (28)*, one can see that* ϕ *is* $\widetilde{L}_k$*-smooth and* σ*-strongly-convex on* $\operatorname{dom} P$ *and the proximal operator of* $\tilde{f}_2$ *can be exactly evaluated. It then follows from this and Theorems 3 and 4 (see Appendix A) that* $y^k_{\rm init}$ *satisfying* (10) *can be successfully found in step 2 of Algorithm 1 by applying Algorithm 2 or 3 to the problem* $\min_z \widetilde{L}(x^k, z, \lambda^k; \rho_k, \mu_k)$*. In addition, by Theorem 5 (see Appendix B), one can see that an* ϵ_k*-primal-dual stationary point of* (12) *can be successfully found in step 3 of Algorithm 1 by applying Algorithm 5 to problem* (12)*. Consequently, Algorithm 1 is well-defined.*

2.1 Complexity results for Algorithm 1

In this subsection we study *iteration and operation complexity* for Algorithm 1. In particular, in order to characterize the approximate solution found by Algorithm 1, we first introduce a notion called an

ε-KKT solution of problem (1). Then we establish iteration and operation complexity of Algorithm 1 for finding an $\mathcal{O}(\varepsilon)$-KKT solution of (1).

For notational convenience, we define

$$\tilde{f}^*(x) := \min_z \{\tilde{f}(x,z) | \tilde{g}(x,z) \le 0\}. \tag{14}$$

Observe that problem (1) can be equivalently reformulated as

$$\min_{x,y} \{f(x,y) | \tilde{f}(x,y) \le \tilde{f}^*(x),\ \tilde{g}(x,y) \le 0\}. \tag{15}$$

The Lagrangian function associated with (15) is given by

$$\widehat{\mathcal{L}}(x,y,\rho,\lambda_{\mathbf{y}}) = f(x,y) + \rho(\tilde{f}(x,y) - \tilde{f}^*(x)) + \langle \lambda_{\mathbf{y}}, \tilde{g}(x,y)\rangle. \tag{16}$$

In the same spirit of classical constrained optimization, one would naturally be interested in a KKT solution (x,y) of (15), namely, (x,y) satisfies

$$\tilde{f}(x,y) \le \tilde{f}^*(x), \quad \tilde{g}(x,y) \le 0, \quad \rho(\tilde{f}(x,y) - \tilde{f}^*(x)) = 0, \quad \langle \lambda_{\mathbf{y}}, \tilde{g}(x,y)\rangle = 0, \tag{17}$$

and moreover (x,y) is a stationary point of the problem

$$\min_{x',y'} \widehat{\mathcal{L}}(x',y',\rho,\lambda_{\mathbf{y}}) \tag{18}$$

for some $\rho \ge 0$ and $\lambda_{\mathbf{y}} \in \mathbb{R}^l_+$. Yet, due to the sophisticated problem structure, characterizing a stationary point of (18) is generally difficult. On the other hand, notice from Lemma 1 and (16) that problem (18) is equivalent to the minimax problem

$$\min_{x',y',\lambda'_{\mathbf{z}}} \max_{z'} \big\{ f(x',y') + \rho\big(\tilde{f}(x',y') - \tilde{f}(x',z') - \langle \lambda'_{\mathbf{z}}, \tilde{g}(x',z')\rangle\big) + \langle \lambda_{\mathbf{y}}, \tilde{g}(x',y')\rangle + \mathscr{I}_{\mathbb{R}^l_+}(\lambda'_{\mathbf{z}}) \big\},^{4}$$

whose stationary point $(x,y,\lambda_{\mathbf{z}},z)$, according to Definition 1 and Assumption 1, satisfies

$$0 \in \partial f(x,y) + \rho\partial\tilde{f}(x,y) - \rho(\nabla_x \tilde{f}(x,z) + \nabla_x \tilde{g}(x,z)\lambda_{\mathbf{z}}; 0) + \nabla\tilde{g}(x,y)\lambda_{\mathbf{y}}, \tag{19}$$

$$0 \in \rho(\partial_z \tilde{f}(x,z) + \nabla_z \tilde{g}(x,z)\lambda_{\mathbf{z}}), \tag{20}$$

$$\lambda_{\mathbf{z}} \in \mathbb{R}^l_+, \quad \tilde{g}(x,z) \le 0, \quad \langle \lambda_{\mathbf{z}}, \tilde{g}(x,z)\rangle = 0.^{5} \tag{21}$$

Based on this observation, the equivalence of (1) and (15), and also the fact that (17) is equivalent to

$$\tilde{f}(x,y) = \tilde{f}^*(x), \quad \tilde{g}(x,y) \le 0, \quad \langle \lambda_{\mathbf{y}}, \tilde{g}(x,y)\rangle = 0, \tag{22}$$

we are instead interested in a (weak) KKT solution of problem (1) and its inexact counterpart that are defined below.

Definition 2 (KKT solution and ϵ-KKT solution). *The pair (x,y) is said to be a KKT solution of problem (1) if there exists $(z,\rho,\lambda_{\mathbf{y}},\lambda_{\mathbf{z}}) \in \mathbb{R}^m \times \mathbb{R}_+ \times \mathbb{R}^l_+ \times \mathbb{R}^l_+$ such that (19)-(22) hold. In addition, for any $\varepsilon > 0$, (x,y) is said to be an ε-KKT solution of problem (1) if there exists $(z,\rho,\lambda_{\mathbf{y}},\lambda_{\mathbf{z}}) \in \mathbb{R}^m \times \mathbb{R}_+ \times \mathbb{R}^l_+ \times \mathbb{R}^l_+$ such that*

$$\begin{aligned}
&\operatorname{dist}\Big(0, \partial f(x,y) + \rho\partial\tilde{f}(x,y) - \rho(\nabla_x \tilde{f}(x,z) + \nabla_x \tilde{g}(x,z)\lambda_{\mathbf{z}}; 0) + \nabla\tilde{g}(x,y)\lambda_{\mathbf{y}}\Big) \le \varepsilon,\\
&\operatorname{dist}\Big(0, \rho(\partial_z \tilde{f}(x,z) + \nabla_z \tilde{g}(x,z)\lambda_{\mathbf{z}})\Big) \le \varepsilon,\\
&\|[\tilde{g}(x,z)]_+\| \le \varepsilon, \quad |\langle \lambda_{\mathbf{z}}, \tilde{g}(x,z)\rangle| \le \varepsilon,\\
&|\tilde{f}(x,y) - \tilde{f}^*(x)| \le \varepsilon, \quad \|[\tilde{g}(x,y)]_+\| \le \varepsilon, \quad |\langle \lambda_{\mathbf{y}}, \tilde{g}(x,y)\rangle| \le \varepsilon,
\end{aligned}$$

where $\tilde{f}^*$ is defined in (14).

[4] $\mathscr{I}_{\mathbb{R}^l_+}(\cdot)$ denotes the indicator function associated with the set $\mathbb{R}^l_+$.

[5] The relations in (21) are equivalent to $0 \in -\tilde{g}(x,z) + \partial\mathscr{I}_{\mathbb{R}^l_+}(\lambda_{\mathbf{z}})$.

The notions of KKT solution and ϵ-KKT solution were initially introduced in [46, Section 3]. Notably, it was demonstrated in [46, Theorem 2] that under suitable assumptions, an ϵ-KKT solution (x, y) of problem (1), with conditions such as $\tilde{g} = 0$, $f_2 = 0$, $\tilde{f}_2 = 0$, $\tilde{f}_1$ being twice differentiable, and $\tilde{f}_1(x, \cdot)$ being strongly convex, implies that x is an $\mathcal{O}(\epsilon)$-hypergradient-based stationary point of (1).

The notions of KKT solution and ϵ-KKT solution were initially introduced in [46, Section 3]. Notably, it was demonstrated in [46, Theorem 2] that under suitable assumptions, an ϵ-KKT solution (x, y) of problem (1), with conditions such as $\tilde{g} = 0$, $f_2 = 0$, $\tilde{f}_2 = 0$, $\tilde{f}_1$ being twice differentiable, and $\tilde{f}_1(x, \cdot)$ being strongly convex, implies that x is an $\mathcal{O}(\epsilon)$-hypergradient-based stationary point of (1).

We next study iteration and operation complexity for Algorithm 1. To proceed, recall that $\mathcal{X} = \operatorname{dom} f_2$ and $\mathcal{Y} = \operatorname{dom} \tilde{f}_2$. We define

$$\tilde{f}^*_{\rm hi} := \sup\{\tilde{f}^*(x) | x \in \mathcal{X}\}, \tag{23}$$

$$D_{\mathbf{x}} := \max\{\|u - v\| \big| u, v \in \mathcal{X}\}, \quad D_{\mathbf{y}} := \max\{\|u - v\| \big| u, v \in \mathcal{Y}\}, \tag{24}$$

$$f_{\rm hi} := \max\{f(x, y) | (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \quad f_{\rm low} := \min\{f(x, y) | (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \tag{25}$$

$$\tilde{f}_{\rm low} := \min\{\tilde{f}(x, z) | (x, z) \in \mathcal{X} \times \mathcal{Y}\}, \quad \tilde{g}_{\rm hi} := \max\{\|\tilde{g}(x, y)\| \big| (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \tag{26}$$

$$K := \lceil (\log \varepsilon - \log \epsilon_0)/\log \tau \rceil_+, \quad \mathbb{K} := \{0, 1, \ldots, K + 1\}, \quad \mathbb{K} - 1 = \{k - 1 | k \in \mathbb{K}\}. \tag{27}$$

It then follows from Assumption 1(iii) that

$$\|\nabla \tilde{g}(x, y)\| \le L_{\tilde{g}} \qquad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{28}$$

In addition, by Assumption 1 and the compactness of $\mathcal{X}$ and $\mathcal{Y}$, one can observe that $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $f_{\rm hi}$, $f_{\rm low}$, $\tilde{f}_{\rm low}$ and $\tilde{g}_{\rm hi}$ are finite. Moreover, $\tilde{f}^*_{\rm hi}$ is also finite (see Lemma 1(ii) in Section 4).

The following assumption will be used to establish complexity of Algorithm 1.

Assumption 2 (Slater's condition). *There exists $\hat{z}_x \in \mathcal{Y}$ for each $x \in \mathcal{X}$ such that $\tilde{g}_i(x, \hat{z}_x) < 0$ for all $i = 1, 2, \ldots, l$ and $G := \inf\{-\tilde{g}_i(x, \hat{z}_x) | x \in \mathcal{X},\ i = 1, \ldots, l\} > 0$.*[6]

We are now ready to present an *iteration and operation complexity* of Algorithm 1, measured by the amount of evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla \tilde{g}$ and proximal operators of f_2 and $\tilde{f}_2$, for finding an $\mathcal{O}(\varepsilon)$-KKT solution of (1), whose proofs are deferred to Section 4.

Theorem 1 (iteration and operation complexity of Algorithm 1 for problem (1) with $\sigma = 0$). *Suppose that Assumptions 1 and 2 hold with $\sigma = 0$, i.e., $\tilde{f}_1(x, \cdot)$ being convex but not strongly convex for any given $x \in \operatorname{dom} f_2$. Let $\{(x^k, y^k, z^k, \lambda^k)\}_{k \in \mathbb{K}}$ be generated by Algorithm 1, f^*, $\tilde{f}^*_{\rm hi}$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $f_{\rm hi}$, $f_{\rm low}$, $\tilde{f}_{\rm low}$, $\tilde{g}_{\rm hi}$ and K be defined in (1), (23), (24), (25), (26) and (27), L_f, $L_{\tilde{f}}$, $L_{\nabla f_1}$, $L_{\nabla \tilde{f}_1}$, $L_{\tilde{g}}$, $L_{\nabla \tilde{g}}$ and G be given in Assumptions 1 and 2, and ε, ϵ_0, τ, μ_K, ρ_K and λ_0 be given in Algorithm 1. Let*

$$\vartheta = \frac{1}{2}\|\lambda^0\|^2 + \frac{\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}}{1 - \tau^4} + \frac{D_{\mathbf{y}}\epsilon_0}{1 - \tau^3}, \tag{29}$$

$$L = L_{\nabla f_1} + 2L_{\nabla \tilde{f}_1} + 2L^2_{\tilde{g}} + 2\tilde{g}_{\rm hi}L_{\nabla \tilde{g}} + 2\sqrt{2\vartheta}L_{\nabla \tilde{g}}, \quad \widetilde{L} = L_{\nabla \tilde{f}_1} + L^2_{\tilde{g}} + \tilde{g}_{\rm hi}L_{\nabla \tilde{g}} + \sqrt{2\vartheta}L_{\nabla \tilde{g}}, \tag{30}$$

$$\alpha = \min\left\{1, \sqrt{4/(D_{\mathbf{y}}L)}\right\}, \quad \delta = (2 + \alpha^{-1})L(D^2_{\mathbf{x}} + D^2_{\mathbf{y}}) + \max\{1/D_{\mathbf{y}}, L/4\}D^2_{\mathbf{y}}, \tag{31}$$

$$\begin{aligned} M = {} & 16 \max\left\{1/(4L^2_{\tilde{g}}), 2/(\alpha L^2_{\tilde{g}})\right\} \left[(3L + 1/(2D_{\mathbf{y}}))^2/\min\{2L^2_{\tilde{g}}, 1/(2D_{\mathbf{y}})\} + 3L + 1/(2D_{\mathbf{y}})\right]^2 \\ & \times \left(\delta + 2\alpha^{-1}\left(f^* - f_{\rm low} + \tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low} + L_{\tilde{f}}D_{\mathbf{y}} + 3\vartheta + \tilde{g}^2_{\rm hi} + D_{\mathbf{y}}/4 + L(D^2_{\mathbf{x}} + D^2_{\mathbf{y}})\right)\right), \end{aligned} \tag{32}$$

$$T = \left\lceil 16\left(f_{\rm hi} - f_{\rm low} + 1 + D_{\mathbf{y}}/4\right)L + 8(1 + 4D^2_{\mathbf{y}}L^2)\right\rceil_+, \tag{33}$$

$$\lambda^{K+1}_{\mathbf{y}} = [\lambda^K + \mu_K\tilde{g}(x^{K+1}, y^{K+1})]_+, \quad \lambda^{K+1}_{\mathbf{z}} = \rho_K^{-1}[\lambda^K + \mu_K\tilde{g}(x^{K+1}, z^{K+1})]_+. \tag{34}$$

Suppose that $\varepsilon^{-2} - 8\tau^{-3}G^{-2}\vartheta \ge 0$. Then the following statements hold.

(i) Algorithm 1 terminates after $K+1$ outer iterations and outputs an approximate point (x^{K+1}, y^{K+1})

[6] If Assumption 2 fails to hold, one may instead consider the perturbed counterpart of (1) with $\tilde{g}(x, z)$ replaced by $\tilde{g}(x, z) - \epsilon$ for some suitable $\epsilon > 0$, which clearly satisfies Assumption 2.

of (1) *satisfying*

$$\begin{aligned}
&\operatorname{dist}\Big(0, \partial f(x^{K+1}, y^{K+1}) + \rho_K \partial \tilde{f}(x^{K+1}, y^{K+1}) - \rho_K(\nabla_x \tilde{f}(x^{K+1}, z^{K+1}) + \nabla_x \tilde{g}(x^{K+1}, z^{K+1})\lambda_{\mathbf{z}}^{K+1}; 0) \\
&\qquad + \nabla \tilde{g}(x^{K+1}, y^{K+1})\lambda_{\mathbf{y}}^{K+1}\Big) \le \varepsilon, \qquad (35)\\
&\operatorname{dist}\Big(0, \rho_K(\partial_z \tilde{f}(x^{K+1}, z^{K+1}) + \nabla_z \tilde{g}(x^{K+1}, z^{K+1})\lambda_{\mathbf{z}}^{K+1})\Big) \le \varepsilon, \qquad (36)\\
&\|[\tilde{g}(x^{K+1}, z^{K+1})]_+\| \le 2\varepsilon^2 G^{-1}(\epsilon_0 + L_{\tilde{f}}) D_{\mathbf{y}}, \qquad (37)\\
&|\langle \lambda_{\mathbf{z}}^{K+1}, \tilde{g}(x^{K+1}, z^{K+1})\rangle| \le 2\varepsilon^2 G^{-1}(\epsilon_0 + L_{\tilde{f}}) D_{\mathbf{y}} \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + L_{\tilde{f}})D_{\mathbf{y}}\}, \qquad (38)\\
&\|[\tilde{g}(x^{K+1}, y^{K+1})]_+\| \le 2\varepsilon^2 G^{-1}(\epsilon_0 + L_f + L_{\tilde{f}}) D_{\mathbf{y}}, \qquad (39)\\
&|\langle \lambda_{\mathbf{y}}^{K+1}, \tilde{g}(x^{K+1}, z^{K+1})\rangle| \le 2\varepsilon G^{-1}(\epsilon_0 + L_f + L_{\tilde{f}}) D_{\mathbf{y}} \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + L_f + L_{\tilde{f}})D_{\mathbf{y}}\}, \qquad (40)\\
&|\tilde{f}(x^{K+1}, y^{K+1}) - \tilde{f}^*(x^{K+1})| \le \max\Big\{2\varepsilon^2 G^{-2} L_{\tilde{f}}(\epsilon_0 + L_f + L_{\tilde{f}}) D_{\mathbf{y}}^2,\ \varepsilon^3 \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + L_{\tilde{f}})D_{\mathbf{y}}\}/2 \\
&\qquad\qquad + \varepsilon\left(f_{\rm hi} - f_{\rm low} + 1 + D_{\mathbf{y}}/4 + L_{\tilde{g}}^{-2}/4 + 2D_{\mathbf{y}}^2 L\right)\Big\}. \qquad (41)
\end{aligned}$$

(ii) The total number of evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla \tilde{g}$ *and proximal operators of* f_2 *and* $\tilde{f}_2$ *performed in Algorithm 1 is no more than* N, *respectively, where*

$$\begin{aligned}
N = &\left(\left\lceil 96\sqrt{2}\left(1 + (12L + 2/D_y)/L_{\tilde{g}}^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_y L}\right\} T(1-\tau^7)^{-1} \\
&\times (\tau\varepsilon)^{-7}\left(56K\log(1/\tau) + 56\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&+ (\tau\varepsilon)^{-3/2}(1-\tau^{3/2})^{-1} D_{\mathbf{y}}\sqrt{2\widetilde{L}} + K. \qquad (42)
\end{aligned}$$

Remark 2. *One can observe from Theorem 1 that Algorithm 1 enjoys an iteration complexity of* $\mathcal{O}(\log \varepsilon^{-1})$ *and an operation complexity of* $\mathcal{O}(\varepsilon^{-7}\log\varepsilon^{-1})$, *measured by the amount of evaluations of* ∇f_1, $\nabla \tilde{f}_1$, $\nabla \tilde{g}$ *and proximal operators of* f_2 *and* $\tilde{f}_2$, *for finding an* $\mathcal{O}(\varepsilon)$*-KKT solution* (x^{K+1}, y^{K+1}) *of* (1) *satisfying*

$$\begin{aligned}
&\operatorname{dist}\Big(0, \partial f(x^{K+1}, y^{K+1}) + \rho_K \partial \tilde{f}(x^{K+1}, y^{K+1}) + \nabla \tilde{g}(x^{K+1}, y^{K+1})\lambda_{\mathbf{y}}^{K+1} \\
&\qquad - \rho_K(\nabla_x \tilde{f}(x^{K+1}, z^{K+1}) + \nabla_x \tilde{g}(x^{K+1}, z^{K+1})\tilde{\lambda}_{\mathbf{z}}^{K+1}; 0)\Big) \le \varepsilon, \qquad (43)\\
&\operatorname{dist}\Big(0, \rho_K(\partial_z \tilde{f}(x^{K+1}, z^{K+1}) + \nabla_z \tilde{g}(x^{K+1}, z^{K+1})\lambda_{\mathbf{z}}^{K+1})\Big) \le \varepsilon, \qquad (44)\\
&\|[\tilde{g}(x^{K+1}, z^{K+1})]_+\| = \mathcal{O}(\varepsilon^2), \quad |\langle \lambda_{\mathbf{z}}^{K+1}, \tilde{g}(x^{K+1}, z^{K+1})\rangle| = \mathcal{O}(\varepsilon^2), \qquad (45)\\
&\|[\tilde{g}(x^{K+1}, y^{K+1})]_+\| = \mathcal{O}(\varepsilon^2), \quad |\langle \lambda_{\mathbf{y}}^{K+1}, \tilde{g}(x^{K+1}, z^{K+1})\rangle| = \mathcal{O}(\varepsilon), \qquad (46)\\
&|\tilde{f}(x^{K+1}, y^{K+1}) - \tilde{f}^*(x^{K+1})| = \mathcal{O}(\varepsilon), \qquad (47)
\end{aligned}$$

where $\tilde{f}^*$ *is defined in* (14), $\rho_K = (\epsilon_0\tau^K)^{-1}$, *and* $\lambda_{\mathbf{y}}^{K+1}, \lambda_{\mathbf{z}}^{K+1} \in \mathbb{R}_+^l$ *are given in* (34).

Theorem 2 (iteration and operation complexity of Algorithm 1 for problem (1) with $\sigma > 0$). *Suppose that Assumptions 1 and 2 hold with* $\sigma > 0$, *i.e.,* $\tilde{f}_1(x, \cdot)$ *being strongly convex with parameter* σ *for any given* $x \in \operatorname{dom} f_2$. *Let* $\{(x^k, y^k, z^k, \lambda^k)\}_{k\in\mathbb{K}}$ *be generated by Algorithm 1,* f^*, $\tilde{f}_{\rm hi}^*$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $\tilde{f}_{\rm low}$, $f_{\rm low}$, $f_{\rm hi}$, $\tilde{g}_{\rm hi}$, K, ϑ, L, $\widetilde{L}$, $\lambda_{\mathbf{y}}^{K+1}$ *and* $\lambda_{\mathbf{z}}^{K+1}$ *be defined in* (1), (23), (24), (25), (26), (27), (29), (30) *and* (34), σ, L_f, $L_{\tilde{f}}$, $L_{\nabla f_1}$, $L_{\nabla \tilde{f}_1}$, $L_{\tilde{g}}$, $L_{\nabla\tilde{g}}$ *and* G *be given in Assumptions 1 and 2, and* ε, ϵ_0, τ, μ_K, ρ_K *and* λ_0 *be given in Algorithm 1. Let*

$$\begin{aligned}
&\tilde{\alpha} = \min\left\{1, \sqrt{8\sigma/L}\right\}, \quad \tilde{\delta} = (2 + \tilde{\alpha}^{-1})(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)L + \max\{1/D_{\mathbf{y}}, L/4\}D_{\mathbf{y}}^2, \qquad (48)\\
&\widetilde{M} = 16\max\left\{1/(4L_{\tilde{g}}^2), 2/(\tilde{\alpha}L_{\tilde{g}}^2)\right\}\left[9L^2/\min\{2L_{\tilde{g}}^2, \sigma\} + 3L\right]^2 \\
&\qquad \times\left(\tilde{\delta} + 2\tilde{\alpha}^{-1}\Big(f^* - f_{\rm low} + \tilde{f}_{\rm hi}^* - \tilde{f}_{\rm low} + L_{\tilde{f}}D_{\mathbf{y}} + 3\vartheta + \tilde{g}_{\rm hi}^2 + L(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\Big)\right), \qquad (49)\\
&\widetilde{T} = \left\lceil 16\left(f_{\rm hi} - f_{\rm low} + 1\right)L + 8(1 + \sigma^{-2}L^2)\right\rceil_+. \qquad (50)
\end{aligned}$$

Suppose that $\varepsilon^{-2} - 8\tau^{-3}G^{-2}\vartheta \ge 0$. *Then the following statements hold.*

(i) Algorithm 1 terminates after $K+1$ outer iterations and outputs an approximate point (x^{K+1}, y^{K+1}) of problem (1) satisfying (35)-(40) and

$$|\tilde{f}(x^{K+1}, y^{K+1}) - \tilde{f}^*(x^{K+1})| \le \max\Big\{2\varepsilon^2 G^{-2} L_{\tilde{f}}(\epsilon_0 + L_f + L_{\tilde{f}}) D_{\mathbf{y}}^2,\ \varepsilon^3 \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + L_{\tilde{f}}) D_{\mathbf{y}}\}/2 + \varepsilon\left(f_{\rm hi} - f_{\rm low} + 1 + L_{\tilde{g}}^{-2}/4 + \sigma^{-2}L/2\right)\Big\}. \tag{51}$$

(ii) The total number of evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla \tilde{g}$ and proximal operators of f_2 and $\tilde{f}_2$ performed in Algorithm 1 is no more than $\widetilde{N}$, respectively, where

$$\begin{aligned}\widetilde{N} =& 3397 \max\Big\{2, \sqrt{L/(2\sigma)}\Big\}\widetilde{T}(1-\tau^6)^{-1} \\ &\times (\tau\varepsilon)^{-6}\left(38K\log(1/\tau) + 38\log(1/\epsilon_0) + 2(\log\widetilde{M})_+ + 2 + 2\log(2\widetilde{T})\right) \\ &+ 2(\tau\varepsilon)^{-1}(1-\tau)\left\lceil\sqrt{\frac{\widetilde{L}}{\sigma}} + 1\right\rceil \max\Big\{1, \Big\lceil 2\log(2\widetilde{L}D_{\mathbf{y}}^2) + 6K\log(1/\tau) - 6\log\epsilon_0\Big\rceil\Big\} + K. \end{aligned} \tag{52}$$

Remark 3. *One can observe from Theorem 2 that Algorithm 1 enjoys an iteration complexity of $\mathcal{O}(\log\varepsilon^{-1})$ and an operation complexity of $\mathcal{O}(\varepsilon^{-6}\log\varepsilon^{-1})$, measured by the amount of evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla \tilde{g}$ and proximal operators of f_2 and $\tilde{f}_2$, for finding an $\mathcal{O}(\varepsilon)$-KKT solution (x^{K+1}, y^{K+1}) of (1) satisfying (43)-(47). Such an operation complexity significantly improves the previous best-known operation complexity $\mathcal{O}(\varepsilon^{-7}\log\varepsilon^{-1})$ established in [46, Theorem 5] by a factor of ε^{-1}.*

3 Numerical results

In this section, we conduct some preliminary experiments to test the performance of our SMO method (Algorithm 1), and compare it with a first-order penalty (FOP) method ([46, Algorithm 4]). Both algorithms are coded in Matlab and all the computations are performed on a laptop with a 2.30 GHz Intel i9-9880H 8-core processor and 16 GB of RAM.

3.1 Constrained bilevel linear optimization

In this subsection, we consider constrained bilevel linear optimization in the form of

$$\begin{aligned} \min\quad & c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x) \\ \text{s.t.}\quad & y \in \operatorname*{argmin}_z \left\{\tilde{d}^T z + \mathscr{I}_{[-1,1]^m}(z) \,\middle|\, \widetilde{A}x + \widetilde{B}z - \tilde{b} \le 0\right\}, \end{aligned} \tag{53}$$

where $c \in \mathbb{R}^n$, $d, \tilde{d} \in \mathbb{R}^m$, $\tilde{b} \in \mathbb{R}^l$, $\widetilde{A} \in \mathbb{R}^{l\times n}$, $\widetilde{B} \in \mathbb{R}^{l\times m}$, and $\mathscr{I}_{[-1,1]^n}(\cdot)$ and $\mathscr{I}_{[-1,1]^m}(\cdot)$ are the indicator functions of $[-1,1]^n$ and $[-1,1]^m$ respectively.

For each triple (n, m, l), we randomly generate 10 instances of problem (53). Specifically, we first randomly generate c and d with all the entries independently chosen from the standard normal distribution. We then randomly generate $\widetilde{A}$ and $\widetilde{B}$ with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.01. In addition, we randomly generate $\hat{y} \in [-1,1]^m$ with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.1 and then projected to $[-1,1]^m$ and choose $\tilde{d}$ and $\tilde{b}$ such that $\hat{y}$ is an optimal solution of the lower-level optimization of (53) with $x = 0$.

Notice that (53) is a special case of (1) with $f(x,y) = c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x)$, $\tilde{f}(x,z) = \tilde{d}^T z + \mathscr{I}_{[-1,1]^m}(z)$ and $\tilde{g}(x,z) = \widetilde{A}x + \widetilde{B}z - \tilde{b}$. We now apply SMO and FOP to solve (53). In particular, we choose 0 as the initial point for both methods. In addition, we set $(\varepsilon, \epsilon_0, \tau) = (10^{-2}, 1, 0.8)$ for SMO. To enhance the efficiency of FOP, we adopt a dynamic updating scheme on its penalty and tolerance parameters. Specifically, we set $\rho_k = 5^{k-1}$, $\varepsilon_k = \rho_k^{-1}$ and $x_{-1} = 0$ for [46, Algorithm 2]. For each $k > 1$, let (x^{k-1}, y^{k-1}) be the output of [46, Algorithm 2] with $(\varepsilon, \rho) = (\varepsilon_{k-1}, \rho_{k-1})$. We run [46, Algorithm 2] with $(\varepsilon, \rho) = (\varepsilon_k, \rho_k)$ and $(x^{k-1}, \tilde{y}^{k-1})$ as the initial point to generate (x^k, y^k), where $\tilde{y}^{k-1} \in \operatorname{argmin}_z \tilde{f}(x^{k-1}, z)$ is found by CVX [20]. We terminate both algorithms once $\epsilon_k \le 10^{-2}$ for SMO, $\varepsilon_k \le 10^{-2}$ for FOP, and (x^k, y^k) satisfies

$$\|[\tilde{g}(x^k, y^k)]_+\| \le 10^{-2}, \quad \tilde{f}(x^k, y^k) - \tilde{f}^*(x^k) \le 10^{-2}$$

for some k, and output (x^k, y^k) as an approximate solution of(53), where $\tilde{f}^*$ is defined in (14) and the value $\tilde{f}^*(x^k)$ is computed by CVX [20].

The computational results of SMO and FOP for problem (53) with the instances randomly generated above are presented in Table 1. In detail, the values of n, m and l are listed in the first three columns. For each triple (n, m, l), the average initial objective value $f(x^0, \hat{y})$ with $\hat{y}$ being generated above,[7] the average final objective value $f(x^k, y^k)$ and the average CPU time (in seconds) over 10 random instances are given in the rest of the columns. One can observe that both SMO and FOP find an approximate solution with much lower objective value than the initial objective value. Moreover, SMO outputs an approximate solution with a similar objective value as FOP, while SMO significantly outperforms FOP in terms of average CPU time.

			Initial objective value	Final objective value		CPU time (seconds)	
n	m	l		SMO	FOP	SMO	FOP
100	100	5	0.22	−75.78	−75.75	7.4	22.1
200	200	10	−0.38	−154.20	−154.18	12.5	107.9
300	300	15	−0.11	−246.70	−246.66	24.1	267.8
400	400	20	0.34	−305.97	−305.89	37.0	561.6
500	500	25	0.96	−394.74	−394.70	63.8	719.8

Table 1: Numerical results for problem (53)

3.2 Constrained bilevel optimization with quadratic upper level and linear lower level

In this subsection, we consider constrained bilevel optimization with quadratic upper level and linear lower level in the form of

$$\begin{array}{ll} \min & x^T Ax + x^T By + y^T Cy + c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x) \\ \text{s.t.} & y \in \underset{z}{\operatorname{argmin}} \left\{ \tilde{d}^T z + \mathscr{I}_{[-1,1]^m}(z) \big| \widetilde{A}x + \widetilde{B}z - \tilde{b} \le 0 \right\}, \end{array} \tag{54}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{m\times m}$, $c \in \mathbb{R}^n$, $d, \tilde{d} \in \mathbb{R}^m$, $\tilde{b} \in \mathbb{R}^l$, $\widetilde{A} \in \mathbb{R}^{l\times n}$, $\widetilde{B} \in \mathbb{R}^{l\times m}$, and $\mathscr{I}_{[-1,1]^n}(\cdot)$ and $\mathscr{I}_{[-1,1]^m}(\cdot)$ are the indicator functions of $[-1,1]^n$ and $[-1,1]^m$ respectively.

For each triple (n, m, l), we randomly generate 10 instances of problem (54). Specifically, we first randomly generate A, B, C, c and d with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.1. We then randomly generate $\widetilde{A}$ and $\widetilde{B}$ with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.01. In addition, we randomly generate $\hat{y} \in [-1,1]^m$ with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.1 and then projected to $[-1,1]^m$ and choose $\tilde{d}$ and $\tilde{b}$ such that $\hat{y}$ is an optimal solution of the lower-level optimization of (54) with $x = 0$.

Notice that (54) is a special case of (1) with $f(x,y) = x^T Ax + x^T By + y^T Cy + c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x)$, $\tilde{f}(x,z) = \tilde{d}^T z + \mathscr{I}_{[-1,1]^m}(z)$ and $\tilde{g}(x,z) = \widetilde{A}x + \widetilde{B}z - \tilde{b}$. We now apply our SMO method (Algorithm 1) to solve (54).[8] Specifically, we choose 0 as the initial point and set the parameters of SMO as $(\varepsilon, \epsilon_0, \tau) = (10^{-2}, 1, 0.8)$. We terminate SMO once $\epsilon_k \le 10^{-2}$ and (x^k, y^k) satisfies

$$\|[\tilde{g}(x^k, y^k)]_+\| \le 10^{-2}, \quad \tilde{f}(x^k, y^k) - \tilde{f}^*(x^k) \le 10^{-2}$$

for some k and output (x^k, y^k) as an approximate solution of (54), where $\tilde{f}^*$ is defined in (14) and the value $\tilde{f}^*(x^k)$ is computed by CVX [20].

The computational results of SMO for problem (54) with the instances randomly generated above are presented in Table 2. In detail, the values of n, m and l are listed in the first three columns. For each triple (n, m, l), the average initial objective value $f(x^0, \hat{y})$ with $\hat{y}$ being generated above and the average final objective value $f(x^k, y^k)$ over 10 random instances are given in the rest of the columns. One can see that the approximate solution (x^k, y^k) found by SMO significantly reduces the initial objective value.

[7] Note that (x^0, y^0_{init}) may not be a feasible point of (53). Nevertheless, $(x^0, \hat{y})$ is a feasible point of (53) due to $x^0 = 0$ and the particular way for generating instances of (53). Besides, (53) can be viewed as an implicit optimization problem in terms of the variable x. It is thus reasonable to use $f(x^0, \hat{y})$ as the initial objective value for the purpose of comparison.

[8] Clearly, problem (54) is more sophisticated than (53). As shown in Table 1, problem (53) already poses significant challenges for FOP [46] when the dimension n is relatively large. Therefore, we do not apply FOP to solve (54).

n	m	l	Initial objective value	Final objective value
100	100	5	-0.04	-95.70
200	200	10	0.03	-275.34
300	300	15	0.15	-487.64
400	400	20	0.20	-749.02
500	500	25	0.13	-1085.57

Table 2: Numerical results for problem (54)

3.3 Hyperparameter tuning for support vector machine

In this subsection, we consider a hyperparameter tuning model for support vector machine (SVM):

$$\begin{aligned}\min_{c,w,b,\xi} \ & \frac{1}{m}\sum_{n<i\le n+m} \ell\left(\hat{y}_i, w^T\hat{x}_i + b\right) + \mathscr{I}_{[0,10]^n}(c) \\ \text{s.t. } (w,b,\xi) \in \operatorname*{argmin}_{\tilde{w},\tilde{b},\tilde{\xi}} \Big\{ & \sum_{1\le i\le n} \ell\left(\hat{y}_i, \tilde{w}^T\hat{x}_i + \tilde{b}\right) + c^T\tilde{\xi} + \mathscr{I}_{[-1,1]^{q+1}}(\tilde{w},\tilde{b}) + \mathscr{I}_{[0,20]^n}(\tilde{\xi}) \Big| \\ & \hat{y}_i(\tilde{w}^T\hat{x}_i + \tilde{b}) \ge 1 - \tilde{\xi}_i,\ i = 1,\dots,n\Big\},\end{aligned} \tag{55}$$

where $\{(\hat{x}_i,\hat{y}_i)\}_{1\le i\le n}$ is the training set, $\{(\hat{x}_i,\hat{y}_i)\}_{n<i\le n+m}$ is the validation set, $c\in\mathbb{R}^n$, $w\in\mathbb{R}^q$, $b\in\mathbb{R}$, $\xi\in\mathbb{R}^n$, $\ell(u,v)=\log(1+e^{-uv})$ is the binomial deviance loss function [24], and $\mathscr{I}_{[0,10]^n}$, $\mathscr{I}_{[-1,1]^{q+1}}$, and $\mathscr{I}_{[0,20]^n}$ are the indicator functions of $[0,10]^n$, $[-1,1]^{q+1}$ and $[0,20]^n$, respectively. Specifically, at the lower level of (55), we train a linear SVM on the training set with a decision hyperplane of the form $\{x\in\mathbb{R}^q : w^Tx+b=0\}$, where the binomial deviance is used as the loss function, and a slack variable ξ and penalty parameter c are introduced to handle non-separable datasets.[9] At the upper level of (55), we minimize the validation loss to select the hyperparameter c and the corresponding (w,b,ξ). Similar bilevel SVM models have been widely studied in the literature (e.g., [3, 54, 76]).

In our experiments, we solve problem (55) on five datasets from the LIBSVM repository [5]. For each dataset, one-fourth of the samples are randomly selected as the validation set, and the remainder are used as the training set. Notice that (55) is a special case of (1) with $x=c$, $y=(w,b,\xi)$, $z=(\tilde{w},\tilde{b},\tilde{\xi})$,

$$\begin{aligned} f(x,y) &= \frac{1}{m}\sum_{n<i\le n+m} \ell\left(\hat{y}_i, w^T\hat{x}_i + b\right) + \mathscr{I}_{[0,10]^n}(c), \\ \tilde{f}(x,z) &= \sum_{1\le i\le n} \ell\left(\hat{y}_i, w^T\hat{x}_i + b\right) + c^T\tilde{\xi} + \mathscr{I}_{[-1,1]^{q+1}}(\tilde{w},\tilde{b}) + \mathscr{I}_{[0,20]^n}(\tilde{\xi}), \\ \tilde{g}_i(x,z) &= 1 - \tilde{\xi}_i - \hat{y}_i(\tilde{w}^T\hat{x}_i + \tilde{b}), i = 1\dots,n. \end{aligned}$$

As a result, both SMO (Algorithm 1) and FOP ([46, Algorithm 2]) are suitable for solving (55). We now apply both methods to solve (55), starting from $x^0=0$ and y^0 with the entries independently drawn from the uniform distribution on $[0,1]$. In addition, we set $(\varepsilon,\epsilon_0,\tau)=(10^{-2},1,0.9)$ for SMO. To enhance the efficiency of FOP, we adopt a dynamic updating scheme on its penalty and tolerance parameters. Specifically, we set $\rho_k=5^{k-1}$, $\varepsilon_k=\rho_k^{-1}$ for [46, Algorithm 2]. For each $k>1$, let (x^{k-1},y^{k-1}) be the output of [46, Algorithm 2] with $(\varepsilon,\rho)=(\varepsilon_{k-1},\rho_{k-1})$. We run [46, Algorithm 2] with $(\varepsilon,\rho)=(\varepsilon_k,\rho_k)$ and $(x^{k-1},\tilde{y}^{k-1})$ as the initial point to generate (x^k,y^k), where $\tilde{y}^{k-1}\in\operatorname{argmin}_z \tilde{f}(x^{k-1},z)$ is found by CVX [20]. We terminate both algorithms once $\epsilon_k\le 10^{-2}$ for SMO, $\varepsilon_k\le 10^{-2}$ for FOP, and (x^k,y^k) satisfies

$$\|[\tilde{g}(x^k,y^k)]_+\|\le 10^{-2},\quad \tilde{f}(x^k,y^k)-\tilde{f}^*(x^k)\le 10^{-2}$$

for some k, and output (x^k,y^k) as an approximate solution of (55), where $\tilde{f}^*$ is defined in (14) and the value $\tilde{f}^*(x^k)$ is computed by CVX [20].

The computational results of SMO and FOP for problem (55) are presented in Table 3. In detail, the names of the datasets are listed in the first column. For each dataset, the initial objective value $f(x^0,y_0^*)$, where y_0^* is the lower-level optimal solution with $x=x^0$ computed by CVX [20], the final objective value $f(x^k,y^k)$, validation accuracy, and the CPU time (in seconds) are given in the rest of the columns. It can be observed that both SMO and FOP find approximate solutions with objective values substantially

[9] The vector c is introduced to assign weights to individual data points in the weighted SVM formulation. This technique has been studied in the literature to capture the relative importance of data points in the training set (see [59, 70, 72]).

lower than the initial value, and the resulting support vector machine achieves good validation accuracy. Moreover, SMO produces an approximate solution with an objective value similar to that of FOP but significantly outperforms FOP in terms of CPU time.

Dataset	Initial objective value	Final objective value		Validation accuracy		CPU time (seconds)	
		SMO	FOP	SMO	FOP	SMO	FOP
breast-cancer_scale	2.894	0.336	0.352	100.0%	98.5%	175.8	1537.9
heart_scale	1.359	0.556	0.551	75.9%	72.4%	402.6	3335.5
ionosphere_scale	1.063	0.443	0.447	88.9%	86.7%	792.5	7520.7
german.numer_scale	1.609	0.544	0.562	87.7%	84.2%	278.6	2453.0
australian_scale	2.565	0.654	0.678	72.5%	72.5%	224.6	2138.0

Table 3: Numerical results for problem (55)

4 Proof of main results

In this section, we provide a proof of our main results presented in Subsection 2.1, which are particularly Theorems 1 and 2.

It should be noted that while the first-order penalty method (FOP) ([46, Algorithm 4]) and the SMO method (Algorithm 1) both require solving minimax subproblems, they differ substantially in several aspects: (i) FOP solves a single minimax subproblem with fixed penalty parameters and without a warm-start strategy, whereas SMO solves a sequence of minimax subproblems with dynamically updated penalty parameters and a tailored warm-start strategy; (ii) The minimax subproblem in FOP arises from a quadratic penalty scheme applied to the lower-level problem, while in SMO it is derived from a modified augmented Lagrangian scheme; (iii) FOP is designed for problems with a merely convex lower-level objective and cannot exploit strong convexity, whereas SMO can leverage it to achieve stronger complexity results. As a result, the convergence proofs for FOP and SMO are fundamentally different. Moreover, due to SMO's more intricate structure, establishing its convergence is significantly more challenging.

To proceed, one can observe from (9) and (14) that

$$\min_z \widetilde{\mathcal{L}}(x,z,\lambda;\rho,\mu) \le \tilde{f}^*(x) + \frac{\|\lambda\|^2}{2\rho\mu} \qquad \forall x\in\mathcal{X}, \lambda\in\mathbb{R}^l_+, \rho,\mu>0, \tag{56}$$

which will be frequently used later.

We now introduce several technical lemmas that will be utilized to prove Theorems 1 and 2 subsequently. The following lemma presents several properties of the lower-level problem of (1), whose proof can be found in [46].

Lemma 1 ([46, Lemma 3]). *Suppose that Assumptions 1 and 2 hold. Let $\tilde{f}^*$, $\tilde{f}^*_{\rm hi}$, $D_{\mathbf{y}}$, $L_{\tilde{f}}$ and G be given in* (14), (23), (24)*, and Assumptions 1 and 2, respectively. Then the following statements hold.*

(i) $\lambda^*\ge 0$ and $\|\lambda^*\|\le G^{-1}L_{\tilde{f}}D_{\mathbf{y}}$ for all $\lambda^*\in\Lambda^*(x)$ and $x\in\mathcal{X}$, where $\Lambda^*(x)$ denotes the set of optimal Lagrangian multipliers of problem (14) *for any $x\in\mathcal{X}$.*

(ii) The function $\tilde{f}^*$ is Lipschitz continuous on $\mathcal{X}$ and $\tilde{f}^*_{\rm hi}$ is finite.

(iii) It holds that

$$\tilde{f}^*(x) = \max_\lambda \min_z \tilde{f}(x,z) + \langle\lambda, \tilde{g}(x,z)\rangle - \mathscr{I}_{\mathbb{R}^l_+}(\lambda) \qquad \forall x\in\mathcal{X},$$

where $\mathscr{I}_{\mathbb{R}^l_+}(\cdot)$ is the indicator function associated with $\mathbb{R}^l_+$.

The next lemma provides an upper bound on $\|\lambda^k\|$ for all $0\le k\in\mathbb{K}-1$.

Lemma 2. *Suppose that Assumption 1 holds. Let $\mathbb{K}$ and ϑ be defined in* (27) *and* (29)*, μ_k and ρ_k be given in Algorithm 1, and $\{\lambda^k\}_{k\in\mathbb{K}}$ be generated by Algorithm 1. Then we have*

$$\|\lambda^k\|^2 \le 2\rho_k\mu_k\vartheta \qquad \forall 0\le k\in\mathbb{K}-1. \tag{57}$$

Proof. One can observe from (23), (26) and Algorithm 1 that $\tilde{f}^*_{\rm hi}\ge\tilde{f}_{\rm low}$ and $\mu_0\ge\rho_0\ge 1>\tau>0$, which together with (29) imply that (57) holds for $k=0$. It remains to show that (57) holds for all $1\le k\in\mathbb{K}-1$.

Since $(x^{t+1}, y^{t+1}, z^{t+1})$ is an ϵ_t-primal-dual stationary point of (12) for all $0 \le t \in \mathbb{K}-1$, it follows from Definition 1 that there exists some $u \in \partial_z \mathcal{L}(x^{t+1}, y^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t)$ with $\|u\| \le \epsilon_t$. Notice from (8) and (9) that $\partial_z \mathcal{L}(x^{t+1}, y^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t) = -\rho_t \partial_z \widetilde{\mathcal{L}}(x^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t)$. Hence, $-\rho_t^{-1}u \in \partial_z \widetilde{\mathcal{L}}(x^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t)$. Also, observe from (9) and Assumption 1 that $\widetilde{\mathcal{L}}(x^{t+1}, \cdot, \lambda^t; \rho_t, \mu_t)$ is convex. Using this, (24), $-\rho_t^{-1}u \in \partial_z \widetilde{\mathcal{L}}(x^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t)$ and $\|u\| \le \epsilon_t$, we obtain

$$\begin{aligned}\widetilde{\mathcal{L}}(x^{t+1}, z, \lambda^t; \rho_t, \mu_t) &\ge \widetilde{\mathcal{L}}(x^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t) + \langle -\rho_t^{-1}u, z - z^{t+1}\rangle. \\ &\ge \widetilde{\mathcal{L}}(x^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t) - \rho_t^{-1} D_{\mathbf{y}} \epsilon_t \qquad \forall z \in \mathcal{Y},\end{aligned}$$

which implies that

$$\min_z \widetilde{\mathcal{L}}(x^{t+1}, z, \lambda^t; \rho_t, \mu_t) \ge \widetilde{\mathcal{L}}(x^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t) - \rho_t^{-1} D_{\mathbf{y}} \epsilon_t. \tag{58}$$

By this, (9) and (56), one has

$$\begin{aligned}\tilde{f}^*(x^{t+1}) &\overset{(56)}{\ge} \min_z \widetilde{\mathcal{L}}(x^{t+1}, z, \lambda^t; \rho_t, \mu_t) - \frac{\|\lambda^t\|^2}{2\rho_t\mu_t} \\ &\overset{(9)(58)}{\ge} \tilde{f}(x^{t+1}, z^{t+1}) + \frac{1}{2\rho_t\mu_t}\left(\|[\lambda^t + \mu_t \tilde{g}(x^{t+1}, z^{t+1})]_+\|^2 - \|\lambda^t\|^2\right) - \rho_t^{-1} D_{\mathbf{y}} \epsilon_t \\ &= \tilde{f}(x^{t+1}, z^{t+1}) + \frac{1}{2\rho_t\mu_t}\left(\|\lambda^{t+1}\|^2 - \|\lambda^t\|^2\right) - \rho_t^{-1} D_{\mathbf{y}} \epsilon_t,\end{aligned}$$

where the equality follows from the relation $\lambda^{t+1} = [\lambda^t + \mu_t \tilde{g}(x^{t+1}, z^{t+1})]_+$ (see Algorithm 1). Using this inequality, (23), (26) and $\epsilon_t \le \epsilon_0$ (see Algorithm 1), we have

$$\|\lambda^{t+1}\|^2 - \|\lambda^t\|^2 \le 2\rho_t\mu_t(\tilde{f}^*(x^{t+1}) - \tilde{f}(x^{t+1}, y^{t+1})) + 2\mu_t D_{\mathbf{y}} \epsilon_t \le 2\rho_t\mu_t(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}) + 2\mu_t D_{\mathbf{y}} \epsilon_0.$$

Summing up this inequality for $t = 0, \ldots, k-1$ with $1 \le k \in \mathbb{K}-1$ yields

$$\|\lambda^k\|^2 \le \|\lambda^0\|^2 + 2(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}) \sum_{t=0}^{k-1} \rho_t\mu_t + 2D_{\mathbf{y}}\epsilon_0 \sum_{t=0}^{k-1} \mu_t. \tag{59}$$

Recall from Algorithm 1 that $\epsilon_t = \epsilon_0\tau^t$, $\mu_t = \epsilon_t^{-3}$ and $\rho_t = \epsilon_t^{-1}$. It is not hard to verify that $\sum_{t=0}^{k-1} \rho_t\mu_t \le \rho_{k-1}\mu_{k-1}/(1-\tau^4)$ and $\sum_{t=0}^{k-1} \mu_t \le \mu_{k-1}/(1-\tau^3)$. Using these, (59), $\rho_k > \rho_{k-1} \ge 1$ and $\mu_k > \mu_{k-1} \ge 1$ (see Algorithm 1), we obtain that for all $1 \le k \in \mathbb{K}-1$,

$$\begin{aligned}\rho_k^{-1}\mu_k^{-1}\|\lambda^k\|^2 &\le \rho_k^{-1}\mu_k^{-1}\left(\|\lambda^0\|^2 + \frac{2\rho_{k-1}\mu_{k-1}(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})}{1-\tau^4} + \frac{2\mu_{k-1}D_{\mathbf{y}}\epsilon_0}{1-\tau^3}\right) \\ &\le \|\lambda^0\|^2 + \frac{2(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})}{1-\tau^4} + \frac{2D_{\mathbf{y}}\epsilon_0}{1-\tau^3} \overset{(29)}{=} 2\vartheta.\end{aligned}$$

It implies that the conclusion of this lemma holds. □

The following lemma provides an upper bound on $\|[\tilde{g}(x^{k+1}, z^{k+1})]_+\|$ and $\|[\tilde{g}(x^{k+1}, y^{k+1})]_+\|$.

Lemma 3. *Suppose that Assumptions 1 and 2 hold. Let $D_{\mathbf{y}}$, $\mathbb{K}$ and ϑ be defined in (24), (27) and (29), L_f, $L_{\tilde{f}}$ and G be given in Assumptions 1 and 2, and ϵ_0, ρ_k and μ_k be given in Algorithm 1. Suppose that $(x^{k+1}, y^{k+1}, z^{k+1}, \lambda^{k+1})$ is generated by Algorithm 1 for some $0 \le k \in \mathbb{K}-1$ with*

$$\rho_k^{-1}\mu_k \ge 8G^{-2}\vartheta. \tag{60}$$

Then we have

$$\|[\tilde{g}(x^{k+1}, z^{k+1})]_+\| \le \mu_k^{-1}\|\lambda^{k+1}\| \le 2\mu_k^{-1}G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}, \tag{61}$$

$$\|[\tilde{g}(x^{k+1}, y^{k+1})]_+\| \le \mu_k^{-1}\|[\lambda^k + \mu_k \tilde{g}(x^{k+1}, y^{k+1})]_+\| \le 2\mu_k^{-1}G^{-1}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})D_{\mathbf{y}}. \tag{62}$$

Proof. Suppose that $(x^{k+1}, y^{k+1}, z^{k+1}, \lambda^{k+1})$ is generated by Algorithm 1 for some $0 \le k \in \mathbb{K} - 1$ satisfying (60). Notice that $(x^{k+1}, y^{k+1}, z^{k+1})$ is an ϵ_k-primal-dual stationary point of (12). It then follows from (8), Definition 1 and Assumption 1 that

$$\operatorname{dist}\Big(0, \nabla_y f(x^{k+1}, y^{k+1}) + \rho_k \partial_y \tilde{f}(x^{k+1}, y^{k+1}) + \nabla_y \tilde{g}(x^{k+1}, y^{k+1})[\lambda^k + \mu_k \tilde{g}(x^{k+1}, y^{k+1})]_+\Big) \le \epsilon_k, \quad (63)$$

$$\operatorname{dist}\Big(0, -\rho_k \partial_z \tilde{f}(x^{k+1}, z^{k+1}) - \nabla_z \tilde{g}(x^{k+1}, z^{k+1})[\lambda^k + \mu_k \tilde{g}(x^{k+1}, z^{k+1})]_+\Big) \le \epsilon_k. \quad (64)$$

We first show that (61) holds. Notice from Algorithm 1 that $\lambda^{k+1} = [\lambda^k + \mu_k \tilde{g}(x^{k+1}, z^{k+1})]_+$. Hence, it follows from (64) that there exists some $u \in \partial_z \tilde{f}(x^{k+1}, z^{k+1})$ such that

$$\|\rho_k u + \nabla_z \tilde{g}(x^{k+1}, z^{k+1})\lambda^{k+1}\| \le \epsilon_k. \quad (65)$$

By Assumption 2, there exists some $\hat{z}^{k+1} \in \mathcal{Y}$ such that $-\tilde{g}_i(x^{k+1}, \hat{z}^{k+1}) \ge G$ for all i. Observe that $\langle \lambda^{k+1}, \lambda^k + \mu_k \tilde{g}(x^{k+1}, z^{k+1})\rangle = \|[\lambda^k + \mu_k \tilde{g}(x^{k+1}, z^{k+1})]_+\|^2 \ge 0$, which implies that

$$-\langle \lambda^{k+1}, \mu_k^{-1}\lambda^k\rangle \le \langle \lambda^{k+1}, \tilde{g}(x^{k+1}, z^{k+1})\rangle. \quad (66)$$

Using these, (65), $\lambda^{k+1} \ge 0$ and $u \in \partial_z \tilde{f}(x^{k+1}, z^{k+1})$, we have

$$\begin{aligned}
&\rho_k \tilde{f}(x^{k+1}, z^{k+1}) - \rho_k \tilde{f}(x^{k+1}, \hat{z}^{k+1}) + G\|\lambda^{k+1}\|_1 - \langle \lambda^{k+1}, \mu_k^{-1}\lambda^k\rangle \\
&\le \rho_k \tilde{f}(x^{k+1}, z^{k+1}) - \rho_k \tilde{f}(x^{k+1}, \hat{z}^{k+1}) + \langle \lambda^{k+1}, -\tilde{g}(x^{k+1}, \hat{z}^{k+1}) - \mu_k^{-1}\lambda^k\rangle \\
&\overset{(66)}{\le} \rho_k \tilde{f}(x^{k+1}, z^{k+1}) - \rho_k \tilde{f}(x^{k+1}, \hat{z}^{k+1}) + \langle \lambda^{k+1}, \tilde{g}(x^{k+1}, z^{k+1}) - \tilde{g}(x^{k+1}, \hat{z}^{k+1}))\rangle \\
&\le \langle \rho_k u, z^{k+1} - \hat{z}^{k+1}\rangle + \langle \nabla_z \tilde{g}(x^{k+1}, z^{k+1})\lambda^{k+1}, z^{k+1} - \hat{z}^{k+1}\rangle \\
&= \langle \rho_k u + \nabla_z \tilde{g}(x^{k+1}, z^{k+1})\lambda^{k+1}, z^{k+1} - \hat{z}^{k+1}\rangle \le D_{\mathbf{y}}\epsilon_k, \quad (67)
\end{aligned}$$

where the first inequality is due to $\lambda^{k+1} \ge 0$ and $-\tilde{g}_i(x^{k+1}, \hat{z}^{k+1}) \ge G$ for all i, the third inequality follows from $u \in \partial_z \tilde{f}(x^{k+1}, z^{k+1})$, $\lambda^{k+1} \ge 0$ and the convexity of $\tilde{f}(x^{k+1}, \cdot)$ and $\tilde{g}_i(x^{k+1}, \cdot)$ for all i, and the last inequality is due to (24), (65) and $z^{k+1}, \hat{z}^{k+1} \in \mathcal{Y}$.

In view of (24), (67), $z^{k+1}, \hat{z}^{k+1} \in \mathcal{Y}$, and the Lipschitz continuity of $\tilde{f}$, one has

$$\begin{aligned}
D_{\mathbf{y}}\epsilon_k + \rho_k L_{\tilde{f}} D_{\mathbf{y}} &\overset{(24)}{\ge} D_{\mathbf{y}}\epsilon_k + \rho_k L_{\tilde{f}}\|z^{k+1} - \hat{z}^{k+1}\| \ge D_{\mathbf{y}}\epsilon_k + \rho_k(\tilde{f}(x^{k+1}, \hat{z}^{k+1}) - \tilde{f}(x^{k+1}, z^{k+1})) \\
&\overset{(67)}{\ge} G\|\lambda^{k+1}\|_1 - \langle \lambda^{k+1}, \mu_k^{-1}\lambda^k\rangle \ge (G - \mu_k^{-1}\|\lambda^k\|)\|\lambda^{k+1}\|, \quad (68)
\end{aligned}$$

where the first inequality is due to (24) and $z^{k+1}, \hat{z}^{k+1} \in \mathcal{Y}$, the second inequality follows from $L_{\tilde{f}}$-Lipschitz continuity of $\tilde{f}$, and the last inequality is due to $\|\lambda^{k+1}\|_1 \ge \|\lambda^{k+1}\|$. In addition, it follows from (57) and (60) that

$$G - \mu_k^{-1}\|\lambda^k\| \overset{(57)}{\ge} G - \sqrt{2\rho_k \mu_k^{-1}\vartheta} \overset{(60)}{\ge} G/2,$$

which together with (68) yields

$$\|\lambda^{k+1}\| \le 2G^{-1}(\epsilon_k + \rho_k L_{\tilde{f}})D_{\mathbf{y}}.$$

The statement (61) then follows from this, $\epsilon_k \le \epsilon_0$, and

$$\|[\tilde{g}(x^{k+1}, z^{k+1})]_+\| \le \mu_k^{-1}\|[\lambda^k + \mu_k \tilde{g}(x^{k+1}, z^{k+1})]_+\| = \mu_k^{-1}\|\lambda^{k+1}\|.$$

We next show that (62) holds. Indeed, let $\tilde{\lambda}^{k+1} = [\lambda^k + \mu_k \tilde{g}(x^{k+1}, y^{k+1})]_+$. It then follows from (63) that

$$\operatorname{dist}\Big(0, \nabla_y f(x^{k+1}, y^{k+1}) + \rho_k \partial_y \tilde{f}(x^{k+1}, y^{k+1}) + \nabla_y \tilde{g}(x^{k+1}, y^{k+1})\tilde{\lambda}^{k+1}\Big) \le \epsilon_k.$$

Hence, there exists some $v \in \rho_k^{-1}\nabla_y f(x^{k+1}, y^{k+1}) + \partial_y \tilde{f}(x^{k+1}, y^{k+1})$ such that

$$\|\rho_k v + \nabla_y \tilde{g}(x^{k+1}, y^{k+1})\tilde{\lambda}^{k+1}\| \le \epsilon_k.$$

The rest of the proof of (62) is similar to the one of (61) with u, z^{k+1} and λ^{k+1} being replaced with v, y^{k+1} and $\tilde{\lambda}^{k+1}$ respectively and thus omitted. □

The next lemma provides an upper bound on the quantities associated with an approximate KKT solution (x^{k+1}, y^{k+1}).

Lemma 4. *Suppose that Assumptions 1 and 2 hold. Let $D_{\mathbf{y}}$, $\mathbb{K}$ and ϑ be defined in (24), (27) and (29), L_f, $L_{\tilde{f}}$ and G be given in Assumptions 1 and 2, and ϵ_0, τ, ρ_k and μ_k be given in Algorithm 1. Suppose that $(x^{k+1}, y^{k+1}, z^{k+1}, \lambda^{k+1})$ is generated by Algorithm 1 for some $0 \le k \in \mathbb{K} - 1$ with*

$$\rho_k^{-1}\mu_k \ge 8\tau^{-2}G^{-2}\vartheta. \tag{69}$$

Let

$$\lambda_{\mathbf{y}}^{k+1} = [\lambda^k + \mu_k \tilde{g}(x^{k+1}, y^{k+1})]_+, \qquad \lambda_{\mathbf{z}}^{k+1} = \rho_k^{-1}[\lambda^k + \mu_k \tilde{g}(x^{k+1}, z^{k+1})]_+. \tag{70}$$

Then we have

$$\begin{aligned}
&\mathrm{dist}\Big(0, \partial f(x^{k+1}, y^{k+1}) + \rho_k \partial \tilde{f}(x^{k+1}, y^{k+1}) - \rho_k\big(\nabla_x \tilde{f}(x^{k+1}, z^{k+1}) + \nabla_x \tilde{g}(x^{k+1}, z^{k+1})\lambda_{\mathbf{z}}^{k+1}; 0\big) \\
&\quad + \nabla \tilde{g}(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1}\Big) \le \epsilon_k, \qquad (71)\\
&\mathrm{dist}\Big(0, \rho_k(\partial_z \tilde{f}(x^{k+1}, z^{k+1}) + \nabla_z \tilde{g}(x^{k+1}, z^{k+1})\lambda_{\mathbf{z}}^{k+1})\Big) \le \epsilon_k, \qquad (72)\\
&\|[\tilde{g}(x^{k+1}, z^{k+1})]_+\| \le 2\mu_k^{-1}G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}, \qquad (73)\\
&|\langle \lambda_{\mathbf{z}}^{k+1}, \tilde{g}(x^{k+1}, z^{k+1})\rangle| \le 2\rho_k^{-1}\mu_k^{-1}G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}} \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\}, \qquad (74)\\
&\|[\tilde{g}(x^{k+1}, y^{k+1})]_+\| \le 2\mu_k^{-1}G^{-1}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})D_{\mathbf{y}}, \qquad (75)\\
&|\langle \lambda_{\mathbf{y}}^{k+1}, \tilde{g}(x^{k+1}, y^{k+1})\rangle| \le 2\mu_k^{-1}G^{-1}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})D_{\mathbf{y}} \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\}. \qquad (76)
\end{aligned}$$

Proof. Suppose that $(x^{k+1}, y^{k+1}, z^{k+1}, \lambda^{k+1})$ is generated by Algorithm 1 for some $0 \le k \in \mathbb{K} - 1$ satisfying (69). Notice that $(x^{k+1}, y^{k+1}, z^{k+1})$ is an ϵ_k-primal-dual stationary point of (12). It then follows from Definition 1 that

$$\begin{aligned}
&\mathrm{dist}\big(0, \partial_{(x,y)}\mathcal{L}(x^{k+1}, y^{k+1}, z^{k+1}, \lambda^k; \rho_k, \mu_k)\big) \le \epsilon_k, \qquad (77)\\
&\mathrm{dist}\big(0, \partial_z \mathcal{L}(x^{k+1}, y^{k+1}, z^{k+1}, \lambda^k; \rho_k, \mu_k)\big) \le \epsilon_k. \qquad (78)
\end{aligned}$$

In view of these, (8) and (70), one has

$$\begin{aligned}
&\partial_{(x,y)}\mathcal{L}(x^{k+1}, y^{k+1}, z^{k+1}, \lambda^k; \rho_k, \mu_k)\\
&= \partial f(x^{k+1}, y^{k+1}) + \rho_k \partial \tilde{f}(x^{k+1}, y^{k+1}) + \nabla \tilde{g}(x^{k+1}, y^{k+1})[\lambda^k + \mu_k \tilde{g}(x^{k+1}, y^{k+1})]_+\\
&\quad - \Big(\rho_k \nabla_x \tilde{f}(x^{k+1}, z^{k+1}) + \nabla_x \tilde{g}(x^{k+1}, z^{k+1})[\lambda^k + \mu_k \tilde{g}(x^{k+1}, z^{k+1})]_+; 0\Big)\\
&= \partial f(x^{k+1}, y^{k+1}) + \rho_k \partial \tilde{f}(x^{k+1}, y^{k+1})\\
&\quad - \rho_k\Big(\nabla_x \tilde{f}(x^{k+1}, z^{k+1}) + \nabla_x \tilde{g}(x^{k+1}, z^{k+1})\lambda_{\mathbf{z}}^{k+1}; 0\Big) + \nabla \tilde{g}(x^{k+1}, y^{k+1})\lambda_{\mathbf{y}}^{k+1},
\end{aligned}$$

$$\begin{aligned}
\partial_z \mathcal{L}(x^{k+1}, y^{k+1}, z^{k+1}, \lambda^k; \rho_k, \mu_k) &= -\rho_k \partial_z \tilde{f}(x^{k+1}, z^{k+1}) - \nabla_z \tilde{g}(x^{k+1}, z^{k+1})[\lambda^k + \mu_k \tilde{g}(x^{k+1}, z^{k+1})]_+\\
&= -\rho_k\Big(\partial_z \tilde{f}(x^{k+1}, z^{k+1}) + \nabla_z \tilde{g}(x^{k+1}, z^{k+1})\lambda_{\mathbf{z}}^{k+1}\Big).
\end{aligned}$$

These relations together with (77) and (78) imply that (71) and (72) hold.

Notice from Algorithm 1 that $0 < \tau < 1$, which together with (69) implies that (60) holds for μ_k and ρ_k. It then follows that (61) and (62) hold, which immediately yields (73), (75), and

$$\|\lambda^{k+1}\| \le 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}, \quad \|[\lambda^k + \mu_k \tilde{g}(x^{k+1}, y^{k+1})]_+\| \le 2G^{-1}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})D_{\mathbf{y}}. \tag{79}$$

Also, notice from (70) and $\lambda^{k+1} = [\lambda^k + \mu_k \tilde{g}(x^{k+1}, z^{k+1})]_+$ that $\lambda_{\mathbf{z}}^{k+1} = \rho_k^{-1}\lambda^{k+1}$. By this, (70) and (79), one has

$$\|\lambda_{\mathbf{z}}^{k+1}\| \le 2\rho_k^{-1}G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}, \qquad \|\lambda_{\mathbf{y}}^{k+1}\| \le 2G^{-1}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})D_{\mathbf{y}}. \tag{80}$$

Observe from (70) that $\langle \lambda_{\mathbf{y}}^{k+1}, \lambda^k + \mu_k \tilde{g}(x^{k+1}, y^{k+1})\rangle = \|[\lambda^k + \mu_k \tilde{g}(x^{k+1}, y^{k+1})]_+\|^2 \ge 0$, which implies that

$$-\langle \lambda_{\mathbf{y}}^{k+1}, \mu_k^{-1}\lambda^k\rangle \le \langle \lambda_{\mathbf{y}}^{k+1}, \tilde{g}(x^{k+1}, y^{k+1})\rangle. \tag{81}$$

In addition, we claim that

$$\|\lambda^k\| \le \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\}. \tag{82}$$

Indeed, (82) clearly holds if $k = 0$. We now assume that $k > 0$. Notice from Algorithm 1 that $\mu_{k-1} = \tau^3\mu_k$ and $\rho_{k-1} = \tau\rho_k$, which along with (69) imply that $\rho_{k-1}^{-1}\mu_{k-1} \ge 8G^{-2}\vartheta$. By this and Lemma 2 with k replaced by $k-1$, one can conclude that $\|\lambda^k\| \le 2G^{-1}(\epsilon_0 + \rho_{k-1}L_{\tilde{f}})D_{\mathbf{y}}$. This together with $\rho_{k-1} < \rho_k$ implies that (82) holds as desired.

We next show that (74) and (76) hold. By $\lambda_{\mathbf{y}}^{k+1}, \lambda_{\mathbf{z}}^{k+1} \ge 0$, (73), (75), (80), (81) and (82), one has

$$\begin{aligned}
\langle \lambda_{\mathbf{z}}^{k+1}, \tilde{g}(x^{k+1}, z^{k+1})\rangle &\le \langle \lambda_{\mathbf{z}}^{k+1}, [\tilde{g}(x^{k+1}, z^{k+1})]_+\rangle \le \|\lambda_{\mathbf{z}}^{k+1}\|\|[\tilde{g}(x^{k+1}, z^{k+1})]_+\| \\
&\overset{(73)(80)}{\le} 4\rho_k^{-1}\mu_k^{-1}G^{-2}(\epsilon_0 + \rho_k L_{\tilde{f}})^2 D_{\mathbf{y}}^2, \\
\langle \lambda_{\mathbf{z}}^{k+1}, \tilde{g}(x^{k+1}, z^{k+1})\rangle &= \rho_k^{-1}\langle \lambda^{k+1}, \tilde{g}(x^{k+1}, z^{k+1})\rangle \overset{(66)}{\ge} -\rho_k^{-1}\langle \lambda^{k+1}, \mu_k^{-1}\lambda^k\rangle \ge -\rho_k^{-1}\mu_k^{-1}\|\lambda^{k+1}\|\|\lambda^k\| \\
&\overset{(79)(82)}{\ge} -2\rho_k^{-1}\mu_k^{-1}G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}} \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\}, \\
\langle \lambda_{\mathbf{y}}^{k+1}, \tilde{g}(x^{k+1}, y^{k+1})\rangle &\le \langle \lambda_{\mathbf{y}}^{k+1}, [\tilde{g}(x^{k+1}, y^{k+1})]_+\rangle \le \|\lambda_{\mathbf{y}}^{k+1}\|\|[\tilde{g}(x^{k+1}, y^{k+1})]_+\| \\
&\overset{(75)(80)}{\le} 4\mu_k^{-1}G^{-2}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})^2 D_{\mathbf{y}}^2, \\
\langle \lambda_{\mathbf{y}}^{k+1}, \tilde{g}(x^{k+1}, y^{k+1})\rangle &\overset{(81)}{\ge} \langle \lambda_{\mathbf{y}}^{k+1}, -\mu_k^{-1}\lambda^k\rangle \ge -\mu_k^{-1}\|\lambda_{\mathbf{y}}^{k+1}\|\|\lambda^k\| \\
&\overset{(80)(82)}{\ge} -2\mu_k^{-1}G^{-1}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})D_{\mathbf{y}} \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\}.
\end{aligned}$$

These relations imply that (74) and (76) hold. □

The following lemma provides an estimate on operation complexity at step 3 of Algorithm 1 for problem (1) with $\sigma = 0$, i.e., $\tilde{f}_1(x,\cdot)$ being convex but not strongly convex for any given $x \in \mathrm{dom}\, f_2$.

Lemma 5. *Suppose that Assumption 1 holds with $\sigma = 0$, i.e., $\tilde{f}_1(x,\cdot)$ being convex but not strongly convex for any given $x \in \mathrm{dom}\, f_2$. Let f^*, L_k, $\tilde{f}_{\rm hi}^*$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $f_{\rm hi}$, $f_{\rm low}$ $\tilde{f}_{\rm low}$, $\tilde{g}_{\rm hi}$, $\mathbb{K}$ and ϑ be defined in (1), (13), (23), (24), (25), (26), (27) and (29), $L_{\tilde{f}}$ be given in Assumption 1, ϵ_k, ρ_k and μ_k be given in Algorithm 1, and*

$$\begin{aligned}
\alpha_k &= \min\left\{1, \sqrt{4\epsilon_k/(D_{\mathbf{y}}L_k)}\right\}, && (83)\\
\delta_k &= (2+\alpha_k^{-1})L_k(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2) + \max\{\epsilon_k/D_{\mathbf{y}}, \alpha_k L_k/4\}\, D_{\mathbf{y}}^2, && (84)\\
M_k &= \frac{16\max\{1/(2L_k), \min\{D_{\mathbf{y}}/\epsilon_k, 4/(\alpha_k L_k)\}\}\,\mu_k}{[(3L_k + \epsilon_k/(2D_{\mathbf{y}}))^2/\min\{L_k, \epsilon_k/(2D_{\mathbf{y}})\} + 3L_k + \epsilon_k/(2D_{\mathbf{y}})]^{-2}\,\epsilon_k^2} \times \Big(\delta_k + 2\alpha_k^{-1} \\
&\quad \times \Big(f^* - f_{\rm low} + \rho_k(\tilde{f}_{\rm hi}^* - \tilde{f}_{\rm low}) + \rho_k L_{\tilde{f}} D_{\mathbf{y}} + 3\rho_k\vartheta + \mu_k\tilde{g}_{\rm hi}^2 + \epsilon_k D_{\mathbf{y}}/4 + L_k(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\Big)\Big), && (85)\\
T_k &= \left\lceil 16\left(f_{\rm hi} - f_{\rm low} + \rho_k\epsilon_k + \epsilon_k D_{\mathbf{y}}/4\right) L_k\epsilon_k^{-2} + 8(1 + 4D_{\mathbf{y}}^2 L_k^2\epsilon_k^{-2})\mu_k^{-1} - 1\right\rceil_+, && (86)\\
N_k &= \left(\left\lceil 96\sqrt{2}\left(1 + (24L_k + 4\epsilon_k/D_{\mathbf{y}})\, L_k^{-1}\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}}L_k\epsilon_k^{-1}}\right\} \\
&\quad \times ((T_k+1)(\log M_k)_+ + T_k + 1 + 2T_k\log(T_k+1)). && (87)
\end{aligned}$$

Then for all $0 \le k \in \mathbb{K} - 1$, an ϵ_k-primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of problem (12) is successfully found at step 3 of Algorithm 1 that satisfies

$$\max_z \mathcal{L}(x^{k+1}, y^{k+1}, z, \lambda^k; \rho_k, \mu_k) \le f_{\rm hi} + \rho_k\epsilon_k + \frac{\epsilon_k D_{\mathbf{y}}}{4} + \frac{1}{2\mu_k}\left(L_k^{-1}\epsilon_k^2 + 4D_{\mathbf{y}}^2 L_k\right). \tag{88}$$

Moreover, the total number of evaluations of ∇f_1, $\nabla\tilde{f}_1$, $\nabla\tilde{g}$ and proximal operators of f_2 and $\tilde{f}_2$ performed at step 3 in iteration k of Algorithm 1 is no more than N_k, respectively.

Proof. Observe from (8) and Assumption 1 that problem (12) can be viewed as

$$\min_{x,y}\max_z\{h(x,y,z) + p(x,y) - q(z)\}$$

with

$$h(x,y,z) = f_1(x,y) + \rho_k \tilde{f}_1(x,y) + \frac{1}{2\mu_k}\|[\lambda^k + \mu_k \tilde{g}(x,y)]_+\|^2 - \rho_k \tilde{f}_1(x,z) - \frac{1}{2\mu_k}\|[\lambda^k + \mu_k \tilde{g}(x,z)]_+\|^2,$$
$$p(x,y) = f_2(x) + \rho_k \tilde{f}_2(y), \quad q(z) = \rho_k \tilde{f}_2(z).$$

By (28) and Assumption 1, it can be verified that $\|[\lambda^k+\mu_k\tilde{g}(x,y)]_+\|^2/(2\mu_k)$ and $\|[\lambda^k+\mu_k\tilde{g}(x,z)]_+\|^2/(2\mu_k)$ are both $(\mu_k L_{\tilde{g}}^2 + \mu_k \tilde{g}_{\rm hi} L_{\nabla\tilde{g}} + \|\lambda^k\| L_{\nabla\tilde{g}})$-smooth on $\mathcal{X}\times\mathcal{Y}$. Using this and the fact that f_1 and $\tilde{f}_1$ are respectively $L_{\nabla f_1}$- and $L_{\nabla\tilde{f}_1}$-smooth on $\mathcal{X}\times\mathcal{Y}$, we can see that $h(x,y,z)$ is L_k-smooth on $\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}$ for all $0 \le k \in \mathbb{K}-1$, where L_k is given in (13). In addition, it follows from Assumption 1 and $\sigma = 0$ that $h(x,y,\cdot)$ is concave but not strongly concave. Consequently, it follows from Theorem 5 (see Appendix B) that an ϵ_k-primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of problem (12) is successfully found by Algorithm 5 at step 3 of Algorithm 1.

In addition, by (8), (9) and (25), one has

$$\begin{aligned}\min_{x,y}\max_z \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k) &\overset{(8)(9)}{=} \min_{x,y}\left\{f(x,y) + \rho_k\widetilde{\mathcal{L}}(x,y,\lambda^k;\rho_k,\mu_k) - \min_z \rho_k\widetilde{\mathcal{L}}(x,z,\lambda^k;\rho_k,\mu_k)\right\}\\ &\ge \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x,y) \overset{(25)}{=} f_{\rm low}. \end{aligned} \tag{89}$$

Let (x^*,y^*) be an optimal solution of (1). It then follows that $f(x^*,y^*) = f^*$, $\tilde{f}(x^*,y^*) = \tilde{f}^*(x^*)$ and $\tilde{g}(x^*,y^*) \le 0$, where f^* and $\tilde{f}^*$ are defined in (1) and (14), respectively. Using these, (8), (9), (23), (26) and (57), we obtain that

$$\begin{aligned}&\min_{x,y}\max_z \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k) \le \max_z \mathcal{L}(x^*,y^*,z,\lambda^k;\rho_k,\mu_k)\\ &\overset{(8)(9)}{=} f(x^*,y^*) + \rho_k\tilde{f}(x^*,y^*) + \frac{1}{2\mu_k}\|[\lambda^k+\mu_k\tilde{g}(x^*,y^*)]_+\|^2 - \min_z \rho_k\widetilde{L}(x^*,z,\lambda^k;\rho_k,\mu_k)\\ &\le f^* + \rho_k\tilde{f}^*(x^*) + \frac{1}{2\mu_k}\|\lambda^k\|^2 - \min_z\left\{\rho_k\tilde{f}(x^*,z) + \frac{1}{2\mu_k}\|[\lambda^k+\mu_k\tilde{g}(x^*,z)]_+\|^2\right\}\\ &\overset{(23)(26)}{\le} f^* + \rho_k(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}) + \frac{1}{2\mu_k}\|\lambda^k\|^2 \overset{(57)}{\le} f^* + \rho_k(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}) + \rho_k\vartheta, \end{aligned} \tag{90}$$

where the second inequality is due to $\tilde{f}(x^*,y^*) = \tilde{f}^*(x^*)$, $\tilde{g}(x^*,y^*) \le 0$, and (9). Also, by (8), (24), (25), (26) and (57), one has

$$\begin{aligned}&\min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k)\\ &\overset{(8)}{\ge} \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}}\left\{f(x,y) + \rho_k(\tilde{f}(x,y) - \tilde{f}(x,z)) - \frac{1}{2\mu_k}\|[\lambda^k+\mu_k\tilde{g}(x,z)]_+\|^2\right\}\\ &\ge \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}}\left\{f(x,y) - \rho_k L_{\tilde{f}}\|y-z\| - \frac{1}{2\mu_k}\left(\|\lambda^k\| + \mu_k\|[\tilde{g}(x,z)]_+\|\right)^2\right\}\\ &\ge \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}}\left\{f(x,y) - \rho_k L_{\tilde{f}}\|y-z\| - \frac{1}{\mu_k}\|\lambda^k\|^2 - \mu_k\|[\tilde{g}(x,z)]_+\|^2\right\}\\ &\ge f_{\rm low} - \rho_k L_{\tilde{f}} D_{\mathbf{y}} - 2\rho_k\vartheta - \mu_k\tilde{g}^2_{\rm hi}, \end{aligned} \tag{91}$$

where the second inequality is due to $\lambda^k \in \mathbb{R}^l_+$ and $L_{\tilde{f}}$-Lipschitz continuity of $\tilde{f}$ (see Assumption 1(i)), and the last inequality is due to (24), (25), (26) and (57). Notice from step 2 of Algorithm 1 that $y^k_{\rm init}$ is an approximate solution of $\min_z \widetilde{L}(x^k,z,\lambda^k;\rho_k,\mu_k)$ satisfying (10). It then follows from (8), (9), (10) and (25) that

$$\begin{aligned}\max_z \mathcal{L}(x^k,y^k_{\rm init},z,\lambda^k;\rho_k,\mu_k) &\overset{(8)(9)}{=} f(x^k,y^k_{\rm init}) + \rho_k\left(\widetilde{\mathcal{L}}(x^k,y^k_{\rm init},\lambda^k;\rho_k,\mu_k) - \min_z \widetilde{\mathcal{L}}(x^k,z,\lambda^k;\rho_k,\mu_k)\right)\\ &\overset{(10)}{\le} f(x^k,y^k_{\rm init}) + \rho_k\epsilon_k \overset{(25)}{\le} f_{\rm hi} + \rho_k\epsilon_k. \end{aligned} \tag{92}$$

To complete the rest of the proof, let

$$H(x,y,z) = \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k), \quad H^* = \min_{x,y}\max_z \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k), \tag{93}$$

$$H_{\rm low} = \min\left\{\mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k) | (x,y,z) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\right\}. \tag{94}$$

In view of these, (89), (90), (91) and (92), we obtain that

$$\begin{aligned}
&\max_z H(x^k, y^k_{\rm init}, z) \overset{(92)}{\le} f_{\rm hi} + \rho_k\epsilon_k,\\
&f_{\rm low} \overset{(89)}{\le} H^* \overset{(90)}{\le} f^* + \rho_k(\tilde f^*_{\rm hi} - \tilde f_{\rm low}) + \rho_k\vartheta,\\
&H_{\rm low} \overset{(91)}{\ge} f_{\rm low} - \rho_k L_{\tilde f} D_{\mathbf{y}} - 2\rho_k\vartheta - \mu_k\tilde g^2_{\rm hi}.
\end{aligned}$$

Using these and Theorem 5 (see Appendix B) with $\hat x^0 = (x^k, y^k_{\rm init})$, $\epsilon = \epsilon_k$, $\hat\epsilon_0 = \epsilon_k/(2\sqrt{\mu_k})$, $L_{\nabla h} = L_k$, $\widehat L = 3L_k + \epsilon_k/(2D_{\mathbf{y}})$, $\sigma_y = 0$, $\hat\sigma_y = \epsilon_k/(2D_{\mathbf{y}})$, $\hat\alpha = \alpha_k$, $\hat\delta = \delta_k$, $D_p = \sqrt{D^2_{\mathbf{x}} + D^2_{\mathbf{y}}}$, $D_q = D_{\mathbf{y}}$, and H, H^*, $H_{\rm low}$ given in (93) and (94), we can conclude that the ϵ_k-primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of problem (12) found at step 3 of Algorithm 1 satisfies (88). Moreover, the total number of evaluations of ∇f_1, $\nabla\tilde f_1$, $\nabla\tilde g$ and proximal operators of f_2 and $\tilde f_2$ performed by Algorithm 5 at step 3 of Algorithm 1 is no more than N_k, respectively. □

The next lemma presents an upper bound on the optimality violation of y^{k+1} for the lower-level problem of (1) when $\sigma = 0$ and $x = x^{k+1}$.

Lemma 6. *Suppose that Assumptions 1 and 2 hold with $\sigma = 0$, i.e., $\tilde f_1(x,\cdot)$ being convex but not strongly convex for any given $x \in \operatorname{dom} f_2$. Let $\tilde f^*$, L_k, $D_{\mathbf{y}}$, $f_{\rm hi}$, $f_{\rm low}$ and $\mathbb{K}$ be defined in* (14), (13), (24), (25) *and* (27), *L_f, $L_{\tilde f}$ and G be given in Assumptions 1 and 2, and ϵ_k, ρ_k, μ_k and λ^0 be given in Algorithm 1. Suppose that $(x^{k+1}, y^{k+1}, \lambda^{k+1})$ is generated by Algorithm 1 for some $0 \le k \in \mathbb{K} - 1$ satisfying* (69). *Then we have*

$$\begin{aligned}
|\tilde f(x^{k+1}, y^{k+1}) - \tilde f^*(x^{k+1})| \le \max\Bigg\{&2\mu_k^{-1}G^{-2}L_{\tilde f}(\epsilon_0 + L_f + \rho_k L_{\tilde f})D^2_{\mathbf{y}},\\
&\rho_k^{-1}\mu_k^{-1}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde f})D_{\mathbf{y}}\}/2\\
&+\rho_k^{-1}\left(f_{\rm hi} - f_{\rm low} + \rho_k\epsilon_k + \frac{\epsilon_k D_{\mathbf{y}}}{4} + \frac{1}{2\mu_k}\left(L_k^{-1}\epsilon_k^2 + 4D^2_{\mathbf{y}}L_k\right)\right)\Bigg\}.
\end{aligned}$$

Proof. Notice from (69) and the proof of Lemma 4 that (82) holds. Using this, (8), (9), (25) and (56), we have

$$\begin{aligned}
&\max_z \mathcal{L}(x^{k+1}, y^{k+1}, z, \lambda^k; \rho_k, \mu_k)\\
&\overset{(8)(9)}{=} f(x^{k+1}, y^{k+1}) + \rho_k\tilde f(x^{k+1}, y^{k+1}) + \frac{1}{2\mu_k}\|[\lambda^k + \mu_k\tilde g(x^{k+1}, y^{k+1})]_+\|^2 - \min_z \rho_k\widetilde{\mathcal{L}}(x^{k+1}, z, \lambda^k; \rho_k, \mu_k)\\
&\ge f(x^{k+1}, y^{k+1}) + \rho_k\tilde f(x^{k+1}, y^{k+1}) - \min_z \rho_k\widetilde{\mathcal{L}}(x^{k+1}, z, \lambda^k; \rho_k, \mu_k)\\
&\overset{(25)(56)}{\ge} f_{\rm low} + \rho_k\big(\tilde f(x^{k+1}, y^{k+1}) - \tilde f^*(x^{k+1})\big) - \frac{1}{2\mu_k}\|\lambda^k\|^2\\
&\overset{(82)}{\ge} f_{\rm low} + \rho_k\big(\tilde f(x^{k+1}, y^{k+1}) - \tilde f^*(x^{k+1})\big) - \mu_k^{-1}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde f})D_{\mathbf{y}}\}/2.
\end{aligned}$$

This together with (88) implies that

$$\begin{aligned}
\tilde f(x^{k+1}, y^{k+1}) - \tilde f^*(x^{k+1}) \ \le\ &\rho_k^{-1}\left(f_{\rm hi} - f_{\rm low} + \rho_k\epsilon_k + \frac{\epsilon_k D_{\mathbf{y}}}{4} + \frac{1}{2\mu_k}\left(L_k^{-1}\epsilon_k^2 + 4D^2_{\mathbf{y}}L_k\right)\right)\\
&+\rho_k^{-1}\mu_k^{-1}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde f})D_{\mathbf{y}}\}/2. \qquad (95)
\end{aligned}$$

On the other hand, let $\lambda^* \in \mathbb{R}^l_+$ be an optimal Lagrangian multiplier of problem (14) with $x = x^{k+1}$. It then follows from Lemma 1(i) that $\|\lambda^*\| \le G^{-1}L_{\tilde f}D_{\mathbf{y}}$. Using these, (14) and (75), we have

$$\begin{aligned}
\tilde f^*(x^{k+1}) &= \min_y\left\{\tilde f(x^{k+1}, y) + \langle\lambda^*, \tilde g(x^{k+1}, y)\rangle\right\} \le \tilde f(x^{k+1}, y^{k+1}) + \langle\lambda^*, \tilde g(x^{k+1}, y^{k+1})\rangle\\
&\le \tilde f(x^{k+1}, y^{k+1}) + \|\lambda^*\|\|[\tilde g(x^{k+1}, y^{k+1})]_+\| \le \tilde f(x^{k+1}, y^{k+1}) + 2\mu_k^{-1}G^{-2}L_{\tilde f}(\epsilon_0 + L_f + \rho_k L_{\tilde f})D^2_{\mathbf{y}}.
\end{aligned}$$

The conclusion of this lemma then follows from this and (95). □

The following lemma provides an operation complexity of step 2 of Algorithm 1 for problem (1) with $\sigma = 0$, i.e., $\tilde{f}_1(x,\cdot)$ being convex but not strongly convex for any given $x \in \mathrm{dom}\, f_2$.

Lemma 7. *Suppose that Assumption 1 holds with $\sigma = 0$, i.e., $\tilde{f}_1(x,\cdot)$ being convex but not strongly convex for any given $x \in \mathrm{dom}\, f_2$. Let $\widetilde{L}_k$, $D_{\mathbf{y}}$ and $\mathbb{K}$ be defined in (11), (24) and (27), ϵ_k be given in Algorithm 1, and*

$$N_k' = \left\lceil D_{\mathbf{y}} \sqrt{2\epsilon_k^{-1}\widetilde{L}_k} \right\rceil. \tag{96}$$

Then for all $0 \le k \in \mathbb{K}-1$, y_{init}^k satisfying (10) is found at step 2 of Algorithm 1 by Algorithm 2 in no more than N_k' evaluations of ∇f_1, $\nabla \tilde{g}$ and the proximal operator of $\tilde{f}_2$, respectively.

Proof. Notice from (9) and Algorithm 1 that y_{init}^k satisfying (10) is found by Algorithm 2 applied to the problem

$$\min_y \left\{ \widetilde{L}(x^k, y, \lambda^k; \rho_k, \mu_k) = \phi(y) + P(y) \right\},$$

where $\phi(y) = \tilde{f}_1(x^k, y) + \|[\lambda^k + \mu_k \tilde{g}(x^k, y)]_+\|^2/(2\rho_k\mu_k)$ and $P(y) = \tilde{f}_2(y)$. By Assumption 1, $\sigma = 0$ and (28), one can see that ϕ is convex but not strongly convex and $\widetilde{L}_k$-smooth on $\mathrm{dom}\, P$, where $\widetilde{L}_k$ is given in (11). It then follows from this and Theorem 3 (see Appendix A) with $\tilde{\epsilon} = \epsilon_k$, $D_P = D_{\mathbf{y}}$ and $L_{\nabla\phi} = \widetilde{L}_k$ that Algorithm 2 finds y_{init}^k satisfying (10) in no more than N_k' iterations. Notice that each iteration of Algorithm 2 requires one evaluation of $\nabla\phi$ and the proximal operator of P, respectively. Hence, the conclusion of this lemma holds. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. (i) Observe from the definition of K in (27) and $\epsilon_k = \epsilon_0\tau^k$ that K is the smallest nonnegative integer such that $\epsilon_K \le \varepsilon$. Hence, Algorithm 1 terminates and outputs (x^{K+1}, y^{K+1}) after $K+1$ outer iterations. Also, one can see from Algorithm 1 that

$$\rho_K = \epsilon_K^{-1}, \quad \mu_K = \epsilon_K^{-3}, \quad \eta_K = \epsilon_K. \tag{97}$$

Moreover, notice from the assumption of Theorem 1 that $\varepsilon^{-2} - 8\tau^{-2}G^{-2}\vartheta \ge 0$. It then follows from this and (97) that

$$\rho_K^{-1}\mu_K = \epsilon_K^{-2} \ge \varepsilon^{-2} \ge 8\tau^{-2}G^{-2}\vartheta,$$

which implies that (69) holds for $k = K$. In addition, by (13), (30), (57) and $\mu_k \ge \rho_k \ge 1$, one has that for all $0 \le k \in \mathbb{K}-1$,

$$\begin{aligned} 2\mu_k L_{\tilde{g}}^2 \le L_k &\overset{(13)}{=} L_{\nabla f_1} + 2\rho_k L_{\nabla \tilde{f}_1} + 2\mu_k L_{\tilde{g}}^2 + 2\mu_k \tilde{g}_{\mathrm{hi}} L_{\nabla\tilde{g}} + 2\|\lambda^k\| L_{\nabla\tilde{g}} \\ &\overset{(57)}{\le} L_{\nabla f_1} + 2\rho_k L_{\nabla \tilde{f}_1} + 2\mu_k L_{\tilde{g}}^2 + 2\mu_k \tilde{g}_{\mathrm{hi}} L_{\nabla\tilde{g}} + 2\sqrt{2\rho_k\mu_k\vartheta} L_{\nabla\tilde{g}} \le \mu_k L. \end{aligned} \tag{98}$$

It then follows from $\epsilon_K \le \varepsilon$, (97) and Lemmas 4 and 6 that (35)-(41) hold, which proves statement (i) of Theorem 1.

(ii) Let K and N be given in (27) and (42). Recall from Lemmas 5 and 7 that the number of evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla\tilde{g}$, proximal operators of f_2 and $\tilde{f}_2$ performed by Algorithms 2 and 5 at iteration k of Algorithm 1 is at most $N_k + N_k'$, where N_k and N_k' are given in (87) and (96), respectively. By this and statement (i) of this theorem, one can observe that the total number of evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla\tilde{g}$ and proximal operators of f_2 and $\tilde{f}_2$ performed in Algorithm 1 is no more than $\sum_{k=0}^{K}(N_k + N_k')$, respectively. As a result, to prove statement (ii) of this theorem, it suffices to show that $\sum_{k=0}^{K}(N_k + N_k') \le N$.

To this end, using $\mu_k \geq 1 \geq \epsilon_k$, (31), (32), (33), (83), (84), (85), (86) and (98), we obtain that

$$1 \geq \alpha_k \geq \min\left\{1, \sqrt{4\epsilon_k/(\mu_k D_{\mathbf{y}} L)}\right\} \geq \epsilon_k^{1/2}\mu_k^{-1/2}\alpha, \tag{99}$$

$$\delta_k \leq (2+\epsilon_k^{-1/2}\mu_k^{1/2}\alpha^{-1})\mu_k L(D_{\mathbf{x}}^2+D_{\mathbf{y}}^2) + \max\{1/D_{\mathbf{y}}, \mu_k L/4\}D_{\mathbf{y}}^2 \leq \epsilon_k^{-1/2}\mu_k^{3/2}\delta, \tag{100}$$

$$\begin{aligned} M_k &\leq \frac{16\max\left\{1/(4\mu_k L_{\tilde{g}}^2), 2/(\epsilon_k^{1/2}\mu_k^{-1/2}\alpha\mu_k L_{\tilde{g}}^2)\right\}\mu_k}{\left[(3\mu_k L + 1/(2D_{\mathbf{y}}))^2/\min\{2\mu_k L_{\tilde{g}}^2, \epsilon_k/(2D_{\mathbf{y}})\} + 3\mu_k L + 1/(2D_{\mathbf{y}})\right]^{-2}\epsilon_k^2} \times \Bigg(\epsilon_k^{-1/2}\mu_k^{3/2}\delta \\ &\quad + 2\epsilon_k^{-1/2}\mu_k^{1/2}\alpha^{-1}\Big(f^* - f_{\rm low} + \rho_k(\tilde{f}_{\rm hi}^* - \tilde{f}_{\rm low}) + \rho_k L_{\tilde{f}} D_{\mathbf{y}} + 3\rho_k\vartheta + \mu_k\tilde{g}_{\rm hi}^2 + \frac{D_{\mathbf{y}}}{4} + \mu_k L(D_{\mathbf{x}}^2+D_{\mathbf{y}}^2)\Big)\Bigg) \end{aligned} \tag{101}$$

$$\begin{aligned} &\leq \frac{16\epsilon_k^{-1/2}\mu_k^{-1/2}\max\left\{1/(4L_{\tilde{g}}^2), 2/(\alpha L_{\tilde{g}}^2)\right\}\mu_k}{\epsilon_k^2\mu_k^{-4}\left[(3L+1/(2D_{\mathbf{y}}))^2/\min\{2L_{\tilde{g}}^2, 1/(2D_{\mathbf{y}})\} + 3L + 1/(2D_{\mathbf{y}})\right]^{-2}\epsilon_k^2} \times (\epsilon_k^{-1/2}\mu_k^{3/2}) \\ &\quad \times \left(\delta + 2\alpha^{-1}\Big(f^* - f_{\rm low} + \tilde{f}_{\rm hi}^* - \tilde{f}_{\rm low} + L_{\tilde{f}}D_{\mathbf{y}} + 3\vartheta + \tilde{g}_{\rm hi}^2 + \frac{D_{\mathbf{y}}}{4} + L(D_{\mathbf{x}}^2+D_{\mathbf{y}}^2)\Big)\right) \;=\; \epsilon_k^{-5}\mu_k^6 M, \end{aligned} \tag{102}$$

$$T_k \leq \left\lceil 16\left(f_{\rm hi} - f_{\rm low} + \rho_k\epsilon_k + \frac{D_{\mathbf{y}}}{4}\right)\epsilon_k^{-2}\mu_k L + 8(1+4D_{\mathbf{y}}^2\mu_k^2L^2\epsilon_k^{-2})\mu_k^{-1} - 1\right\rceil_+ \leq \epsilon_k^{-2}\mu_k T, \tag{103}$$

where (99) follows from (31), (83) and (98); (100) is due to (31), (84), (99) and $\mu_k \geq 1 \geq \epsilon_k$; (101) is due to (85), (98), (99), (100) and $\epsilon_k \in (0,1]$; (102) follows from $\mu_k \geq \rho_k \geq 1 \geq \epsilon_k$ and (32); and (103) is due to (98), (33) and the fact that $\epsilon_k \in (0,1]$ and $\rho_k\epsilon_k = 1$. By the above inequalities, (87), (98), $T > 1$ and $\mu_k \geq 1 \geq \epsilon_k$, one has

$$\begin{aligned} \sum_{k=0}^K N_k &\leq \sum_{k=0}^K \left(\left\lceil 96\sqrt{2}\left(1 + (24\mu_k L + 4/D_{\mathbf{y}})/(2\mu_k L_{\tilde{g}}^2)\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}}\mu_k L\epsilon_k^{-1}}\right\} \\ &\qquad \times \left((\epsilon_k^{-2}\mu_k T + 1)(\log(\epsilon_k^{-5}\mu_k^6 M))_+ + \epsilon_k^{-2}\mu_k T + 1 + 2\epsilon_k^{-2}\mu_k T\log(\epsilon_k^{-2}\mu_k T + 1)\right) \\ &\leq \sum_{k=0}^K \left(\left\lceil 96\sqrt{2}\left(1 + (12L + 2/D_{\mathbf{y}})/L_{\tilde{g}}^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}}L}\right\}\epsilon_k^{-1/2}\mu_k^{1/2} \\ &\quad \times \epsilon_k^{-2}\mu_k\left((T+1)(\log(\epsilon_k^{-5}\mu_k^6 M))_+ + T + 1 + 2T\log(\epsilon_k^{-2}\mu_k T + 1)\right) \\ &\leq \sum_{k=0}^K \left(\left\lceil 96\sqrt{2}\left(1 + (12L + 2/D_{\mathbf{y}})/L_{\tilde{g}}^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}}L}\right\} \\ &\quad \times \epsilon_k^{-5/2}\mu_k^{3/2}T\left(2(\log(\epsilon_k^{-5}\mu_k^6 M))_+ + 2 + 2\log(2\epsilon_k^{-2}\mu_k T)\right) \\ &\leq \sum_{k=0}^K \left(\left\lceil 96\sqrt{2}\left(1 + (12L + 2/D_{\mathbf{y}})/L_{\tilde{g}}^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}}L}\right\}T \\ &\quad \times \epsilon_k^{-5/2}\mu_k^{3/2}\left(14\log\mu_k - 14\log\epsilon_k + 2(\log M)_+ + 2 + 2\log(2T)\right), \end{aligned} \tag{104}$$

where the first inequality follows from $\epsilon_k \in (0,1]$, (87), (98), (102) and (103), and the second and third inequalities are due to the fact that $\mu_k \geq 1 \geq \epsilon_k$ and $T > 1$. By the definition of K in (27), one has $\tau^K \geq \tau\varepsilon/\epsilon_0$. Also, notice from Algorithm 1 that $\mu_k = \epsilon_k^{-3} = (\epsilon_0\tau^k)^{-3}$. It then follows from these and (104) that

$$\begin{aligned} \sum_{k=0}^K N_k &\leq \sum_{k=0}^K \left(\left\lceil 96\sqrt{2}\left(1 + (12L + 2/D_{\mathbf{y}})/L_{\tilde{g}}^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}}L}\right\}T \\ &\qquad \times \epsilon_k^{-7}\left(56\log(1/\epsilon_k) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\ &= \left(\left\lceil 96\sqrt{2}\left(1 + (12L + 2/D_{\mathbf{y}})/L_{\tilde{g}}^2\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}}L}\right\}T \\ &\quad \times \sum_{k=0}^K \epsilon_0^{-7}\tau^{-7k}\left(56k\log(1/\tau) + 56\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right) \end{aligned}$$

$$
\begin{aligned}
&\le \left(\left\lceil 96\sqrt{2}\left(1+(12L+2/D_{\mathbf{y}})/L_{\tilde{g}}^2\right)\right\rceil+2\right)\max\left\{2,\sqrt{D_{\mathbf{y}}L}\right\}T\\
&\quad\times\sum_{k=0}^{K}\epsilon_0^{-7}\tau^{-7k}\left(56K\log(1/\tau)+56\log(1/\epsilon_0)+2(\log M)_+ +2+2\log(2T)\right)\\
&\le \left(\left\lceil 96\sqrt{2}\left(1+(12L+2/D_{\mathbf{y}})/L_{\tilde{g}}^2\right)\right\rceil+2\right)\max\left\{2,\sqrt{D_{\mathbf{y}}L}\right\}T\epsilon_0^{-7}\\
&\quad\times\tau^{-7K}(1-\tau^7)^{-1}\left(56K\log(1/\tau)+56\log(1/\epsilon_0)+2(\log M)_+ +2+2\log(2T)\right)\\
&\le \left(\left\lceil 96\sqrt{2}\left(1+(12L+2/D_{\mathbf{y}})/L_{\tilde{g}}^2\right)\right\rceil+2\right)\max\left\{2,\sqrt{D_{\mathbf{y}}L}\right\}T\epsilon_0^{-7}(1-\tau^7)^{-1}\\
&\quad\times(\tau\varepsilon/\epsilon_0)^{-7}\left(56K\log(1/\tau)+56\log(1/\epsilon_0)+2(\log M)_+ +2+2\log(2T)\right), \qquad (105)
\end{aligned}
$$

where the second last inequality is due to $\sum_{k=0}^{K}\tau^{-7k}\le\tau^{-7K}/(1-\tau^7)$, and the last inequality follows from $\tau^K\ge\tau\varepsilon/\epsilon_0$.

In addition, observe from (11), (30), (57) and $\rho_k^{-1}\mu_k\ge 1$, one has that for all $0\le k\in\mathbb{K}-1$,

$$
\widetilde{L}_k = L_{\nabla\tilde{f}_1}+\rho_k^{-1}(\mu_k L_{\tilde{g}}^2+\mu_k\tilde{g}_{\rm hi}L_{\nabla\tilde{g}}+\|\lambda^k\|L_{\nabla\tilde{g}})\le L_{\nabla\tilde{f}_1}+\rho_k^{-1}(\mu_k L_{\tilde{g}}^2+\mu_k\tilde{g}_{\rm hi}L_{\nabla\tilde{g}}+\sqrt{2\rho_k\mu_k\vartheta}L_{\nabla\tilde{g}})\le\rho_k^{-1}\mu_k\widetilde{L}.
$$

Using this, (96), $\epsilon_k=\epsilon_0\tau^k$, $\rho_k=\epsilon_k^{-1}$, and $\mu_k=\epsilon_k^{-3}$, we have

$$
\begin{aligned}
&\sum_{k=1}^{K}N_k'\le\sum_{k=1}^{K}D_{\mathbf{y}}\sqrt{2\mu_k(\rho_k\epsilon_k)^{-1}\widetilde{L}}+K=\sum_{k=1}^{K}\epsilon_k^{-3/2}D_{\mathbf{y}}\sqrt{2\widetilde{L}}+K=\sum_{k=1}^{K}\epsilon_0^{-3/2}\tau^{-3k/2}D_{\mathbf{y}}\sqrt{2\widetilde{L}}+K\\
&\le\epsilon_0^{-3/2}\tau^{-3K/2}(1-\tau^{3/2})^{-1}D_{\mathbf{y}}\sqrt{2\widetilde{L}}+K\le\epsilon_0^{-3/2}(\tau\varepsilon/\epsilon_0)^{-3/2}(1-\tau^{3/2})^{-1}D_{\mathbf{y}}\sqrt{2\widetilde{L}}+K,
\end{aligned}
$$

where the second last inequality is due to $\sum_{k=0}^{K}\tau^{-3k/2}\le\tau^{-3K/2}/(1-\tau^{3/2})$, and the last inequality follows from $\tau^K\ge\tau\varepsilon/\epsilon_0$. This together with (42) and (105) implies that $\sum_{k=1}^{K}(N_k+N_k')\le N$. Hence, statement (ii) of Theorem 1 holds. □

In the remainder of this section, we first establish several lemmas and then use them to prove Theorem 2. In particular, the following lemma provides an operation complexity of step 3 of Algorithm 1 for problem (1) with $\sigma>0$, i.e., $\tilde{f}_1(x,\cdot)$ being strongly convex with parameter σ for any given $x\in\mathrm{dom}\,f_2$.

Lemma 8. *Suppose that Assumption 1 holds with $\sigma>0$, i.e., $\tilde{f}_1(x,\cdot)$ being strongly convex with parameter σ for any given $x\in\mathrm{dom}\,f_2$. Let f^*, L_k, $\tilde{f}_{\rm hi}^*$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $f_{\rm hi}$, $f_{\rm low}$ $\tilde{f}_{\rm low}$, $\tilde{g}_{\rm hi}$, $\mathbb{K}$ and ϑ be defined in* (1), (13), (23), (24), (25), (26), (27) *and* (29), *$L_{\tilde{f}}$ and σ be given in Assumption 1, ϵ_k, ρ_k and μ_k be given in Algorithm 1, and*

$$
\begin{aligned}
\tilde{\alpha}_k &= \min\left\{1,\sqrt{8\sigma\rho_k/L_k}\right\}, &(106)\\
\tilde{\delta}_k &= (2+\tilde{\alpha}_k^{-1})(D_{\mathbf{x}}^2+D_{\mathbf{y}}^2)L_k+\max\left\{2\sigma\rho_k,\tilde{\alpha}_kL_k/4\right\}D_{\mathbf{y}}^2, &(107)\\
\widetilde{M}_k &= \frac{16\max\left\{1/(2L_k),\min\left\{1/(2\sigma\rho_k),4/(\tilde{\alpha}_kL_k)\right\}\right\}}{\left[9L_k^2/\min\{L_k,\sigma\rho_k\}+3L_k\right]^{-2}\epsilon_k^2}\\
&\quad\times\left(\tilde{\delta}_k+2\tilde{\alpha}_k^{-1}\left(f^*-f_{\rm low}+\rho_k(\tilde{f}_{\rm hi}^*-\tilde{f}_{\rm low})+\rho_kL_{\tilde{f}}D_{\mathbf{y}}+3\rho_k\vartheta+\mu_k\tilde{g}_{\rm hi}^2+L_k(D_{\mathbf{x}}^2+D_{\mathbf{y}}^2)\right)\right), &(108)\\
\widetilde{T}_k &= \left\lceil 16\left(f_{\rm hi}-f_{\rm low}+\rho_k\epsilon_k\right)L_k\epsilon_k^{-2}+8\sigma^{-2}\rho_k^{-2}L_k^2+7\right\rceil_+, &(109)\\
\widetilde{N}_k &= 3397\max\left\{2,\sqrt{L_k/(2\sigma\rho_k)}\right\}\left((\widetilde{T}_k+1)(\log\widetilde{M}_k)_+ +\widetilde{T}_k+1+2\widetilde{T}_k\log(\widetilde{T}_k+1)\right). &(110)
\end{aligned}
$$

Then for all $0\le k\in\mathbb{K}-1$, an ϵ_k-primal-dual stationary point $(x^{k+1},y^{k+1},z^{k+1})$ of problem (12) *is successfully found at step 3 of Algorithm 1 that satisfies*

$$
\max_z\mathcal{L}(x^{k+1},y^{k+1},z,\lambda^k;\rho_k,\mu_k)\le f_{\rm hi}+\rho_k\epsilon_k+\frac{1}{2\epsilon_k^2}\left(L_k^{-1}+\sigma^{-2}\rho_k^{-2}L_k\right). \qquad (111)
$$

Moreover, the total number of evaluations of ∇f_1, $\nabla\tilde{f}_1$, $\nabla\tilde{g}$ and proximal operators of f_2 and $\tilde{f}_2$ performed at step 3 in iteration k of Algorithm 1 is no more than $\widetilde{N}_k$, respectively.

Proof. Observe from (8) and Assumption 1 that problem (12) can be viewed as

$$
\min_{x,y}\max_z\{h(x,y,z)+p(x,y)-q(z)\}
$$

with

$$h(x,y,z) = f_1(x,y) + \rho_k \tilde{f}_1(x,y) + \frac{1}{2\mu_k}\|[\lambda^k + \mu_k \tilde{g}(x,y)]_+\|^2 - \rho_k \tilde{f}_1(x,z) - \frac{1}{2\mu_k}\|[\lambda^k + \mu_k \tilde{g}(x,z)]_+\|^2,$$
$$p(x,y) = f_2(x) + \rho_k \tilde{f}_2(y), \quad q(z) = \rho_k \tilde{f}_2(z).$$

By (28) and Assumption 1, it can be verified that $\|[\lambda^k+\mu_k\tilde{g}(x,y)]_+\|^2/(2\mu_k)$ and $\|[\lambda^k+\mu_k\tilde{g}(x,z)]_+\|^2/(2\mu_k)$ are both $(\mu_k L_{\tilde{g}}^2 + \mu_k \tilde{g}_{\rm hi} L_{\nabla\tilde{g}} + \|\lambda^k\| L_{\nabla\tilde{g}})$-smooth on $\mathcal{X}\times\mathcal{Y}$. Using this, Assumption 1 with $\sigma > 0$, and the fact that f_1 and $\tilde{f}_1$ are respectively $L_{\nabla f_1}$- and $L_{\nabla \tilde{f}_1}$-smooth on $\mathcal{X}\times\mathcal{Y}$, we can see that $h(x,y,\cdot)$ is $\sigma\rho_k$-strongly-concave on $\mathcal{Y}$, and $h(x,y,z)$ is L_k-smooth on $\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}$ for all $0 \le k \in \mathbb{K}-1$, where L_k is given in (13). Consequently, it follows from Theorem 5 (see Appendix B) that an ϵ_k-primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of problem (12) is successfully found by Algorithm 5 at step 3 of Algorithm 1.

In addition, by (8), (9) and (25), one has

$$\begin{aligned}\min_{x,y}\max_z \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k) &\overset{(8)(9)}{=} \min_{x,y}\left\{f(x,y) + \rho_k \widetilde{L}(x,y,\lambda^k;\rho_k,\mu_k) - \min_z \rho_k \widetilde{L}(x,z,\lambda^k;\rho_k,\mu_k)\right\}\\ &\ge \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x,y) \overset{(25)}{=} f_{\rm low}. \end{aligned} \tag{112}$$

Let (x^*, y^*) be an optimal solution of (1). It then follows that $f(x^*,y^*) = f^*$, $\tilde{f}(x^*,y^*) = \tilde{f}^*(x^*)$ and $\tilde{g}(x^*,y^*) \le 0$, where f^* and $\tilde{f}^*$ are defined in (1) and (14), respectively. Using these, (8), (9), (23), (26) and (57), we obtain that

$$\begin{aligned}&\min_{x,y}\max_z \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k) \le \max_z \mathcal{L}(x^*,y^*,z,\lambda^k;\rho_k,\mu_k)\\ &\overset{(8)(9)}{=} f(x^*,y^*) + \rho_k\tilde{f}(x^*,y^*) + \frac{1}{2\mu_k}\|[\lambda^k+\mu_k\tilde{g}(x^*,y^*)]_+\|^2 - \min_z \rho_k\widetilde{L}(x^*,z,\lambda^k;\rho_k,\mu_k)\\ &\le f^* + \rho_k\tilde{f}^*(x^*) + \frac{1}{2\mu_k}\|\lambda^k\|^2 - \min_z\left\{\rho_k\tilde{f}(x^*,z) + \frac{1}{2\mu_k}\|[\lambda^k+\mu_k\tilde{g}(x^*,z)]_+\|^2\right\}\\ &\overset{(23)(26)}{\le} f^* + \rho_k(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}) + \frac{1}{2\mu_k}\|\lambda^k\|^2 \overset{(57)}{\le} f^* + \rho_k(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}) + \rho_k\vartheta, \end{aligned} \tag{113}$$

where the second inequality is due to $\tilde{f}(x^*,y^*) = \tilde{f}^*(x^*)$, $\tilde{g}(x^*,y^*) \le 0$ and (9). Also, by (8), (24), (25), (26) and (57), one has

$$\begin{aligned}&\min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k)\\ &\overset{(8)}{\ge} \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}}\left\{f(x,y) + \rho_k(\tilde{f}(x,y) - \tilde{f}(x,z)) - \frac{1}{2\mu_k}\|[\lambda^k+\mu_k\tilde{g}(x,z)]_+\|^2\right\}\\ &\ge \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}}\left\{f(x,y) - \rho_k L_{\tilde{f}}\|y-z\| - \frac{1}{2\mu_k}\left(\|\lambda^k\| + \mu_k\|[\tilde{g}(x,z)]_+\|\right)^2\right\}\\ &\ge \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}}\left\{f(x,y) - \rho_k L_{\tilde{f}}\|y-z\| - \frac{1}{\mu_k}\|\lambda^k\|^2 - \mu_k\|[\tilde{g}(x,z)]_+\|^2\right\}\\ &\ge f_{\rm low} - \rho_k L_{\tilde{f}} D_{\mathbf{y}} - 2\rho_k\vartheta - \mu_k\tilde{g}^2_{\rm hi}, \end{aligned} \tag{114}$$

where the second inequality is due to $\lambda^k \in \mathbb{R}^l_+$ and $L_{\tilde{f}}$-Lipschitz continuity of $\tilde{f}$ (see Assumption 1(i)), and the last inequality is due to (24), (25), (26) and (57). Notice from step 2 of Algorithm 1 that $y^k_{\rm init}$ is an approximate solution of $\min_z \widetilde{L}(x^k,z,\lambda^k;\rho_k,\mu_k)$ satisfying (10). It then follows from (8), (9), (10) and (25) that

$$\begin{aligned}\max_z \mathcal{L}(x^k, y^k_{\rm init}, z, \lambda^k;\rho_k,\mu_k) &\overset{(8)(9)}{=} f(x^k,y^k_{\rm init}) + \rho_k\left(\widetilde{\mathcal{L}}(x^k,y^k_{\rm init},\lambda^k;\rho_k,\mu_k) - \min_z \widetilde{\mathcal{L}}(x^k,z,\lambda^k;\rho_k,\mu_k)\right)\\ &\overset{(10)}{\le} f(x^k,y^k_{\rm init}) + \rho_k\epsilon_k \overset{(25)}{\le} f_{\rm hi} + \rho_k\epsilon_k. \end{aligned} \tag{115}$$

To complete the rest of the proof, let

$$H(x,y,z) = \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k), \quad H^* = \min_{x,y}\max_z \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k), \tag{116}$$

$$H_{\rm low} = \min\left\{\mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k) | (x,y,z) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\right\}. \tag{117}$$

In view of these, (112), (113), (114) and (115), we obtain that

$$\begin{aligned}
&\max_z H(x^k, y^k_{\rm init}, z) \overset{(115)}{\leq} f_{\rm hi} + \rho_k\epsilon_k,\\
&f_{\rm low} \overset{(112)}{\leq} H^* \overset{(90)}{\leq} f^* + \rho_k(\tilde f^*_{\rm hi} - \tilde f_{\rm low}) + \rho_k\vartheta,\\
&H_{\rm low} \overset{(114)}{\geq} f_{\rm low} - \rho_k L_{\tilde f} D_{\mathbf{y}} - 2\rho_k\vartheta - \mu_k\tilde g^2_{\rm hi}.
\end{aligned}$$

Using these and Theorem 5 (see Appendix B) with $\hat x^0 = (x^k, y^k_{\rm init})$, $\epsilon = \epsilon_k$, $\hat\epsilon_0 = \epsilon_k/2$, $\hat\sigma_y = \sigma_y = \sigma\rho_k$, $L_{\nabla h} = L_k$, $\widehat L = 3L_k$, $\hat\alpha = \alpha_k$, $\hat\delta = \delta_k$, $D_p = \sqrt{D^2_{\mathbf{x}} + D^2_{\mathbf{y}}}$, $D_q = D_{\mathbf{y}}$, and H, H^*, $H_{\rm low}$ given in (116) and (117), we can conclude that the ϵ_k-primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of problem (12) found at step 3 of Algorithm 1 satisfies (88). Moreover, the total number of evaluations of ∇f_1, $\nabla\tilde f_1$, $\nabla\tilde g$ and proximal operators of f_2 and $\tilde f_2$ performed by Algorithm 5 at step 3 of Algorithm 1 is no more than $\widetilde N_k$, respectively. □

The next lemma presents an upper bound on the optimality violation of y^{k+1} for the lower-level problem of (1) when $\sigma > 0$ and $x = x^{k+1}$.

Lemma 9. *Suppose that Assumptions 1 and 2 hold with $\sigma > 0$, i.e., $\tilde f_1(x,\cdot)$ is strongly convex with parameter σ for any given $x \in \operatorname{dom} f_2$. Let $\tilde f^*$, L_k, $D_{\mathbf{y}}$, $f_{\rm hi}$, $f_{\rm low}$ and $\mathbb{K}$ be defined in* (14), (13), (24), (25) *and* (27), *L_f, $L_{\tilde f}$, σ and G be given in Assumptions 1 and 2, and ϵ_k, ρ_k, μ_k and λ^0 be given in Algorithm 1. Suppose that $(x^{k+1}, y^{k+1}, \lambda^{k+1})$ is generated by Algorithm 1 for some $0 \leq k \in \mathbb{K} - 1$ satisfying* (69). *Then we have*

$$\begin{aligned}
|\tilde f(x^{k+1}, y^{k+1}) - \tilde f^*(x^{k+1})| \leq \max\Bigg\{ &2\mu_k^{-1}G^{-2}L_{\tilde f}(\epsilon_0 + L_f + \rho_k L_{\tilde f})D^2_{\mathbf{y}},\\
&\rho_k^{-1}\mu_k^{-1}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde f})D_{\mathbf{y}}\}/2\\
&+ \rho_k^{-1}\left(f_{\rm hi} - f_{\rm low} + \rho_k\epsilon_k + \frac{1}{2\epsilon_k^2}\left(L_k^{-1} + \sigma^{-2}\rho_k^{-2}L_k\right)\right)\Bigg\}.
\end{aligned}$$

Proof. Using (8), (9), (25), (56), and (82), we have

$$\begin{aligned}
&\max_z \mathcal{L}(x^{k+1}, y^{k+1}, z, \lambda^k; \rho_k, \mu_k)\\
&\overset{(8)(9)}{=} f(x^{k+1}, y^{k+1}) + \rho_k\tilde f(x^{k+1}, y^{k+1}) + \frac{1}{2\mu_k}\|[\lambda^k + \mu_k\tilde g(x^{k+1}, y^{k+1})]_+\|^2 - \min_z \rho_k\widetilde{\mathcal{L}}(x^{k+1}, z, \lambda^k; \rho_k, \mu_k)\\
&\geq f(x^{k+1}, y^{k+1}) + \rho_k\tilde f(x^{k+1}, y^{k+1}) - \min_z \rho_k\widetilde{\mathcal{L}}(x^{k+1}, z, \lambda^k; \rho_k, \mu_k)\\
&\overset{(25)(56)}{\geq} f_{\rm low} + \rho_k\big(\tilde f(x^{k+1}, y^{k+1}) - \tilde f^*(x^{k+1})\big) - \frac{1}{2\mu_k}\|\lambda^k\|^2\\
&\overset{(82)}{\geq} f_{\rm low} + \rho_k\big(\tilde f(x^{k+1}, y^{k+1}) - \tilde f^*(x^{k+1})\big) - \mu_k^{-1}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde f})D_{\mathbf{y}}\}/2.
\end{aligned}$$

This together with (111) implies that

$$\begin{aligned}
\tilde f(x^{k+1}, y^{k+1}) - \tilde f^*(x^{k+1}) \ \leq\ &\rho_k^{-1}\left(f_{\rm hi} - f_{\rm low} + \rho_k\epsilon_k + \frac{1}{2\epsilon_k^2}\left(L_k^{-1} + \sigma^{-2}\rho_k^{-2}L_k\right)\right)\\
&+ \rho_k^{-1}\mu_k^{-1}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde f})D_{\mathbf{y}}\}/2. \qquad (118)
\end{aligned}$$

On the other hand, let $\lambda^* \in \mathbb{R}^l_+$ be an optimal Lagrangian multiplier of problem (14) with $x = x^{k+1}$. It then follows from Lemma 1(i) that $\|\lambda^*\| \leq G^{-1}L_{\tilde f}D_{\mathbf{y}}$. Using these, (14) and (75), we have

$$\begin{aligned}
\tilde f^*(x^{k+1}) =& \min_y\left\{\tilde f(x^{k+1}, y) + \langle\lambda^*, \tilde g(x^{k+1}, y)\rangle\right\} \leq \tilde f(x^{k+1}, y^{k+1}) + \langle\lambda^*, \tilde g(x^{k+1}, y^{k+1})\rangle\\
&\leq \tilde f(x^{k+1}, y^{k+1}) + \|\lambda^*\|\|[\tilde g(x^{k+1}, y^{k+1})]_+\| \leq \tilde f(x^{k+1}, y^{k+1}) + 2\mu_k^{-1}G^{-2}L_{\tilde f}(\epsilon_0 + L_f + \rho_k L_{\tilde f})D^2_{\mathbf{y}}.
\end{aligned}$$

The conclusion of this lemma then follows from this and (118). □

The following lemma provides an estimate on operation complexity at step 2 of Algorithm 1 for problem (1) with $\sigma > 0$, i.e., $\tilde{f}_1(x,\cdot)$ being strongly convex with parameter σ for any given $x \in \mathrm{dom}\, f_2$.

Lemma 10. *Suppose that Assumptions 1 and 2 hold with $\sigma > 0$, i.e., $\tilde{f}_1(x,\cdot)$ being strongly convex with parameter σ for any given $x \in \mathrm{dom}\, f_2$. Let $\widetilde{L}_k$, $D_{\mathbf{y}}$ and $\mathbb{K}$ be defined in* (11), (24) *and* (27), *σ be given in Assumption 1, ϵ_k be given in Algorithm 1, and*

$$\widetilde{N}'_k = 2\left\lceil \sqrt{\widetilde{L}_k \sigma^{-1}} \right\rceil \max\left\{1, \left\lceil 2\log(2\epsilon_k^{-1}\widetilde{L}_k D_{\mathbf{y}}^2)\right\rceil\right\} + 1. \tag{119}$$

Then for all $0 \le k \in \mathbb{K}-1$, y_{init}^k satisfying (10) *is found at step 2 of Algorithm 1 by Algorithm 2 in no more than $\widetilde{N}'_k$ evaluations of $\nabla \tilde{f}_1$, $\nabla \tilde{g}$ and the proximal operator of $\tilde{f}_2$, respectively.*

Proof. Notice from (9) and Algorithm 1 that y_{init}^k satisfying (10) is found by Algorithm 2 applied to the problem

$$\min_y \left\{ \widetilde{L}(x^k, y, \lambda^k; \rho_k, \mu_k) = \phi(y) + P(y) \right\},$$

where $\phi(y) = \tilde{f}_1(x^k, y) + \|[\lambda^k + \mu_k \tilde{g}(x^k, y)]_+\|^2/(2\rho_k\mu_k)$ and $P(y) = \tilde{f}_2(y)$. By Assumption 1 with $\sigma > 0$ and (28), one can see that ϕ is σ-strongly-convex and $\widetilde{L}_k$-smooth on $\mathrm{dom}\, P$ with $\widetilde{L}_k$ given in (11). It then follows from this, Theorem 3 (see Appendix A) and (129) with $\tilde{\epsilon} = \epsilon_k$, $D_P = D_{\mathbf{y}}$, $\sigma_\phi = \sigma$ and $L_{\nabla\phi} = \widetilde{L}_k$ that Algorithm 3 finds y_{init}^k satisfying (10) in no more than $\widetilde{T}'_k$ iterations, where

$$\widetilde{T}'_k = \left\lceil \sqrt{\widetilde{L}_k \sigma^{-1}} \right\rceil \max\left\{1, \left\lceil 2\log(2\epsilon_k^{-1}\widetilde{L}_k D_{\mathbf{y}}^2)\right\rceil\right\}.$$

Notice that the first step of Algorithm 2 requires one evaluation of $\nabla\phi$ and the proximal operator of P, respectively, and each iteration of Algorithm 2 requires two evaluation of $\nabla\phi$ and the proximal operator of P, respectively. Hence, the conclusion of this lemma holds. □

We are now ready to prove Theorem 2.

Proof of Theorem 2. (i) Recall from the proof of Theorem 1 that (98) holds. It then follows from this, $\epsilon_K \le \varepsilon$, (97) and Lemmas 4 and 9 that (35)-(40) and (51) hold, which proves statement (i) of Theorem 2.

(ii) Let K, $\tilde{\alpha}$, $\tilde{\delta}$, $\widetilde{M}$, $\widetilde{T}$ and $\widetilde{N}$ be given in (27), (48), (49), (50) and (52), respectively. Recall from Lemmas 8 and 10 that the number of evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla \tilde{g}$, proximal operators of f_2 and $\tilde{f}_2$ performed by Algorithms 3 and 5 at iteration k of Algorithm 1 is at most $\widetilde{N}_k + \widetilde{N}'_k$, where $\widetilde{N}_k$ and $\widetilde{N}'_k$ are given in (110) and (119), respectively. By this and statement (i) of this theorem, one can observe that the total number of evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla \tilde{g}$ and proximal operators of f_2 and $\tilde{f}_2$ performed in Algorithm 1 is no more than $\sum_{k=0}^K (\widetilde{N}_k + \widetilde{N}'_k)$, respectively. As a result, to prove statement (ii) of this theorem, it suffices to show that $\sum_{k=0}^K (\widetilde{N}_k + \widetilde{N}'_k) \le \widetilde{N}$.

To this end, using $\mu_k \ge \rho_k \ge 1 \ge \epsilon_k$, (48), (49), (50), (98) (106), (107), (108) and (109), we obtain that

$$1 \ge \tilde{\alpha}_k \ge \min\left\{1, \sqrt{8\sigma\rho_k/(\mu_k L)}\right\} \ge \rho_k^{1/2}\mu_k^{-1/2}\tilde{\alpha}, \tag{120}$$

$$\tilde{\delta}_k \le (2 + \rho_k^{-1/2}\mu_k^{1/2}\tilde{\alpha}^{-1})(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\mu_k L + \max\{2\sigma\rho_k, \mu_k L/4\} D_{\mathbf{y}}^2 \le \rho_k^{-1/2}\mu_k^{3/2}\tilde{\delta}, \tag{121}$$

$$\begin{aligned}\widetilde{M}_k \le{}& \frac{16\max\left\{1/(4\mu_k L_{\tilde{g}}^2), 2/(\rho_k^{1/2}\mu_k^{-1/2}\tilde{\alpha}\mu_k L_{\tilde{g}}^2)\right\}}{\left[9\mu_k^2 L^2/\min\{2\mu_k L_{\tilde{g}}^2, \sigma\rho_k\} + 3\mu_k L\right]^{-2}\epsilon_k^2} \times \Bigg(\rho_k^{-1/2}\mu_k^{3/2}\tilde{\delta} \\ &+ 2\rho_k^{-1/2}\mu_k^{1/2}\tilde{\alpha}^{-1}\Big(f^* - f_{\mathrm{low}} + \rho_k(\tilde{f}_{\mathrm{hi}}^* - \tilde{f}_{\mathrm{low}}) + \rho_k L_{\tilde{f}} D_{\mathbf{y}} + 3\rho_k\vartheta + \mu_k\tilde{g}_{\mathrm{hi}}^2 + \mu_k L(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\Big)\Bigg)\end{aligned} \tag{122}$$

$$\begin{aligned}\le{}& \frac{16\rho_k^{-1/2}\mu_k^{-1/2}\max\left\{1/(4L_{\tilde{g}}^2), 2/(\tilde{\alpha}L_{\tilde{g}}^2)\right\}}{\rho_k^2\mu_k^{-4}\left[9L^2/\min\{2L_{\tilde{g}}^2, \sigma\} + 3L\right]^{-2}\epsilon_k^2} \times \rho_k^{-1/2}\mu_k^{3/2} \\ &\times \left(\tilde{\delta} + 2\tilde{\alpha}^{-1}\Big(f^* - f_{\mathrm{low}} + \tilde{f}_{\mathrm{hi}}^* - \tilde{f}_{\mathrm{low}} + L_{\tilde{f}} D_{\mathbf{y}} + 3\vartheta + \tilde{g}_{\mathrm{hi}}^2 + L(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\Big)\right) = \epsilon_k^{-2}\rho_k^{-3}\mu_k^5\widetilde{M},\end{aligned} \tag{123}$$

$$\widetilde{T}_k \le \left\lceil 16\left(f_{\rm hi} - f_{\rm low} + \rho_k\epsilon_k\right)\epsilon_k^{-2}\mu_k L + 8\sigma^{-2}\rho_k^{-2}\mu_k^2 L^2 + 7 \right\rceil_+ \le \epsilon_k^{-2}\mu_k\widetilde{T}, \tag{124}$$

where (120) follows from (48), (98) and (106); (121) is due to (48), (107), (120) and $\mu_k \ge 1 \ge \epsilon_k$; (122) is due to (98), (108), (120), (121) and $\epsilon_k \in (0,1]$; (123) follows from $\mu_k \ge \rho_k \ge 1 \ge \epsilon_k$ and (49); and (124) is due to (50), (98) and the fact that $\epsilon_k \in (0,1]$ and $\rho_k\epsilon_k = 1$. By the above inequalities, (98), (110), $\widetilde{T} > 1$ and $\mu_k \ge 1 \ge \epsilon_k$, one has

$$\begin{aligned}
\sum_{k=0}^{K}\widetilde{N}_k &\le \sum_{k=0}^{K} 3397\max\left\{2, \sqrt{\mu_k L/(2\sigma\rho_k)}\right\} \\
&\quad\times\left((\epsilon_k^{-2}\mu_k\widetilde{T}+1)(\log(\epsilon_k^{-2}\rho_k^{-3}\mu_k^5\widetilde{M}))_+ + \epsilon_k^{-2}\mu_k\widetilde{T} + 1 + 2\epsilon_k^{-2}\mu_k\widetilde{T}\log(\epsilon_k^{-2}\mu_k\widetilde{T}+1)\right) \\
&\le \sum_{k=0}^{K} 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\}\rho_k^{-1/2}\mu_k^{1/2}\times\epsilon_k^{-2}\mu_k\left((\widetilde{T}+1)(\log(\epsilon_k^{-2}\rho_k^{-3}\mu_k^5\widetilde{M}))_+ + \widetilde{T} + 1 + 2\widetilde{T}\log(\epsilon_k^{-2}\mu_k\widetilde{T}+1)\right) \\
&\le \sum_{k=0}^{K} 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\}\epsilon_k^{-2}\rho_k^{-1/2}\mu_k^{3/2}\widetilde{T}\left(2(\log(\epsilon_k^{-2}\rho_k^{-3}\mu_k^5\widetilde{M}))_+ + 2 + 2\log(2\epsilon_k^{-2}\mu_k\widetilde{T})\right) \\
&\le \sum_{k=0}^{K} 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\}\widetilde{T}\epsilon_k^{-2}\rho_k^{-1/2}\mu_k^{3/2}\left(12\log\mu_k - 6\log\rho_k - 8\log\epsilon_k + 2(\log\widetilde{M})_+ + 2 + 2\log(2\widetilde{T})\right),
\end{aligned} \tag{125}$$

where the first inequality follows from $\epsilon_k \in (0,1]$, (98), (110), (123) and (124), and the second and third inequalities are due to the fact that $\mu_k \ge 1 \ge \epsilon_k$ and $\widetilde{T} > 1$. By the definition of K in (27), one has $\tau^K \ge \tau\varepsilon/\epsilon_0$. Also, notice from Algorithm 1 that $\rho_k = \epsilon_k^{-1} = (\epsilon_0\tau^k)^{-1}$ and $\mu_k = \epsilon_k^{-3} = (\epsilon_0\tau^k)^{-3}$. It then follows from these and (125) that

$$\begin{aligned}
\sum_{k=0}^{K}\widetilde{N}_k &\le \sum_{k=0}^{K} 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\}\widetilde{T}\epsilon_k^{-6}\left(38\log(1/\epsilon_k) + 2(\log\widetilde{M})_+ + 2 + 2\log(2\widetilde{T})\right) \\
&= 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\}\widetilde{T}\sum_{k=0}^{K}\epsilon_0^{-6}\tau^{-6k}\left(38k\log(1/\tau) + 38\log(1/\epsilon_0) + 2(\log\widetilde{M})_+ + 2 + 2\log(2\widetilde{T})\right) \\
&\le 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\}\widetilde{T}\sum_{k=0}^{K}\epsilon_0^{-6}\tau^{-6k}\left(38K\log(1/\tau) + 38\log(1/\epsilon_0) + 2(\log\widetilde{M})_+ + 2 + 2\log(2\widetilde{T})\right) \\
&\le 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\}\widetilde{T}\epsilon_0^{-6}\tau^{-6K}(1-\tau^6)^{-1}\left(38K\log(1/\tau) + 38\log(1/\epsilon_0) + 2(\log\widetilde{M})_+ + 2 + 2\log(2\widetilde{T})\right) \\
&\le 3397\max\left\{2, \sqrt{L/(2\sigma)}\right\}\widetilde{T}\epsilon_0^{-6}(1-\tau^6)^{-1} \\
&\quad\times(\tau\varepsilon/\epsilon_0)^{-6}\left(38K\log(1/\tau) + 38\log(1/\epsilon_0) + 2(\log\widetilde{M})_+ + 2 + 2\log(2\widetilde{T})\right),
\end{aligned} \tag{126}$$

where the second last inequality is due to $\sum_{k=0}^{K}\tau^{-6k} \le \tau^{-6K}/(1-\tau^6)$, and the last inequality follows from $\tau^K \ge \tau\varepsilon/\epsilon_0$.

In addition, observe from (11), (30), (57) and $\rho_k^{-1}\mu_k \ge 1$, one has that for all $0 \le k \in \mathbb{K}-1$,

$$\widetilde{L}_k = L_{\nabla\tilde{f}_1} + \rho_k^{-1}(\mu_k L_{\tilde{g}}^2 + \mu_k\tilde{g}_{\rm hi}L_{\nabla\tilde{g}} + \|\lambda^k\|L_{\nabla\tilde{g}}) \le L_{\nabla\tilde{f}_1} + \rho_k^{-1}(\mu_k L_{\tilde{g}}^2 + \mu_k\tilde{g}_{\rm hi}L_{\nabla\tilde{g}} + \sqrt{2\rho_k\mu_k\vartheta}L_{\nabla\tilde{g}}) \le \rho_k^{-1}\mu_k\widetilde{L}.$$

Using this, (119), $\epsilon_k = \epsilon_0\tau^k$, $\rho_k = \epsilon_k^{-1}$, and $\mu_k = \epsilon_k^{-3}$, we have

$$\begin{aligned}
\sum_{k=1}^{K}\widetilde{N}'_k &\le \sum_{k=1}^{K}\left(2\left\lceil\sqrt{\frac{\rho_k^{-1}\mu_k\widetilde{L}}{\sigma}}\right\rceil\max\left\{1, \left\lceil 2\log\left(2\epsilon_k^{-1}\rho_k^{-1}\mu_k\widetilde{L}D_{\mathbf{y}}^2\right)\right\rceil\right\} + 1\right) \\
&= \sum_{k=1}^{K} 2\left\lceil(\epsilon_0\tau^k)^{-1}\sqrt{\frac{\widetilde{L}}{\sigma}}\right\rceil\max\left\{1, \left\lceil 2\log(2\widetilde{L}D_{\mathbf{y}}^2) + 6k\log(1/\tau) - 6\log\epsilon_0\right\rceil\right\} + K
\end{aligned}$$

$$\leq \sum_{k=1}^{K} 2(\epsilon_0\tau^k)^{-1}\left\lceil\sqrt{\frac{\widetilde{L}}{\sigma}}+1\right\rceil \max\left\{1,\left\lceil 2\log(2\widetilde{L}D_{\mathbf{y}}^2)+6K\log(1/\tau)-6\log\epsilon_0\right\rceil\right\}+K$$

$$\leq 2\epsilon_0^{-1}\tau^{-K}(1-\tau)\left\lceil\sqrt{\frac{\widetilde{L}}{\sigma}}+1\right\rceil \max\left\{1,\left\lceil 2\log(2\widetilde{L}D_{\mathbf{y}}^2)+6K\log(1/\tau)-6\log\epsilon_0\right\rceil\right\}+K$$

$$\leq 2(\tau\varepsilon)^{-1}(1-\tau)\left\lceil\sqrt{\frac{\widetilde{L}}{\sigma}}+1\right\rceil \max\left\{1,\left\lceil 2\log(2\widetilde{L}D_{\mathbf{y}}^2)+6K\log(1/\tau)-6\log\epsilon_0\right\rceil\right\}+K$$

where the second last inequality is due to $\sum_{k=0}^{K}\tau^{-k}\leq\tau^{-K}/(1-\tau)$, and the last inequality follows from $\tau^K\geq\tau\varepsilon/\epsilon_0$. This together with (52) and (126) implies that $\sum_{k=1}^{K}(\widetilde{N}_k+\widetilde{N}'_k)\leq\widetilde{N}$. Hence, statement (ii) of Theorem 2 holds. □

References

[1] G. B. Allende and G. Still. Solving bilevel programs with the KKT-approach. *Mathematical Programming*, 138(1):309–332, 2013.

[2] J. F. Bard. *Practical bilevel optimization: algorithms and applications*, volume 30. Springer Science & Business Media, 2013.

[3] K. P. Bennett, G. Kunapuli, J. Hu, and J.-S. Pang. Bilevel optimization and machine learning. In *IEEE World Congress on Computational Intelligence*, pages 25–47. Springer, 2008.

[4] L. Bertinetto, J. F. Henriques, P. Torr, and A. Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2018.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[6] L. Chen, J. Xu, and J. Zhang. Bilevel optimization without lower-level strong convexity from the hyper-objective perspective. *arXiv preprint arXiv:2301.00712*, 2023.

[7] T. Chen, Y. Sun, Q. Xiao, and W. Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2466–2488, 2022.

[8] F. H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[9] B. Colson, P. Marcotte, and G. Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153(1):235–256, 2007.

[10] C. Crockett, J. A. Fessler, et al. Bilevel methods for image reconstruction. *Foundations and Trends® in Signal Processing*, 15(2-3):121–289, 2022.

[11] Y.-H. Dai, J. Wang, and L. Zhang. Optimality conditions and numerical algorithms for a class of linearly constrained minimax optimization problems. *SIAM Journal on Optimization*, 34(3):2883–2916, 2024.

[12] Y.-H. Dai and L. Zhang. Optimality conditions for constrained minimax optimization. *arXiv preprint arXiv:2004.09730*, 2020.

[13] S. Dempe. *Foundations of bilevel programming*. Springer Science & Business Media, 2002.

[14] S. Dempe, V. Kalashnikov, G. A. Pérez-Valdés, and N. Kalashnykova. Bilevel programming problems. *Energy Systems. Springer, Berlin*, 10:978–3, 2015.

[15] S. Dempe and A. Zemkoho. Bilevel optimization. In *Springer optimization and its applications. Vol. 161*. Springer, 2020.

[16] S. Dempe and A. B. Zemkoho. The bilevel programming problem: reformulations, constraint qualifications and optimality conditions. *Mathematical Programming*, 138(1):447–473, 2013.

[17] M. Feurer and F. Hutter. Hyperparameter optimization. In *Automated machine learning*, pages 3–33. Springer, Cham, 2019.

[18] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pages 1165–1173, 2017.

[19] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577, 2018.

[20] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, 2014.

[21] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758, 2020.

[22] Z. Guo, Q. Hu, L. Zhang, and T. Yang. Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

[23] P. Hansen, B. Jaumard, and G. Savard. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on Scientific and Statistical Computing*, 13(5):1194–1217, 1992.

[24] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[25] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

[26] X. Hu, N. Xiao, X. Liu, and K.-C. Toh. An improved unconstrained approach for bilevel optimization. *SIAM Journal on Optimization*, 33(4):2801–2829, 2023.

[27] F. Huang, J. Li, and S. Gao. Biadam: Fast adaptive bilevel optimization methods. *arXiv preprint arXiv:2106.11396*, 2021.

[28] M. Huang, X. Chen, K. Ji, S. Ma, and L. Lai. Efficiently escaping saddle points in bilevel optimization. *Journal of Machine Learning Research*, 26(1):1–61, 2025.

[29] Y. Ishizuka and E. Aiyoshi. Double penalty method for bilevel optimization problems. *Annals of Operations Research*, 34(1):73–88, 1992.

[30] K. Ji, J. D. Lee, Y. Liang, and H. V. Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. *Advances in Neural Information Processing Systems*, 33:11490–11500, 2020.

[31] K. Ji, J. Yang, and Y. Liang. Bilevel optimization: convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892, 2021.

[32] P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in Neural Information Processing Systems*, 34:30271–30283, 2021.

[33] V. Konda and J. Tsitsiklis. Actor-critic algorithms. *Advances in Neural Information Processing Systems*, 12:1008–1014, 1999.

[34] W. Kong and R. D. Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *SIAM Journal on Optimization*, 31(4):2558–2585, 2021.

[35] D. Kovalev and A. Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. *Advances in Neural Information Processing Systems*, 35:14691–14703, 2022.

[36] J. Kwon, D. Kwon, S. Wright, and R. D. Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages 18083–18113, 2023.

[37] J. Li, B. Gu, and H. Huang. A fully single loop algorithm for bilevel optimization without hessian inverse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7426–7434, 2022.

[38] Y. Li, G.-H. Lin, J. Zhang, and X. Zhu. A novel approach for bilevel programs based on Wolfe duality. *arXiv preprint arXiv:2302.06838*, 2023.

[39] Q. Lin, Z. Lu, and L. Xiao. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4):2244–2273, 2015.

[40] B. Liu, M. Ye, S. Wright, P. Stone, and Q. Liu. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in Neural Information Processing Systems*, 35:17248–17262, 2022.

[41] H. Liu, K. Simonyan, and Y. Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations*, 2018.

[42] R. Liu, J. Gao, J. Zhang, D. Meng, and Z. Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10045–10067, 2021.

[43] D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*, 30:6470–6479, 2017.

[44] Z. Lu and S. Mei. A first-order augmented Lagrangian method for constrained minimax optimization. *Mathematical Programming*, pages 1–42, 2024.

[45] Z. Lu and S. Mei. A first-order method for nonconvex-strongly-concave constrained minimax optimization. *Preprint*, 2024. https://zhaosong-lu.github.io/ResearchPapers/strongly-cvx-minimax.pdf.

[46] Z. Lu and S. Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024.

[47] Z. Lu and Z. Zhou. Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. *SIAM Journal on Optimization*, 33(2):1159–1190, 2023.

[48] Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical programs with equilibrium constraints*. Cambridge University Press, 1996.

[49] X. Ma, W. Yao, J. J. Ye, and J. Zhang. Combined approach with second-order optimality conditions for bilevel programming problems. *arXiv preprint arXiv:2108.00179*, 2021.

[50] D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*, pages 2113–2122, 2015.

[51] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[52] J. A. Mirrlees. The theory of moral hazard and unobservable behaviour: Part I. *The Review of Economic Studies*, 66(1):3–21, 1999.

[53] Y. E. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.

[54] T. Okuno, A. Takeda, A. Kawana, and M. Watanabe. On ℓ_p-hyperparameter learning via bilevel nonsmooth optimization. *Journal of Machine Learning Research*, 22(1):11093–11139, 2021.

[55] J. Outrata, M. Kocvara, and J. Zowe. *Nonsmooth approach to optimization problems with equilibrium constraints: theory, applications and numerical results*, volume 28. Springer Science & Business Media, 1998.

[56] F. Pedregosa. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning*, pages 737–746, 2016.

[57] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. *Advances in Neural Information Processing Systems*, 32:113–124, 2019.

[58] H. Shen and T. Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, pages 30992–31015, 2023.

[59] X. Shen and L. Yu. CU splitting early termination based on weighted SVM. *EURASIP journal on image and video processing*, 2013(1):4, 2013.

[60] C. Shi, J. Lu, and G. Zhang. An extended Kuhn–Tucker approach for linear bilevel programming. *Applied Mathematics and Computation*, 162(1):51–63, 2005.

[61] K. Shimizu, Y. Ishizuka, and J. F. Bard. *Nondifferentiable and two-level mathematical programming*. Springer Science & Business Media, 2012.

[62] A. Sinha, P. Malo, and K. Deb. A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22(2):276–295, 2017.

[63] D. Sow, K. Ji, Z. Guan, and Y. Liang. A primal-dual approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022.

[64] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[65] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Manuscript, May 2008.

[66] L. N. Vicente and P. H. Calamai. Bilevel and multilevel programming: A bibliography review. *Journal of Global optimization*, 5(3):291–306, 1994.

[67] H. Von Stackelberg. *Market structure and equilibrium*. Springer Science & Business Media, 2010.

[68] X. Wang, R. Pan, R. Pi, and T. Zhang. Effective bilevel optimization via minimax reformulation. *arXiv preprint arXiv:2305.13153*, 2023.

[69] D. Ward and J. M. Borwein. Nonsmooth calculus in finite dimensions. *SIAM Journal on Control and Optimization*, 25(5):1312–1340, 1987.

[70] P. Xanthopoulos and T. Razzaghi. A weighted support vector machine method for control chart pattern recognition. *Computers & Industrial Engineering*, 70:134–149, 2014.

[71] J. Yang, K. Ji, and Y. Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34:13670–13682, 2021.

[72] X. Yang, Q. Song, and Y. Wang. A weighted support vector machine for data classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 21(05):961–976, 2007.

[73] W. Yao, H. Yin, S. Zeng, and J. Zhang. Overcoming lower-level constraints in bilevel optimization: A novel approach with regularized gap functions. *arXiv preprint arXiv:2406.01992*, 2024.

[74] W. Yao, C. Yu, S. Zeng, and J. Zhang. Constrained bi-level optimization: proximal Lagrangian value function approach and Hessian-free algorithm. *arXiv preprint arXiv:2401.16164*, 2024.

[75] J. J. Ye. Constraint qualifications and optimality conditions in bilevel optimization. In *Bilevel Optimization*, pages 227–251. Springer, 2020.

[76] J. J. Ye, X. Yuan, S. Zeng, and J. Zhang. Difference of convex algorithms for bilevel programs with applications in hyperparameter selection. *Mathematical Programming*, 198(2):1583–1616, 2023.

A Optimal first-order methods for unconstrained convex optimization problems

In this part we review optimal first-order methods for solving convex optimization problem

$$\Psi^* = \min_x \{\Psi(x) \coloneqq \phi(x) + P(x)\}, \tag{127}$$

where $P : \mathbb{R}^m \to (-\infty, \infty]$ are closed convex functions, $\phi : \mathbb{R}^m \to (-\infty, \infty]$ is a σ_ϕ-strongly-convex function with $\sigma_\phi \geq 0$, and $\nabla\phi$ is $L_{\nabla\phi}$-Lipschitz continuous on dom P. In addition, we assume that dom P is compact and let $D_P := \max_{x,y\in \text{dom } P} \|x - y\|$.

We first present an optimal first-order method in Algorithm 2 for solving problem (127) with $\sigma_\phi = 0$, i.e., ϕ being convex but not strongly convex. It is a variant of Nesterov's optimal first-order method [53] and has been studied in, for example, [65, Section 3].

Algorithm 2 An optimal first-order method for problem (127) with $\sigma_\phi = 0$

Input: $\tilde{\epsilon} > 0$, $\tilde{x}^0 \in \text{dom } P$ and $x^0 = z^0 = \tilde{x}^0$.

1: **for** $k = 0, 1, \ldots$ **do**

2: Set $y^k = (kx^k + 2z^k)/(k + 2)$.

3: Compute z^{k+1} as

$$z^{k+1} = \operatorname*{argmin}_z \left\{ \ell(z; y^k) + \frac{L_{\nabla\phi}}{k+2}\|z - z^k\|^2 \right\},$$

where

$$\ell(x; y) := \phi(y) + \langle \nabla\phi(y), x - y\rangle + P(x). \tag{128}$$

4: Set $x^{k+1} = (kx^k + 2z^{k+1})/(k + 2)$.

5: Terminate the algorithm and output x^{k+1} if

$$\Psi(x^{k+1}) - \underline{\Psi}_{k+1} \leq \tilde{\epsilon}, \qquad \text{where} \quad \underline{\Psi}_{k+1} = \frac{4}{(k+1)(k+3)} \min\left\{ \sum_{i=0}^{k} \frac{i+2}{2} \ell(x; y^i) \right\}.$$

6: **end for**

The following result provides an iteration complexity of Algorithm 2 for finding an $\tilde{\epsilon}$-optimal solution[10] of (127). It is an immediate consequence of [65, Corrolary] and its proof is thus omitted.

Theorem 3. *Let $\{(x^k, y^k)\}$ be generated by Algorithm 2 and $\ell(\cdot;\cdot)$ be defined in* (128)*. Then, $\Psi(x^k) - \Psi^* \leq \Psi(x^k) - \underline{\Psi}_k$ for all $k \geq 1$. Moreover, for any given $\tilde{\epsilon} > 0$, Algorithm 2 finds an approximate solution x^{k+1} of problem* (127) *such that $\Psi(x^{k+1}) - \Psi^* \leq \Psi(x^{k+1}) - \underline{\Psi}_{k+1} \leq \tilde{\epsilon}$ in no more than $\bar{K}$ iterations, where*

$$\bar{K} = \left\lceil D_P \sqrt{2L_{\nabla\phi}\tilde{\epsilon}^{-1}} \right\rceil.$$

We next present an optimal first-order method [47, Algorithm 4] for solving problem (127) with $\sigma_\phi > 0$, i.e., ϕ being strongly convex with parameter σ_ϕ, which is a slight variant of Nesterov's optimal first-order methods [39, 53].

[10] An $\tilde{\epsilon}$-optimal solution of problem (127) is a point x satisfying $\Psi(x) - \Psi^* \leq \tilde{\epsilon}$.

Algorithm 3 An optimal first-order method for problem (127) with $\sigma_\phi > 0$

Input: $\tilde{\epsilon} > 0$ and $\tilde{x}^0 \in \operatorname{dom} P$.

1: Compute
$$x^0 = \operatorname{prox}_{P/L_{\nabla\phi}}\left(\tilde{x}^0 - L_{\nabla\phi}^{-1}\nabla\phi(\tilde{x}^0)\right).$$
2: Set $z^0 = x^0$ and $\alpha = \sqrt{\sigma_\phi/L_{\nabla\phi}}$.
3: **for** $k = 0, 1, \ldots$ **do**
4: Set $y^k = (x^k + \alpha z^k)/(1+\alpha)$.
5: Compute z^{k+1} as
$$z^{k+1} = \operatorname*{argmin}_z \left\{\ell(z; y^k) + \frac{\alpha L_{\nabla\phi}}{2}\|z - \alpha y^k - (1-\alpha)z^k\|^2\right\},$$
where $\ell(x; y)$ is defined in (128).
6: Set $x^{k+1} = (1-\alpha)x^k + \alpha z^{k+1}$.
7: Compute
$$\tilde{x}^{k+1} = \operatorname{prox}_{P/L_{\nabla\phi}}\left(x^{k+1} - L_{\nabla\phi}^{-1}\nabla\phi(x^{k+1})\right).$$
8: Terminate the algorithm and output $\tilde{x}^{k+1}$ if
$$\|\tilde{x}^{k+1} - x^k\| \le \frac{\tilde{\epsilon}}{2L_{\nabla\phi}D_P}.$$
9: **end for**

The following result provides an iteration complexity of Algorithm 3 for finding an approximate optimal solution of problem (127), which was established in [47, Proposition 4].

Theorem 4. *Let $\{\tilde{x}^k\}$ be the sequence generated by Algorithm 3. Then for any given $\tilde{\epsilon} > 0$, an approximate solution $\tilde{x}^{k+1}$ of problem (127) satisfying $\operatorname{dist}(0, \partial\Psi(\tilde{x}^{k+1})) \le 2L_{\nabla\phi}\|\tilde{x}^{k+1} - x^{k+1}\| \le \tilde{\epsilon}/D_P$ is generated by running Algorithm 2 for at most $\widetilde{K}$ iterations, where*
$$\widetilde{K} = \left\lceil \sqrt{\frac{L_{\nabla\phi}}{\sigma_\phi}} \right\rceil \max\left\{1, \left\lceil 2\log\frac{2L_{\nabla\phi}D_P^2}{\tilde{\epsilon}} \right\rceil\right\}.$$

Remark 4. *By the convexity of Ψ, $D_P = \max_{x,y\in\operatorname{dom} P}\|x - y\|$, and Theorem 4, it is not hard to show that the output $\tilde{x}^{k+1}$ of Algorithm 3 satisfies*
$$\Psi(\tilde{x}^{k+1}) - \Psi^* \le \operatorname{dist}(0, \partial\Psi(\tilde{x}^{k+1}))D_P \le \tilde{\epsilon}. \tag{129}$$

B A first-order method for nonconvex-concave minimax problem

In this part, we present a first-order method for finding an ϵ-primal-dual stationary point of the nonconvex-concave minimax problem
$$H^* = \min_x \max_y \{H(x, y) \coloneqq h(x, y) + p(x) - q(y)\}, \tag{130}$$
which has at least one optimal solution and satisfies the following assumptions.

Assumption 3. *(i) $p : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $q : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ are proper convex functions and continuous on $\operatorname{dom} p$ and $\operatorname{dom} q$, respectively, and moreover, $\operatorname{dom} p$ and $\operatorname{dom} q$ are compact.*

(ii) The proximal operators associated with p and q can be exactly evaluated.

(iii) h is $L_{\nabla h}$-smooth on $\operatorname{dom} p \times \operatorname{dom} q$, and moreover, $h(x, \cdot)$ is σ_y-strongly-concave with $\sigma_y \ge 0$ for any given $x \in \operatorname{dom} p$.

For ease of presentation, we define

$$D_p = \max\{\|u-v\| \big| u, v \in \operatorname{dom} p\}, \quad D_q = \max\{\|u-v\| \big| u, v \in \operatorname{dom} q\}, \tag{131}$$

$$H_{\text{low}} = \min\{H(x,y) | (x,y) \in \operatorname{dom} p \times \operatorname{dom} q\}. \tag{132}$$

Recently, a first-order method was proposed in [44, Algorithm 2] for finding an ϵ-primal-dual stationary point of problem (130) with $\sigma_y = 0$, while another first-order method was proposed in [45, Algorithm 1] for finding an ϵ-primal-dual stationary point of problem (130) with $\sigma_y > 0$. We will present a unified first-order method in Algorithm 5 below by combining these two methods. Specifically, given an iterate (x^k, y^k), this unified first-order method finds the next iterate (x^{k+1}, y^{k+1}) by applying [44, Algorithm 1], which is a slight modification of a novel optimal first-order method [35, Algorithm 4] by incorporating a forward-backward splitting scheme and a verifiable termination criterion (see steps 23-25 in Algorithm 4), to the strongly-convex-strongly-concave minimax problem

$$\min_x \max_y \{h_k(x,y) + p(x) - q(y)\},$$

where

$$h_k(x,y) = \begin{cases} h(x,y) - \epsilon\|y-y^0\|^2/(4D_q) + L_{\nabla h}\|x-x^k\|^2, & \text{if} \quad \sigma_y = 0, \\ h(x,y) + L_{\nabla h}\|x-x^k\|^2, & \text{if} \quad \sigma_y > 0. \end{cases} \tag{133}$$

This minimax problem arises from applying a proximal point method to the minimization problem $\min_x\{\max_y h(x,y) - q(y) - \epsilon\|y-y^0\|^2/(4D_q) + p(x)\}$ if $\sigma_y = 0$ or the minimization problem $\min_x\{\max_y h(x,y) - q(y) + p(x)\}$ if $\sigma_y > 0$. One can easily observe that h_k is $L_{\nabla h}$-strongly-convex-$\hat{\sigma}_y$-strongly-concave and $\widehat{L}$-smooth on $\operatorname{dom} p \times \operatorname{dom} q$, where

$$\hat{\sigma}_y = \begin{cases} \epsilon/(2D_q), & \text{if} \quad \sigma_y = 0, \\ \sigma_y, & \text{if} \quad \sigma_y > 0, \end{cases} \qquad \widehat{L} = \begin{cases} 3L_{\nabla h} + \epsilon/(2D_q), & \text{if} \quad \sigma_y = 0, \\ 3L_{\nabla h}, & \text{if} \quad \sigma_y > 0. \end{cases} \tag{134}$$

Before presenting a unified first-order method for problem (130), we first present the modified optimal first-order method [44, Algorithm 1] in Algorithm 4 below for solving a general strongly-convex-strongly-concave minimax problem

$$\min_x \max_y \left\{\bar{h}(x,y) + p(x) - q(y)\right\}, \tag{135}$$

where $\bar{h}(x,y)$ is $\bar{\sigma}_x$-strongly-convex-$\bar{\sigma}_y$-strongly-concave and $L_{\nabla \bar{h}}$-smooth on $\operatorname{dom} p \times \operatorname{dom} q$ for some $\bar{\sigma}_x, \bar{\sigma}_y > 0$. The functions $\hat{h}$, a_x^k and a_y^k arising in Algorithm 4 are defined as follows:

$$\hat{h}(x,y) = \bar{h}(x,y) - \bar{\sigma}_x\|x\|^2/2 + \bar{\sigma}_y\|y\|^2/2,$$
$$a_x^k(x,y) = \nabla_x \hat{h}(x,y) + \bar{\sigma}_x(x - \bar{\sigma}_x^{-1} z_g^k)/2, \quad a_y^k(x,y) = -\nabla_y \hat{h}(x,y) + \bar{\sigma}_y y + \bar{\sigma}_x(y - y_g^k)/8,$$

where y_g^k and z_g^k are generated at iteration k of Algorithm 4 below.

Algorithm 4 A modified optimal first-order method for problem (135)

Input: $\tau > 0$, $\bar{z}^0 = z_f^0 \in -\bar{\sigma}_x \mathrm{dom}\, p$,[11] $\bar{y}^0 = y_f^0 \in \mathrm{dom}\, q$, $(z^0, y^0) = (\bar{z}^0, \bar{y}^0)$, $\bar{\alpha} = \min\left\{1, \sqrt{8\bar{\sigma}_y/\bar{\sigma}_x}\right\}$, $\eta_z = \bar{\sigma}_x/2$, $\eta_y = \min\{1/(2\bar{\sigma}_y), 4/(\bar{\alpha}\bar{\sigma}_x)\}$, $\beta_t = 2/(t+3)$, $\zeta = \left(2\sqrt{5}(1 + 8L_{\nabla\bar{h}}/\bar{\sigma}_x)\right)^{-1}$, $\gamma_x = \gamma_y = 8\bar{\sigma}_x^{-1}$, and $\hat{\zeta} = \min\{\bar{\sigma}_x, \bar{\sigma}_y\}/L_{\nabla\bar{h}}^2$.

for $k = 0, 1, 2, \ldots$ **do**
- $(z_g^k, y_g^k) = \bar{\alpha}(z^k, y^k) + (1-\bar{\alpha})(z_f^k, y_f^k)$.
- $(x^{k,-1}, y^{k,-1}) = (-\bar{\sigma}_x^{-1} z_g^k, y_g^k)$.
- $x^{k,0} = \mathrm{prox}_{\zeta\gamma_x p}(x^{k,-1} - \zeta\gamma_x a_x^k(x^{k,-1}, y^{k,-1}))$.
- $y^{k,0} = \mathrm{prox}_{\zeta\gamma_y q}(y^{k,-1} - \zeta\gamma_y a_y^k(x^{k,-1}, y^{k,-1}))$.
- $b_x^{k,0} = \frac{1}{\zeta\gamma_x}(x^{k,-1} - \zeta\gamma_x a_x^k(x^{k,-1}, y^{k,-1}) - x^{k,0})$.
- $b_y^{k,0} = \frac{1}{\zeta\gamma_y}(y^{k,-1} - \zeta\gamma_y a_y^k(x^{k,-1}, y^{k,-1}) - y^{k,0})$.
- $t = 0$.
- **while** $\gamma_x\|a_x^k(x^{k,t}, y^{k,t}) + b_x^{k,t}\|^2 + \gamma_y\|a_y^k(x^{k,t}, y^{k,t}) + b_y^{k,t}\|^2 > \gamma_x^{-1}\|x^{k,t} - x^{k,-1}\|^2 + \gamma_y^{-1}\|y^{k,t} - y^{k,-1}\|^2$ **do**
  - $x^{k,t+1/2} = x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x(a_x^k(x^{k,t}, y^{k,t}) + b_x^{k,t})$.
  - $y^{k,t+1/2} = y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y(a_y^k(x^{k,t}, y^{k,t}) + b_y^{k,t})$.
  - $x^{k,t+1} = \mathrm{prox}_{\zeta\gamma_x p}(x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x a_x^k(x^{k,t+1/2}, y^{k,t+1/2}))$.
  - $y^{k,t+1} = \mathrm{prox}_{\zeta\gamma_y q}(y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y a_y^k(x^{k,t+1/2}, y^{k,t+1/2}))$.
  - $b_x^{k,t+1} = \frac{1}{\zeta\gamma_x}(x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x a_x^k(x^{k,t+1/2}, y^{k,t+1/2}) - x^{k,t+1})$.
  - $b_y^{k,t+1} = \frac{1}{\zeta\gamma_y}(y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y a_y^k(x^{k,t+1/2}, y^{k,t+1/2}) - y^{k,t+1})$.
  - $t \leftarrow t + 1$.
- **end while**
- $(x_f^{k+1}, y_f^{k+1}) = (x^{k,t}, y^{k,t})$.
- $(z_f^{k+1}, w_f^{k+1}) = (\nabla_x\hat{h}(x_f^{k+1}, y_f^{k+1}) + b_x^{k,t}, -\nabla_y\hat{h}(x_f^{k+1}, y_f^{k+1}) + b_y^{k,t})$.
- $z^{k+1} = z^k + \eta_z\bar{\sigma}_x^{-1}(z_f^{k+1} - z^k) - \eta_z(x_f^{k+1} + \bar{\sigma}_x^{-1} z_f^{k+1})$.
- $y^{k+1} = y^k + \eta_y\bar{\sigma}_y(y_f^{k+1} - y^k) - \eta_y(w_f^{k+1} + \bar{\sigma}_y y_f^{k+1})$.
- $x^{k+1} = -\bar{\sigma}_x^{-1} z^{k+1}$.
- $\hat{x}^{k+1} = \mathrm{prox}_{\hat{\zeta}p}(x^{k+1} - \hat{\zeta}\nabla_x\bar{h}(x^{k+1}, y^{k+1}))$.
- $\hat{y}^{k+1} = \mathrm{prox}_{\hat{\zeta}q}(y^{k+1} + \hat{\zeta}\nabla_y\bar{h}(x^{k+1}, y^{k+1}))$.
- Terminate the algorithm and output $(\hat{x}^{k+1}, \hat{y}^{k+1})$ if

$$\|\hat{\zeta}^{-1}(x^{k+1} - \hat{x}^{k+1}, \hat{y}^{k+1} - y^{k+1}) - (\nabla\bar{h}(x^{k+1}, y^{k+1}) - \nabla\bar{h}(\hat{x}^{k+1}, \hat{y}^{k+1}))\| \le \tau.$$

end for

We now present a first-order method for finding an ϵ-primal-dual stationary point of problem (130) in Algorithm 5 below by unifying [44, Algorithm 2] and [45, Algorithm 1].

Algorithm 5 A first-order method for problem (130)

Input: $\epsilon > 0$, $\hat{\epsilon}_0 \in (0, \epsilon/2]$, $(\hat{x}^0, \hat{y}^0) \in \mathrm{dom}\, p \times \mathrm{dom}\, q$, $(x^0, y^0) = (\hat{x}^0, \hat{y}^0)$, and $\hat{\epsilon}_k = \hat{\epsilon}_0/(k+1)$.

for $k = 0, 1, 2, \ldots$ **do**
- Call Algorithm 4 with $\bar{h} \leftarrow h_k$, $\tau \leftarrow \hat{\epsilon}_k$, $\bar{\sigma}_x \leftarrow L_{\nabla h}$, $\bar{\sigma}_y \leftarrow \hat{\sigma}_y$, $L_{\nabla\bar{h}} \leftarrow \widehat{L}$, $\bar{z}^0 = z_f^0 \leftarrow -\bar{\sigma}_x x^k$, $\bar{y}^0 = y_f^0 \leftarrow y^k$, and denote its output by (x^{k+1}, y^{k+1}), where h_k is given in (133), $\hat{\sigma}_y$ and $\widehat{L}$ are given in (134).
- Terminate the algorithm and output $(x_\epsilon, y_\epsilon) = (x^{k+1}, y^{k+1})$ if

$$\|x^{k+1} - x^k\| \le \epsilon/(4L_{\nabla h}).$$

end for

The following theorem presents complexity results for Algorithm 5, which is a combination of [44,

[11] For convenience, $-\bar{\sigma}_x \mathrm{dom}\, p$ stands for the set $\{-\bar{\sigma}_x u | u \in \mathrm{dom}\, p\}$.

Theorem 2] for $\sigma_y = 0$ and [45, Theorem 1] for $\sigma_y > 0$.

Theorem 5 (Complexity of Algorithm 5). *Suppose that Assumption 3 holds. Let H^*, H, D_p, D_q, H_{low}, $\hat{\sigma}_y$ and $\widehat{L}$ be defined in (130), (131), (132) and (134), $L_{\nabla h}$ and σ_y be given in Assumption 3, ϵ, $\hat{\epsilon}_0$ and $\hat{x}^0$ be given in Algorithm 5, and*

$$
\begin{aligned}
\hat{\alpha} =& \ \min\left\{1, \sqrt{8\hat{\sigma}_y / L_{\nabla h}}\right\}, \\
\hat{\delta} =& \ (2 + \hat{\alpha}^{-1}) L_{\nabla h} D_p^2 + \max\left\{2\hat{\sigma}_y, \hat{\alpha} L_{\nabla h}/4\right\} D_q^2, \\
\widehat{T} =& \ \left\lceil 16(\max_y H(\hat{x}^0, y) - H^* + (\hat{\sigma}_y - \sigma_y) D_q^2/2) L_{\nabla h} \epsilon^{-2} + 32\hat{\epsilon}_0^2 (1 + \hat{\sigma}_y^{-2} L_{\nabla h}^2) \epsilon^{-2} - 1 \right\rceil_+, \\
\widehat{N} =& \ \left(\left\lceil 96\sqrt{2}\left(1 + 8\widehat{L} L_{\nabla h}^{-1}\right)\right\rceil + 2\right) \max\left\{2, \sqrt{L_{\nabla h}/(2\hat{\sigma}_y)}\right\} \\
& \times \left((\widehat{T} + 1) \left(\log \frac{4 \max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{1}{2\hat{\sigma}_y}, \frac{4}{\hat{\alpha} L_{\nabla h}}\right\}\right\} \left(\hat{\delta} + 2\hat{\alpha}^{-1}(H^* - H_{\text{low}} + (\hat{\sigma}_y - \sigma_y) D_q^2/2 + L_{\nabla h} D_p^2)\right)}{\left[\widehat{L}^2 / \min\{L_{\nabla h}, \hat{\sigma}_y\} + \widehat{L}\right]^{-2} \hat{\epsilon}_0^2} \right)_+ \right. \\
& \left. + \widehat{T} + 1 + 2\widehat{T} \log(\widehat{T} + 1) \right).
\end{aligned}
$$

Then Algorithm 5 terminates and outputs an ϵ-primal-dual stationary point (x_ϵ, y_ϵ) of (130) in at most $\widehat{T} + 1$ outer iterations that satisfies

$$
\max_y H(x_\epsilon, y) \le \max_y H(\hat{x}^0, y) + (\hat{\sigma}_y - \sigma_y) D_q^2/2 + 2\hat{\epsilon}_0^2 \left(L_{\nabla h}^{-1} + \hat{\sigma}_y^{-2} L_{\nabla h}\right).
$$

Moreover, the total number of evaluations of ∇h and proximal operators of p and q performed in Algorithm 5 is no more than $\widehat{N}$, respectively.