Sequential minimax optimization methods for bilevel optimization with strongly convex lower-level objective function

Zhaosong Lu * Sanyou Mei * Jin Zhang †

November 16, 2023

Abstract

In this paper we study a class of unconstrained and constrained bilevel optimization problems with strongly convex lower-level objective function. Specifically, we propose sequential minimax optimization methods for solving them, whose subproblems are nonconvex-strongly-concave unconstrained minimax problems and suitably solved by a first-order method developed by the authors that leverages the strong concavity structure. Under suitable assumptions, an *operation complexity* of $\mathcal{O}(\varepsilon^{-3}\log\varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^{-6}\log\varepsilon^{-1})$, measured by their fundamental operations, is established for the proposed methods for finding an ε-KKT solution of the unconstrained and constrained bilevel optimization problems respectively, which improves the previously best known operation complexity by a factor of $1/\varepsilon$.

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

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

1 Introduction

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

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

Bilevel optimization has widely been used in many areas, including adversarial training [44, 45, 57], continual learning [35], hyperparameter tuning [3, 15, 48], image reconstruction [8], meta-learning [4, 24, 51], neural architecture search [13, 33], reinforcement learning [19, 27], and Stackelberg games [59]. More applications about it can be found in [2, 7, 9, 10, 11, 54] and the references therein. Theoretical properties including optimality conditions of (1) have been extensively studied in the literature (e.g., see [11, 12, 42, 58, 62]).

Numerous methods have been developed for solving some special cases of (1). For example, constraint-based methods [18, 53], deterministic gradient-based methods [14, 15, 16, 20, 43, 50, 51], and stochastic gradient-based methods [6, 17, **?**, 21, 22, 25, 26, 28, 29, 61] 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 [5], and also numerical methods were developed in [30, 56, 32] 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, 41, 49]). Recently, difference-of-convex (DC) algorithms were developed in [63] for solving (1) with f being a DC function, and $\tilde{f}$, $\tilde{g}$ being convex functions. In addition, penalty type of methods

*Department of Industrial and Systems Engineering, University of Minnesota, USA (email: zhaosong@umn.edu, mei00035@umn.edu). The work of these authors was partially supported by NSF Award IIS-2211491.

†Department of Mathematics, SUSTech International Center for Mathematics, Southern University of Science and Technology, Shenzhen 518055, China (email: zhangj9@sustech.edu.cn). The work of this author was partially supported by National Natural Science Foundation of China Grant 12271161.

[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.

were proposed in [23, 38, 52] for solving (1). Notably, the paper [38] demonstrates for the first time that bilevel optimization can be approximately solved as minimax optimization. Specifically, it reformulates bilevel optimization as minimax optimization using a novel double penalty scheme and proposes a first-order penalty method with *complexity guarantees* for bilevel optimization via solving a single minimax problem. Subsequently, the work [39] proposed a first-order method for solving (1) via solving a sequence of minimax problems. Lately, a practically efficient multi-stage gradient descent and ascent algorithm (GDA) was developed in [60] 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). More discussion on algorithmic development for bilevel optimization can be found in [2, 7, 11, 34, 55, 58]) and the references therein.

In this paper, we consider problem (1) that 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 continuous on $\mathcal{X} \times \mathcal{Y}$, where $f_2 : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $\tilde{f}_2 : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ are proper closed convex functions, $\tilde{f}_1(x,\cdot)$ is σ-strongly-convex for any given $x \in \mathcal{X}$, and f_1, $\tilde{f}_1$ are respectively $L_{\nabla f_1}$- and $L_{\nabla \tilde{f}_1}$-smooth on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} := \operatorname{dom} f_2$ and $\mathcal{Y} := \operatorname{dom} \tilde{f}_2$.*

(ii) The proximal operator 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 specified in Assumption 1, existing methods are generally not applicable to problem (1) except the methods in [38, 39]. While the latter methods are applicable to problem (1), they do not exploit the strong convexity structure of $\tilde{f}_1(x,\cdot)$. Consequently, they may not be the most efficient methods for solving (1).

In this paper, we propose sequential minimax optimization (SMO) methods for solving problem (1) and its special case with $\tilde{g} \equiv 0$. Our approaches follow a similar framework as [38, Algorithm 2] and [39, Algorithm 1] respectively, but we enhance them by leveraging the strong concavity of $\tilde{f}_2(x,\cdot)$. As a result, our methods achieve a substantially improved operation complexity compared to [38, Algorithm 2] and [39, Algorithm 1].

The main contributions of this paper are summarized below.

- We propose an SMO method for solving an unconstrained bilevel optimization problem with strongly convex lower-level objective function. Under suitable assumptions, we show that this method achieves an operation complexity of $\mathcal{O}(\varepsilon^{-3}\log\varepsilon^{-1})$, measured by its fundamental operations, for finding an ε-KKT solution of the problem, which improves the operation complexity of [38, Algorithm 2] by a factor of $1/\varepsilon$.
- We propose an SMO method for solving a constrained bilevel optimization problem with strongly convex lower-level objective function. Under suitable assumptions, we show that this method achieves an operation complexity of $\mathcal{O}(\varepsilon^{-6}\log\varepsilon^{-1})$, measured by its fundamental operations, for finding an ε-KKT solution of the problem, which improves the operation complexity of [38, Algorithm 4] and [39, Algorithm 1] by a factor of $1/\varepsilon$.

The rest of this paper is organized as follows. In Subsection 1.1, we introduce some notation and terminology. In Section 2, we propose an SMO method for solving an unconstrained bilevel optimization problem with strongly convex lower-level objective function and study its complexity. In Section 3, we propose an SMO method for solving an constrained bilevel optimization problem with strongly convex lower-level objective function and study its complexity. Finally, we provide the proof of the main results in Section 4.

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 and Euclidean norm are denoted by $\langle\cdot,\cdot\rangle$ 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 $\|\phi(x) - \phi(x')\| \le L_\phi \|x - x'\|$ for all $x, x' \in S$. In addition, it is said to be $L_{\nabla\phi}$-*smooth* on S if $\|\nabla\phi(x) - \nabla\phi(x')\| \le L_{\nabla\phi}\|x - x'\|$ for all $x, x' \in S$. A function is said to be σ-strongly-convex if it is strongly convex with modulus $\sigma > 0$. For a closed convex function $p : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$,[2] the *proximal operator* associated with p is denoted by prox_p, that is,

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

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 $x \in \mathbb{R}^n$.

Finally, we introduce two types of approximate solutions for a general minimax problem

$$\Psi^* = \min_x \max_y \Psi(x, y), \tag{2}$$

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

Definition 1. *A point (x, y) is said to be a primal-dual stationary point of the minimax problem (2) 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 (2) if

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

2 Unconstrained bilevel optimization with strongly convex lower-level objective function

In this section, we consider problem (1) with $\tilde{g} \equiv 0$, which is an unconstrained bilevel optimization problem with strongly convex lower-level objective function given by

$$\begin{aligned} f^* = \min \quad & f(x, y) \\ \text{s.t.} \quad & y \in \operatorname*{Argmin}_z \tilde{f}(x, z). \end{aligned} \tag{3}$$

Assume that problem (3) has at least one optimal solution. In addition, f and $\tilde{f}$ satisfy Assumption 1.

A first-order penalty method was recently proposed in [38, Algorithm 2] for solving an unconstrained minimax problem in the form of (3) in which $\tilde{f}_1(x, \cdot)$ is however merely convex for any given $x \in \mathcal{X}$. Specifically, it applies a first-order method to the minimax problem

$$\min_{x,y} \max_z \left\{ P_\rho(x, y, z) := f(x, y) + \rho(\tilde{f}(x, y) - \tilde{f}(x, z)) \right\} \tag{4}$$

for a suitably chosen $\rho > 0$. While this method is applicable to problem (3), it does not exploit the strong convexity structure of $\tilde{f}_1(x, \cdot)$. Besides, it solves a single minimax problem with a large ρ, which is typically an ill-conditioned problem. Consequently, this method may not be the most efficient method for solving (3).

In what follows, we propose a sequential minimax optimization (SMO) method for solving problem (3). Our approach solves a sequence of nonconvex-strongly-concave minimax problems in the form of (4) using the recently developed first-order method [37, Algorithm 1] that leverages the strong convexity of $\tilde{f}_1(x, \cdot)$. As a result, our method achieves a substantially improved operation complexity compared to [38, Algorithm 2].

[2] For convenience, ∞ stands for $+\infty$.

Algorithm 1 A sequential minimax optimization method for problem (3)

Input: $\varepsilon, \tau \in (0,1]$, $\epsilon_0 \in (\tau\varepsilon, 1]$, $\epsilon_k = \epsilon_0\tau^k$, $\rho_k = \epsilon_k^{-1}$, $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$ with $\tilde{f}(x^0, y^0) \le \min_y \tilde{f}(x^0, y) + \tau\varepsilon$.

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

2: Call Algorithm 5 in Appendix B with $\epsilon \leftarrow \epsilon_k$, $\tilde{\epsilon}_0 \leftarrow \epsilon_k/2$, $\tilde{x}^0 \leftarrow (x^k, y^0)$, $\tilde{y}^0 \leftarrow y^0$, $\sigma_y \leftarrow \sigma\rho_k$ and $L_{\nabla h} \leftarrow L_{\nabla f_1} + 2\epsilon_k^{-1} L_{\nabla \tilde{f}_1}$ to find an ϵ_k-primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of problem (4) with $\rho = \rho_k$.

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

4: **end for**

Remark 1. *The initial point (x^0, y^0) of Algorithm 1 can be found by an additional procedure. Indeed, one can first choose any $x^0 \in \mathcal{X}$ and then apply an accelerated proximal gradient method [46] to the problem $\min_y \tilde{f}(x^0, y)$ for finding $y^0 \in \mathcal{Y}$ such that $\tilde{f}(x^0, y^0) \le \min_y \tilde{f}(x^0, y) + \tau\varepsilon$.*

To characterize the approximate solution found by Algorithm 1, we review a terminology called an ε-KKT solution of problem (3), which was introduced in [38, Definition 3].

Definition 2. *For any $\varepsilon > 0$, (x, y) is said to be an ε-KKT solution of problem (3) if there exists $(z, \rho) \in \mathbb{R}^m \times \mathbb{R}_+$ such that*

$$\begin{aligned}
&\operatorname{dist}\Big(0, \partial f(x,y) + \rho\partial\tilde{f}(x,y) - (\rho\nabla_x \tilde{f}(x,z); 0)\Big) \le \varepsilon, \quad \operatorname{dist}\big(0, \rho\partial_z \tilde{f}(x,z)\big) \le \varepsilon,\\
&\tilde{f}(x,y) - \min_{z'} \tilde{f}(x,z') \le \varepsilon.
\end{aligned}$$

We next study operation complexity of Algorithm 1. To proceed, we define

$$D_{\mathbf{x}} := \max\{\|u - v\| \big| u, v \in \mathcal{X}\}, \quad D_{\mathbf{y}} := \max\{\|u - v\| \big| u, v \in \mathcal{Y}\}, \tag{5}$$

$$\tilde{f}_{\rm hi} := \max\{\tilde{f}(x,y) | (x,y) \in \mathcal{X} \times \mathcal{Y}\}, \quad \tilde{f}_{\rm low} := \min\{\tilde{f}(x,y) | (x,y) \in \mathcal{X} \times \mathcal{Y}\}, \tag{6}$$

$$f_{\rm low} := \min\{f(x,y) | (x,y) \in \mathcal{X} \times \mathcal{Y}\}, \tag{7}$$

$$\widehat{K} := \lceil (\log\varepsilon - \log\epsilon_0)/\log\tau \rceil_+, \quad \widehat{\mathbb{K}} := \{0, 1, \ldots, \widehat{K} + 1\}, \quad \widehat{\mathbb{K}} - 1 = \{k - 1 | k \in \widehat{\mathbb{K}}\}. \tag{8}$$

By Assumption 1, one can observe that $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $\tilde{f}_{\rm hi}$, $\tilde{f}_{\rm low}$ and $f_{\rm low}$ are finite.

We are now ready to present a theorem regarding *operation complexity* of Algorithm 1, measured by the amount of evaluations of ∇f_1, $\nabla \tilde{f}_1$ and proximal operator of f_2 and $\tilde{f}_2$, for finding an $\mathcal{O}(\varepsilon)$-KKT solution of (3), whose proof is deferred to Subsection 4.1.

Theorem 1 (Complexity of Algorithm 1). *Suppose that Assumption 1 holds. Let f^*, f, $\tilde{f}$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $\tilde{f}_{\rm hi}$, $\tilde{f}_{\rm low}$, $f_{\rm low}$ and $\widehat{K}$ be defined in (3), (5), (6), (7) and (8), σ, $L_{\nabla f_1}$ and $L_{\nabla \tilde{f}_1}$ be given in Assumption 1, ε, τ, $\{\rho_k\}$, x^0 and y^0 be given in Algorithm 1, and*

$$\widehat{L} = L_{\nabla f_1} + 2L_{\nabla \tilde{f}_1}, \ \hat{\alpha} = \min\left\{1, \sqrt{8\sigma/\widehat{L}}\right\}, \tag{9}$$

$$\hat{\delta} = (2 + \hat{\alpha}^{-1})(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\widehat{L} + \max\left\{2\sigma, \widehat{L}/4\right\} D_{\mathbf{y}}^2, \tag{10}$$

$$\widehat{M} = \frac{16\max\left\{\frac{1}{4L_{\nabla \tilde{f}_1}}, \frac{2}{\hat{\alpha}L_{\nabla \tilde{f}_1}}\right\}}{\left[9\widehat{L}^2/\min\{2L_{\nabla \tilde{f}_1}, \sigma\} + 3\widehat{L}\right]^{-2}} \left(\hat{\delta} + 2\hat{\alpha}^{-1}(f^* - f_{\rm low} + \tilde{f}_{\rm hi} - \tilde{f}_{\rm low} + \widehat{L}(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2))\right), \tag{11}$$

$$\widehat{T} = \left\lceil 16(1 + f(x^0, y^0) - f_{\rm low})\widehat{L} + 8\sigma^{-2}\widehat{L}^2 + 7 \right\rceil_+, \tag{12}$$

$$\begin{aligned}
\widehat{N} = {} & 3397\max\left\{2, \sqrt{\widehat{L}/(2\sigma)}\right\} \widehat{T}\epsilon_0^{-3}(1 - \tau^3)^{-1}\\
& \times (\tau\varepsilon/\epsilon_0)^{-3}\left(14\widehat{K}\log(1/\tau) + 14\log(1/\epsilon_0) + 2(\log\widehat{M})_+ + 2 + 2\log(2\widehat{T})\right).
\end{aligned} \tag{13}$$

Then Algorithm 1 outputs an approximate solution $(x^{\widehat{K}+1}, y^{\widehat{K}+1})$ of problem (3) satisfying

$$\operatorname{dist}\Big(0, \partial f(x^{\widehat{K}+1}, y^{\widehat{K}+1}) + \rho_{\widehat{K}}\partial\tilde{f}(x^{\widehat{K}+1}, y^{\widehat{K}+1}) - (\rho_{\widehat{K}}\nabla_x\tilde{f}(x^{\widehat{K}+1}, z^{\widehat{K}+1}); 0)\Big) \le \varepsilon, \tag{14}$$

$$\operatorname{dist}\Big(0, \rho_{\widehat{K}}\partial\tilde{f}(x^{\widehat{K}+1}, z^{\widehat{K}+1})\Big) \le \varepsilon, \tag{15}$$

$$\tilde{f}(x^{\widehat{K}+1}, y^{\widehat{K}+1}) \le \min_z \tilde{f}(x^{\widehat{K}+1}, z) + \varepsilon\left(1 + f(x^0, y^0) - f_{\rm low} + \varepsilon^2(L_{\nabla f_1}^{-1} + \sigma^{-2}\widehat{L})/2\right), \tag{16}$$

after at most $\widehat{N}$ evaluations of ∇f_1, $\nabla\tilde{f}_1$ and proximal operator of f_2 and $\tilde{f}_2$, respectively.

Remark 2. *(i) One can observe from Theorem 1 that $\widehat{N} = \mathcal{O}(\varepsilon^{-3}\log\varepsilon^{-1})$. As a result, Algorithm 1 enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-3}\log\varepsilon^{-1})$, measured by the amount of evaluations of ∇f_1, $\nabla\tilde{f}_1$ and proximal operator of f_2 and $\tilde{f}_2$, for finding an $\mathcal{O}(\varepsilon)$-KKT solution $(x^{\widehat{K}+1}, y^{\widehat{K}+1})$ of (3) satisfying*

$$\operatorname{dist}\Big(0, \partial f(x^{\widehat{K}+1}, y^{\widehat{K}+1}) + \rho_{\widehat{K}}\partial\tilde{f}(x^{\widehat{K}+1}, y^{\widehat{K}+1}) - (\rho_{\widehat{K}}\nabla_x\tilde{f}(x^{\widehat{K}+1}, z^{\widehat{K}+1}); 0)\Big) \le \varepsilon,$$
$$\operatorname{dist}\Big(0, \rho_{\widehat{K}}\partial\tilde{f}(x^{\widehat{K}+1}, z^{\widehat{K}+1})\Big) \le \varepsilon, \qquad \tilde{f}(x^{\widehat{K}+1}, y^{\widehat{K}+1}) - \min_z \tilde{f}(x^{\widehat{K}+1}, z) = \mathcal{O}(\varepsilon),$$

where $z^{\widehat{K}+1}$ is given in Algorithm 1 and $\rho_{\widehat{K}} = \epsilon_0^{-1}\tau^{-\widehat{K}}$.

(ii) It shall be mentioned that an $\mathcal{O}(\varepsilon)$-KKT solution of (3) can be found by [38, Algorithm 2] with an operation complexity of $\mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$ (see [38, Theorem 2]). As a result, the operation complexity of Algorithm 1 improves that of [38, Algorithm 2] by a factor of $1/\epsilon$.

3 Constrained bilevel optimization with strongly convex lower-level objective function

In this section, we consider problem (1), which is a constrained bilevel optimization problem with strongly convex lower-level objective function satisfying Assumption 1. Recall from Assumption 1 that $\mathcal{X} = \operatorname{dom} f_2$ and $\mathcal{Y} = \operatorname{dom}\tilde{f}_2$. We now make some additional assumptions for problem (1).

Assumption 2. *(i) f and $\tilde{f}$ are L_f- and $L_{\tilde{f}}$-Lipschitz continuous on $\mathcal{X}\times\mathcal{Y}$, respectively.*

(ii) $\tilde{g}_i(x,\cdot)$ is convex and 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 $G := \min\{-\tilde{g}_i(x, \hat{z}_x) | x \in \mathcal{X},\ i = 1, \ldots, l\} > 0$.[3]

First-order penalty type of methods were recently proposed in [38, Algorithm 4] and [39, Algorithm 1] for solving constrained bilevel optimization in the form of (1) in which $\tilde{f}_1(x,\cdot)$ is however merely convex for any given $x \in \mathcal{X}$. Specifically, [38, Algorithm 4] solves constrained bilevel optimization by applying a first-order method to the minimax problem

$$\min_{x,y}\max_z P_{\rho,\mu}(x, y, z)$$

for suitably chosen $\rho, \mu > 0$, where

$$P_{\rho,\mu}(x,y,z) = \big(f_1(x,y) + \rho\tilde{f}_1(x,y) + \rho\mu\,\|[\tilde{g}(x,y)]_+\|^2 - \rho\tilde{f}_1(x,z) - \rho\mu\,\|[\tilde{g}(x,z)]_+\|^2\big) + \big(f_2(x) + \rho\tilde{f}_2(y) - \rho\tilde{f}_2(z)\big).$$

In addition, [39, Algorithm 1] solves constrained bilevel optimization by applying a first-order method to a sequence of minimax problems in the form of

$$\min_{x,y}\max_z \mathcal{L}(x, y, z, \lambda; \rho, \mu), \tag{17}$$

[3] The latter part of this assumption can be weakened to the one that the pointwise Slater's condition holds for the lower-level part of (1), that is, there exists $\hat{z}_x \in \mathcal{Y}$ such that $\tilde{g}(x, \hat{z}_x) < 0$ for each $x \in \mathcal{X}$. Indeed, if $G > 0$, Assumption 2(iii) clearly holds. Otherwise, one can solve the perturbed counterpart of (1) with $\tilde{g}(x,z)$ being replaced by $\tilde{g}(x,z) - \epsilon$ for some suitable $\epsilon > 0$ instead, which satisfies Assumption 2(iii).

where $\mathcal{L}$ is defined as

$$\mathcal{L}(x,y,z,\lambda;\rho,\mu) = f(x,y) + \rho\tilde{f}(x,y) + \frac{1}{2\mu}\|[\lambda+\mu\tilde{g}(x,y)]_+\|^2 - \rho\tilde{f}(x,z) - \frac{1}{2\mu}\|[\lambda+\mu\tilde{g}(x,z)]_+\|^2. \quad (18)$$

While these methods are applicable to problem (3), they do not exploit the strong convexity structure of $\tilde{f}_1(x,\cdot)$. Consequently, these methods may not be the most efficient methods for solving (3).

In what follows, we propose a sequential minimax optimization (SMO) method for solving problem (1). Our approach follows a similar framework as [39, Algorithm 3]. Specifically, at each iteration, our method finds an approximate primal-dual stationary point of a minimax problem in the form of (17). Yet, given that this problem is a nonconvex-strongly-concave unconstrained minimax problem, we solve it using the recently developed first-order method [37, Algorithm 1] that leverages the strong convexity of $\tilde{f}_1(x,\cdot)$. As a result, our method achieves a substantially improved operation complexity compared to [38, Algorithm 4] and [39, Algorithm 1].

To present our method, we define

$$\widetilde{\mathcal{L}}(x,z,,\lambda;\rho,\mu) := \tilde{f}(x,z) + \frac{1}{2\rho\mu}\|[\lambda+\mu\tilde{g}(x,z)]_+\|^2. \quad (19)$$

We are now ready to present our method for solving problem (1).

Algorithm 2 A sequential minimax optimization method for problem (1)

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

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

2: Call Algorithm 3 (see Appendix A) with $\Psi(\cdot)\leftarrow\widetilde{\mathcal{L}}(x^k,\cdot,\lambda^k;\rho_k,\mu_k)$, $\tilde{\epsilon}\leftarrow\eta_k/D_{\mathbf{y}}$, $\sigma_\phi\leftarrow\sigma$, $L_{\nabla\phi}\leftarrow\widetilde{L}_k$, $\tilde{x}^0\leftarrow y^k$ 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

$$\widetilde{\mathcal{L}}(x^k,y^k_{\rm init},\lambda^k;\rho_k,\mu_k) - \min_z\widetilde{\mathcal{L}}(x^k,z,\lambda^k;\rho_k,\mu_k)\le\eta_k, \quad (20)$$

where $\widetilde{\mathcal{L}}$ is given in (19) and

$$\widetilde{L}_k = L_{\nabla\tilde{f}_1} + \rho_k^{-1}(\mu_k L_{\tilde{g}}^2 + \mu_k\tilde{g}_{\rm hi}L_{\nabla\tilde{g}} + \|\lambda^k\|L_{\nabla\tilde{g}}). \quad (21)$$

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

$$\min_{x,y}\max_z\mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k), \quad (22)$$

where

$$L_k = L_{\nabla f_1} + 2\rho_k L_{\nabla\tilde{f}_1} + 2\mu_k L_{\tilde{g}}^2 + 2\mu_k\tilde{g}_{\rm hi}L_{\nabla\tilde{g}} + 2\|\lambda^k\|L_{\nabla\tilde{g}}. \quad (23)$$

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**

Remark 3. *(i) 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 Assumptions 1 and 2, Theorem 3 (see Appendix A), and* (90)*, it is not hard to observe that* $y^k_{\rm init}$ *satisfying* (20) *can be successfully found in step 2 of Algorithm 2 by applying Algorithm 3 to the problem* $\min_z\widetilde{L}(x^k,z,\lambda^k;\rho_k,\mu_k)$.

(ii) In view of Theorem 4 (see Appendix B), one can see that an ϵ_k*-primal-dual stationary point of* (22) *can be successfully found in step 3 of Algorithm 2 by applying Algorithm 5 to problem* (22)*. Consequently, Algorithm 2 is well-defined.*

We next study *iteration and operation complexity* of Algorithm 2. To characterize the approximate solution found by Algorithm 2, we first review a terminology called an ε-KKT solution of problem (1), which was introduced in [38, Definition 4].

Definition 3. *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

$$\tilde{f}^*(x) := \min\{\tilde{f}(x,z) | \tilde{g}(x,z) \le 0\}. \tag{24}$$

To proceed, recall that $\mathcal{X} = \operatorname{dom} f_2$ and $\mathcal{Y} = \operatorname{dom} \tilde{f}_2$. We define

$$\tilde{f}^*_{\rm hi} := \sup\{\tilde{f}^*(x) | x \in \mathcal{X}\}, \tag{25}$$

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

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

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 $f_{\rm hi}$ and $\tilde{g}_{\rm hi}$ are finite. Besides, $\tilde{f}^*_{\rm hi}$ is also finite (see Lemma 3).

We are now ready to present an *iteration and operation complexity* of Algorithm 2, measured by the amount of evaluations of ∇f_1, $\nabla\tilde{f}_1$, $\nabla\tilde{g}$ and proximal operator of f_2 and $\tilde{f}_2$, for finding an $\mathcal{O}(\varepsilon)$-KKT solution of (1), whose proof is deferred to Subsection 4.2.

Theorem 2 (iteration and operation complexity of Algorithm 2). *Suppose that Assumptions 1 and 2 hold. Let $\{(x_k, y_k, z_k, \lambda^k)\}_{k\in\mathbb{K}}$ be generated by Algorithm 2, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $\tilde{f}_{\rm low}$, $f_{\rm low}$, f^*, $\tilde{f}^*_{\rm hi}$, $f_{\rm hi}$, $\tilde{g}_{\rm hi}$ and K be defined in , (5), (6), (7), (1), (25), (26) and (27), 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 2. Let*

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

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

$$\alpha = \min\left\{1, \sqrt{8\sigma/L}\right\}, \quad \delta = (2+\alpha^{-1})(D^2_{\mathbf{x}} + D^2_{\mathbf{y}})L + \max\{1/D_{\mathbf{y}}, L/4\}D^2_{\mathbf{y}}, \tag{31}$$

$$\begin{aligned}
M = {}& 16\max\left\{1/(4L^2_{\tilde{g}}), 2/(\alpha L^2_{\tilde{g}})\right\}\left[9L^2/\min\{2L^2_{\tilde{g}}, \sigma\} + 3L\right]^2\\
&\times\left(\delta + 2\alpha^{-1}\Big(f^* - f_{\rm low} + \tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low} + L_{\tilde{f}}D_{\mathbf{y}} + 3\vartheta + \tilde{g}^2_{\rm hi} + L(D^2_{\mathbf{x}} + D^2_{\mathbf{y}})\Big)\right),
\end{aligned} \tag{32}$$

$$T = \left\lceil 16\left(f_{\rm hi} - f_{\rm low} + 1\right)L + 8(1 + \sigma^{-2}L^2)\right\rceil_+, \tag{33}$$

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

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

(i) Algorithm 2 terminates after $K+1$ outer iterations and outputs an approximate point (x^{K+1}, y^{K+1}) of problem (1) satisfying

$$\begin{aligned}
&\operatorname{dist}\Big(0, \partial f(x^{K+1}, y^{K+1}) + \rho_K\partial\tilde{f}(x^{K+1}, y^{K+1}) - \rho_K(\nabla_x\tilde{f}(x^{K+1}, z^{K+1}) + \nabla_x\tilde{g}(x^{K+1}, z^{K+1})\lambda^{K+1}_{\mathbf{z}}; 0)\\
&\qquad + \nabla\tilde{g}(x^{K+1}, y^{K+1})\lambda^{K+1}_{\mathbf{y}}\Big) \le \varepsilon,
\end{aligned} \tag{35}$$

$$\operatorname{dist}\Big(0, \rho_K(\partial_z\tilde{f}(x^{K+1}, z^{K+1}) + \nabla_z\tilde{g}(x^{K+1}, z^{K+1})\lambda^{K+1}_{\mathbf{z}})\Big) \le \varepsilon, \tag{36}$$

$$\|[\tilde{g}(x^{K+1}, z^{K+1})]_+\| \le 2\varepsilon^2G^{-1}(\epsilon_0 + L_{\tilde{f}})D_{\mathbf{y}}, \tag{37}$$

$$|\langle\lambda^{K+1}_{\mathbf{z}}, \tilde{g}(x^{K+1}, z^{K+1})\rangle| \le 2\varepsilon^2G^{-1}(\epsilon_0 + L_{\tilde{f}})D_{\mathbf{y}}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + L_{\tilde{f}})D_{\mathbf{y}}\}, \tag{38}$$

$$\|[\tilde{g}(x^{K+1},y^{K+1})]_+\| \le 2\varepsilon^2 G^{-1}(\epsilon_0+L_f+L_{\tilde{f}})D_{\mathbf{y}}, \tag{39}$$

$$|\langle\lambda_{\mathbf{y}}^{K+1},\tilde{g}(x^{K+1},z^{K+1})\rangle| \le 2\varepsilon G^{-1}(\epsilon_0+L_f+L_{\tilde{f}})D_{\mathbf{y}}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0+L_f+L_{\tilde{f}})D_{\mathbf{y}}\}, \tag{40}$$

$$\begin{aligned}|\tilde{f}(x^{K+1},y^{K+1})-\tilde{f}^*(x^{K+1})| \le \max\Big\{&2\varepsilon^2 G^{-2}L_{\tilde{f}}(\epsilon_0+L_f+L_{\tilde{f}})D_{\mathbf{y}}^2,\ \varepsilon^3\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0+L_{\tilde{f}})D_{\mathbf{y}}\}/2\\ &+\varepsilon\Big(f_{\rm hi}-f_{\rm low}+1+L_{\tilde{g}}^{-2}/4+\sigma^{-2}L/2\Big)\Big\}.\end{aligned} \tag{41}$$

(ii) The total number of evaluations of ∇f_1, $\nabla\tilde{f}_1$, $\nabla\tilde{g}$ and proximal operator of f_2 and $\tilde{f}_2$ performed in Algorithm 2 is N, respectively, where

$$\begin{aligned}N =&3397\max\Big\{2,\sqrt{L/(2\sigma)}\Big\}T(1-\tau^6)^{-1}\\ &\times(\tau\varepsilon)^{-6}\left(38K\log(1/\tau)+38\log(1/\epsilon_0)+2(\log M)_+ +2+2\log(2T)\right)\\ &+2(\tau\varepsilon)^{-1}(1-\tau)\left\lceil\sqrt{\frac{\widetilde{L}}{\sigma}}+1\right\rceil\max\Big\{1,\Big\lceil 2\log(2\widetilde{L}D_{\mathbf{y}}^2)+6K\log(1/\tau)-6\log\epsilon_0\Big\rceil\Big\}+K.\end{aligned} \tag{42}$$

Remark 4. *(i) One can observe from Theorem 2 that Algorithm 2 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 operator of f_2 and $\tilde{f}_2$, for finding an $\mathcal{O}(\varepsilon)$-KKT solution (x^{K+1},y^{K+1}) of (1) such that*

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

where $\tilde{f}^*$ is defined in (24), $\rho_K=(\epsilon_0\tau^K)^{-1}$, and $\lambda_{\mathbf{y}}^{K+1},\lambda_{\mathbf{z}}^{K+1}\in\mathbb{R}_+^l$ are given in (34).

(ii) It shall be mentioned that an $\mathcal{O}(\varepsilon)$-KKT solution of (1) can be found by [38, Algorithm 4] and [39, Algorithm 1] with an operation complexity of $\mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$ (see [38, Theorem 4] and [39, Theorem 1]). As a result, the operation complexity of Algorithm 2 improves that of [38, Algorithm 4] and [39, Algorithm 1] by a factor of $1/\epsilon$.

4 Proof of the main results

In this section we provide a proof of our main results presented in Sections 2 and 3, which are particularly Theorems 1 and 2.

4.1 Proof of the main results in Section 2

In this subsection we prove Theorem 1. We first establish two lemmas below.

Lemma 1. *Suppose that Assumption 1 holds. Let f^*, f, $\tilde{f}$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $\tilde{f}_{\rm hi}$, $\tilde{f}_{\rm low}$ and $f_{\rm low}$ be defined in (3), (5), (6) and (7), σ, $L_{\nabla f_1}$ and $L_{\nabla\tilde{f}_1}$ be given in Assumption 1, ϵ_k, ρ_k, x^0, y^0 and z_ϵ be given in*

Algorithm 1, and

$$\widehat{L}_k = L_{\nabla f_1} + 2\rho_k L_{\nabla \tilde{f}_1},\ \hat{\alpha}_k = \min\left\{1, \sqrt{8\sigma\rho_k/\widehat{L}_k}\right\}, \tag{43}$$

$$\hat{\delta}_k = (2+\hat{\alpha}_k^{-1})(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\widehat{L}_k + \max\left\{2\sigma\rho_k, \hat{\alpha}_k\widehat{L}_k/4\right\} D_{\mathbf{y}}^2, \tag{44}$$

$$\widehat{M}_k = \frac{16\max\left\{\frac{1}{2\widehat{L}_k}, \min\left\{\frac{1}{2\sigma\rho_k}, \frac{4}{\hat{\alpha}_k\widehat{L}_k}\right\}\right\}\left[\hat{\delta}_k + 2\hat{\alpha}_k^{-1}(f^* - f_{\rm low} + \rho_k(\tilde{f}_{\rm hi} - \tilde{f}_{\rm low}) + \widehat{L}_k(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2))\right]}{\left[9\widehat{L}_k^2/\min\{\widehat{L}_k, \sigma\rho_k\} + 3\widehat{L}_k\right]^{-2}\epsilon_k^2}, \tag{45}$$

$$\widehat{T}_k = \left\lceil 16(1+f(x^0,y^0)-f_{\rm low})\widehat{L}_k\epsilon_k^{-2} + 8\sigma^{-2}\rho_k^{-2}\widehat{L}_k^2 + 7\right\rceil_+, \tag{46}$$

$$\widehat{N}_k = 3397\max\left\{2, \sqrt{\widehat{L}_k/(2\sigma\rho_k)}\right\}\left((\widehat{T}_k+1)(\log\widehat{C}_k)_+ + \widehat{T}_k + 1 + 2\widehat{T}_k\log(\widehat{T}_k+1)\right). \tag{47}$$

Then for all $0 \le k \in \widehat{\mathbb{K}} - 1$, an ϵ_k-primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of problem (4) is successfully found at step 2 of Algorithm 1 that satisfies

$$P_{\rho_k}(x^{k+1}, y^{k+1}, z^{k+1}) \le 1 + f(x^0,y^0) + \epsilon_k^2(\widehat{L}_k^{-1} + \sigma^{-2}\rho_k^{-2}\widehat{L}_k)/2. \tag{48}$$

Moreover, the total number of evaluations of ∇f_1, $\nabla\tilde{f}_1$ and proximal operator of f_2 and $\tilde{f}_2$ performed at step 2 in iteration k of Algorithm 1 is no more than $\widehat{N}_k$.

Proof. Observe that problem (4) with $\rho = \rho_k$ can be viewed as

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

where $h(x,y,z) = f_1(x,y) + \rho_k\tilde{f}_1(x,y) - \rho_k\tilde{f}_1(x,z)$, $p(x,y) = f_2(x) + \rho_k\tilde{f}_2(y)$, and $q(z) = \rho_k\tilde{f}_2(z)$. Hence, problem (4) is in the form of (91) with $H = P_{\rho_k}$. By Assumption 1 and $\rho_k = \epsilon_k^{-1}$, one can see that $h(x,y,\cdot)$ is $\sigma\rho_k$-strongly-convex on $\mathcal{Y}$, and h is $\widehat{L}_k$-smooth on its domain, where $\widehat{L}_k$ is given in (43). Also, notice from Algorithm 1 that $\tilde{\epsilon}_0 = \epsilon_k/2$. Consequently, Algorithm 5 can be suitably applied to problem (4) with $\rho = \rho_k$ for finding an ϵ_k-stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of it.

In addition, notice from Algorithm 1 that $\tilde{f}(x^0,y^0) \le \min_y \tilde{f}(x^0,y) + \tau\varepsilon$ and $\tau\varepsilon \le \epsilon_k$ for all $0 \le k \in \widehat{\mathbb{K}} - 1$. Using this, (4), $\tau\varepsilon \le \epsilon_k$ and $\rho_k = \epsilon_k^{-1}$, we obtain

$$\max_z P_{\rho_k}(x^0,y^0,z) = f(x^0,y^0) + \rho_k(\tilde{f}(x^0,y^0) - \min_z \tilde{f}(x^0,z)) \le f(x^0,y^0) + \rho_k\epsilon_k = f(x^0,y^0) + 1. \tag{49}$$

By this and (96) with $H = P_{\rho_k}$, $\epsilon = \epsilon_k$, $\tilde{\epsilon}_0 = \epsilon_k/2$, $\tilde{x}^0 = (x^0,y^0)$, $D_q = D_{\mathbf{y}}$, $\sigma_y = \sigma\rho_k$ and $L_{\nabla h} = \widehat{L}_k$, one has

$$\begin{aligned} P_{\rho_k}(x^{k+1}, y^{k+1}, z^{k+1}) &\le \max_z P_{\rho_k}(x^0,y^0,z) + \epsilon_k^2(\widehat{L}_k^{-1} + \sigma^{-2}\rho_k^{-2}\widehat{L}_k)/2 \\ &\overset{(49)}{\le} 1 + f(x^0,y^0) + \epsilon_k^2(\widehat{L}_k^{-1} + \sigma^{-2}\rho_k^{-2}\widehat{L}_k)/2. \end{aligned}$$

Hence, (48) holds as desired.

We next show that at most $\widehat{N}_k$ evaluations of ∇f_1, $\nabla\tilde{f}_1$, and proximal operator of f_2 and $\tilde{f}_2$ are respectively performed in step 2 of Algorithm 1 at iteration k. Let (x^*, y^*) be an optimal solution of (3). It then follows that $f(x^*,y^*) = f^*$ and $\tilde{f}(x^*,y^*) = \min_z \tilde{f}(x^*,z)$. By these and the definition of P_ρ in (4), one has

$$\max_z P_{\rho_k}(x^*,y^*,z) = f(x^*,y^*) + \rho_k(\tilde{f}(x^*,y^*) - \min_z \tilde{f}(x^*,z)) = f(x^*,y^*) = f^*,$$

which implies that

$$\min_{x,y}\max_z P_{\rho_k}(x,y,z) \le \max_z P_{\rho_k}(x^*,y^*,z) = f^*. \tag{50}$$

Also, by (4), (6) and (7), one has

$$\min_{x,y}\max_z P_{\rho_k}(x,y,z) \overset{(4)}{=} \min_{x,y}\{f(x,y) + \rho_k(\tilde{f}(x,y) - \min_z \tilde{f}(x,z))\} \ge \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x,y) \overset{(7)}{=} f_{\rm low}, \tag{51}$$

$$\min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} P_{\rho_k}(x,y,z) \overset{(4)}{=} \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}}\{f(x,y) + \rho_k(\tilde{f}(x,y) - \tilde{f}(x,z))\} \overset{(6)(7)}{\ge} f_{\rm low} + \rho_k(\tilde{f}_{\rm low} - \tilde{f}_{\rm hi}). \tag{52}$$

For ease of the rest proof, let

$$H = P_{\rho_k}, \quad H^* = \min_{x,y}\max_{z} P_{\rho_k}(x,y,z), \quad H_{\text{low}} = \min\{P_{\rho_k}(x,y,z)|(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\}. \tag{53}$$

In view of these, (49), (50), (51) and (52), we obtain that

$$\max_z H(x^0,y^0,z) \overset{(49)}{\leq} f(x^0,y^0)+1, \qquad f_{\text{low}} \overset{(51)}{\leq} H^* \overset{(50)}{\leq} f^*, \qquad H_{\text{low}} \overset{(52)}{\geq} f_{\text{low}} + \rho_k(\tilde{f}_{\text{low}} - \tilde{f}_{\text{hi}}).$$

Using these and Theorem 4 with $\epsilon=\epsilon_k$, $\tilde{x}^0=(x^0,y^0)$, $D_p=\sqrt{D_{\mathbf{x}}^2+D_{\mathbf{y}}^2}$, $D_q=D_{\mathbf{y}}$, $\tilde{\epsilon}_0=\epsilon_k/2$, $\sigma_y=\sigma\rho_k$, $L_{\nabla h}=\widehat{L}_k$, $\tilde{\alpha}=\hat{\alpha}$, $\tilde{\delta}=\hat{\delta}$, and H, H^*, H_{low} given in (53), we can conclude that step 2 of Algorithm 1 at iteration k performs at most $\widehat{N}_k$ evaluations of ∇f_1, $\nabla\tilde{f}_1$ and proximal operator of f_2 and $\tilde{f}_2$ respectively for finding an ϵ_k-primal-dual stationary point $(x^{k+1},y^{k+1},z^{k+1})$ of problem (4) satisfying (48). □

Lemma 2. *Suppose that Assumption 1 holds and $(x^{k+1},y^{k+1},z^{k+1})$ is generated by Algorithm 1 for some $0\leq k\in\widehat{\mathbb{K}}$. Let P_ρ, $D_{\mathbf{y}}$, f_{low}, $\tilde{f}$ and ρ_k be given in (4), (5) and (7), respectively. Then we have*

$$\begin{aligned}
&\operatorname{dist}\Big(0, \partial f(x^{k+1},y^{k+1}) + \rho_k\partial\tilde{f}(x^{k+1},y^{k+1}) - (\rho_k\nabla_x\tilde{f}(x^{k+1},z^{k+1});0)\Big) \leq \epsilon_k,\\
&\operatorname{dist}\big(0,\rho_k\partial\tilde{f}(x^{k+1},z^{k+1})\big)\leq\epsilon_k,\\
&\tilde{f}(x^{k+1},y^{k+1}) \leq \min_z \tilde{f}(x^{k+1},z) + \rho_k^{-1}(\max_z P_{\rho_k}(x^{k+1},y^{k+1},z) - f_{\text{low}}).
\end{aligned}$$

Proof. Since $(x^{k+1},y^{k+1},z^{k+1})$ is an ϵ_k-stationary point of (4), it follows from Definition 1 that

$$\operatorname{dist}\big(0,\partial_{x,y}P_{\rho_k}(x^{k+1},y^{k+1},z^{k+1})\big)\leq\epsilon_k, \quad \operatorname{dist}\big(0,\partial_z P_{\rho_k}(x^{k+1},y^{k+1},z^{k+1})\big)\leq\epsilon_k.$$

Using these and the definition of P_ρ in (4), we have

$$\begin{aligned}
&\operatorname{dist}\Big(0, \partial f(x^{k+1},y^{k+1}) + \rho_k\partial\tilde{f}(x^{k+1},y^{k+1}) - (\rho_k\nabla_x\tilde{f}(x^{k+1},z^{k+1});0)\Big) \leq \epsilon_k,\\
&\operatorname{dist}\big(0,\rho_k\partial\tilde{f}(x^{k+1},z^{k+1})\big)\leq\epsilon_k.
\end{aligned}$$

In addition, by (4), we have

$$f(x^{k+1},y^{k+1}) + \rho_k(\tilde{f}(x^{k+1},y^{k+1}) - \min_z\tilde{f}(x^{k+1},z)) = \max_z P_{\rho_k}(x^{k+1},y^{k+1},z),$$

which along with (7) implies that

$$\tilde{f}(x^{k+1},y^{k+1}) - \min_z\tilde{f}(x^{k+1},z) \leq \rho_k^{-1}(\max_z P_{\rho_k}(x^{k+1},y^{k+1},z) - f_{\text{low}}).$$

This completes the proof of this lemma. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. Observe from the definition of $\widehat{K}$ in (8) and $\epsilon_k=\epsilon_0\tau^k$ that $\widehat{K}$ is the smallest nonnegative integer such that $\epsilon_K\leq\varepsilon$. Hence, Algorithm 1 terminates and outputs $(x^{\widehat{K}+1},y^{\widehat{K}+1})$ after $\widehat{K}+1$ outer iterations. In addition, by (9), (43), one has

$$L_{\nabla f_1} \overset{(43)}{\leq} \widehat{L}_k, \qquad 2\rho_k L_{\nabla\tilde{f}_1} \overset{(43)}{\leq} \widehat{L}_k \overset{(9)}{\leq} \rho_k\widehat{L}. \tag{54}$$

It follows from this, $\epsilon_{\widehat{K}}=\rho_k^{-1}\leq\varepsilon$, (48) and Lemma 2 that (14), (15) and (16) hold.

Let $\widehat{K}$ and $\widehat{N}$ be given in (8) and (13). Recall from Lemma 1 that the number of evaluations of ∇f_1, $\nabla\tilde{f}_1$, proximal operator of f_2 and $\tilde{f}_2$ performed by Algorithm 5 at iteration k of Algorithm 1 is at most $\widehat{N}_k$, where $\widehat{N}_k$ is given in (47). It then follows that the total number of evaluations of ∇f_1, $\nabla\tilde{f}_1$, and proximal operator of f_2 and $\tilde{f}_2$ performed in Algorithm 2 is no more than $\sum_{k=0}^{\widehat{K}}\widehat{N}_k$, respectively. As a result, to prove the last statement of this theorem, it suffices to show that $\sum_{k=0}^{\widehat{K}}\widehat{N}_k\leq\widehat{N}$.

To this end, using $\rho_k \geq 1 \geq \epsilon_k$, (9), (10), (11), (12), (43), (44), (45), (46) and (54), we obtain that

$$1 \geq \hat{\alpha}_k \geq \min\left\{1, \sqrt{8\sigma\rho_k/(\rho_k\widehat{L})}\right\} \geq \hat{\alpha}, \tag{55}$$

$$\hat{\delta}_k \leq (2+\hat{\alpha}^{-1})(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\rho_k\widehat{L} + \max\{2\sigma\rho_k, \rho_k\widehat{L}/4\}D_{\mathbf{y}}^2 \leq \rho_k\hat{\delta}, \tag{56}$$

$$\begin{aligned}\widehat{M}_k &\leq \frac{16\max\left\{1/(4\rho_k L_{\nabla\tilde{f}_1}), 2/(\hat{\alpha}\rho_k L_{\nabla\tilde{f}_1})\right\}}{\left[9\rho_k^2\widehat{L}^2/\min\{2\rho_k\widehat{L}_{\nabla\tilde{f}_1}, \sigma\rho_k\} + 3\rho_k\widehat{L}\right]^{-2}\epsilon_k^2} \\ &\quad \times \left(\rho_k\hat{\delta} + 2\hat{\alpha}^{-1}\Big(f^* - f_{\rm low} + \rho_k(\tilde{f}_{\rm hi} - \tilde{f}_{\rm low}) + \rho_k\widehat{L}(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\Big)\right) \qquad (57) \\ &\leq \frac{16\rho_k^{-1}\max\left\{1/(4L_{\nabla\tilde{f}_1}), 2/(\hat{\alpha}L_{\nabla\tilde{f}_1})\right\}}{\rho_k^{-2}\left[9\widehat{L}^2/\min\{2\widehat{L}_{\nabla\tilde{f}_1}, \sigma\} + 3\widehat{L}\right]^{-2}\epsilon_k^2}\rho_k\left(\hat{\delta} + 2\hat{\alpha}^{-1}\Big(f^* - f_{\rm low} + \tilde{f}_{\rm hi} - \tilde{f}_{\rm low} + \widehat{L}(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\Big)\right) \\ &= \epsilon_k^{-2}\rho_k^2\widehat{M}, \qquad (58)\end{aligned}$$

$$\widehat{T}_k \leq \left\lceil 16\left(1 + f(x^0, y^0) - f_{\rm low}\right)\epsilon_k^{-2}\rho_k\widehat{L} + 8\sigma^{-2}\widehat{L}^2 + 7\right\rceil_+ \leq \epsilon_k^{-2}\rho_k\widehat{T}, \tag{59}$$

where (55) follows from (9), (43) and (54); (56) is due to (10), (44), (55) and $\mu_k \geq 1 \geq \epsilon_k$; (57) is due to (45), (54), (55), (56) and $\epsilon_k \in (0,1]$; (58) follows from $\rho_k \geq 1 \geq \epsilon_k$ and (11); and (59) is due to (12) and (54). By the above inequalities, (47), (54), $T > 1$ and $\rho_k \geq 1 \geq \epsilon_k$, one has

$$\begin{aligned}\sum_{k=0}^{\widehat{K}}\widehat{N}_k &\leq \sum_{k=0}^{\widehat{K}} 3397\max\left\{2, \sqrt{\rho_k\widehat{L}/(2\sigma\rho_k)}\right\} \\ &\quad \times \left((\epsilon_k^{-2}\rho_k\widehat{T}+1)(\log(\epsilon_k^{-2}\rho_k^2\widehat{M}))_+ + \epsilon_k^{-2}\rho_k\widehat{T} + 1 + 2\epsilon_k^{-2}\rho_k\widehat{T}\log(\epsilon_k^{-2}\rho_k\widehat{T}+1)\right) \\ &\leq \sum_{k=0}^{\widehat{K}} 3397\max\left\{2, \sqrt{\widehat{L}/(2\sigma)}\right\}\epsilon_k^{-2}\rho_k\left((\widehat{T}+1)(\log(\epsilon_k^{-2}\rho_k^2\widehat{M}))_+ + \widehat{T} + 1 + 2\widehat{T}\log(\epsilon_k^{-2}\rho_k\widehat{T}+1)\right) \\ &\leq \sum_{k=0}^{\widehat{K}} 3397\max\left\{2, \sqrt{\widehat{L}/(2\sigma)}\right\}\epsilon_k^{-2}\rho_k\widehat{T}\left(2(\log(\epsilon_k^{-2}\rho_k^2\widehat{M}))_+ + 2 + 2\log(2\epsilon_k^{-2}\rho_k\widehat{T})\right) \\ &\leq \sum_{k=0}^{\widehat{K}} 3397\max\left\{2, \sqrt{\widehat{L}/(2\sigma)}\right\}\widehat{T}\epsilon_k^{-2}\rho_k\left(6\log\rho_k - 8\log\epsilon_k + (\log\widehat{M})_+ + 2 + 2\log(2\widehat{T})\right), \qquad (60)\end{aligned}$$

where the first inequality follows from $\epsilon_k \in (0,1]$, (47), (54), (58) and (59), and the second and third inequalities are due to the fact that $\rho_k \geq 1 \geq \epsilon_k$ and $T > 1$. By the definition of $\widehat{K}$ in (8), one has $\tau^{\widehat{K}} \geq \tau\varepsilon/\epsilon_0$. Also, notice from Algorithm 2 that $\rho_k = \epsilon_k^{-1} = (\epsilon_0\tau^k)^{-1}$. It then follows from these and (60) that

$$\begin{aligned}&\sum_{k=0}^{\widehat{K}}\widehat{N}_k \leq \sum_{k=0}^{\widehat{K}} 3397\max\left\{2, \sqrt{\widehat{L}/(2\sigma)}\right\}\widehat{T}\times\epsilon_k^{-3}\left(14\log(1/\epsilon_k) + 2(\log\widehat{M})_+ + 2 + 2\log(2\widehat{T})\right) \\ &= 3397\max\left\{2, \sqrt{\widehat{L}/(2\sigma)}\right\}\widehat{T}\sum_{k=0}^{\widehat{K}}\epsilon_0^{-3}\tau^{-3k}\left(14k\log(1/\tau) + 14\log(1/\epsilon_0) + 2(\log\widehat{M})_+ + 2 + 2\log(2\widehat{T})\right) \\ &\leq 3397\max\left\{2, \sqrt{\widehat{L}/(2\sigma)}\right\}\widehat{T}\sum_{k=0}^{\widehat{K}}\epsilon_0^{-3}\tau^{-3k}\left(14\widehat{K}\log(1/\tau) + 14\log(1/\epsilon_0) + 2(\log\widehat{M})_+ + 2 + 2\log(2\widehat{T})\right) \\ &\leq 3397\max\left\{2, \sqrt{\widehat{L}/(2\sigma)}\right\}\widehat{T}\epsilon_0^{-3}\tau^{-3\widehat{K}}(1-\tau^3)^{-1}\left(14\widehat{K}\log(1/\tau) + 14\log(1/\epsilon_0) + 2(\log\widehat{M})_+ + 2 + 2\log(2\widehat{T})\right) \\ &\leq 3397\max\left\{2, \sqrt{\widehat{L}/(2\sigma)}\right\}\widehat{T}\epsilon_0^{-3}(1-\tau^3)^{-1} \\ &\quad \times(\tau\varepsilon/\epsilon_0)^{-3}\left(14\widehat{K}\log(1/\tau) + 14\log(1/\epsilon_0) + 2(\log\widehat{M})_+ + 2 + 2\log(2\widehat{T})\right) \overset{(13)}{\leq} \widehat{N},\end{aligned}$$

where the third inequality is due to $\sum_{k=0}^{\widehat{K}} \tau^{-3k} \leq \tau^{-3\widehat{K}}/(1-\tau^3)$, the fourth inequality follows from $\tau^{\widehat{K}} \geq \tau\varepsilon/\epsilon_0$, and the last inequality is due to (13). □

4.2 Proof of the main results in Section 3

In this subsection we provide a proof of our main result presented in Section 3, which is particularly Theorem 2. Before proceeding, one can observe from (19) and (24) that

$$\min_z \widetilde{\mathcal{L}}(x,z,\lambda;\rho,\mu) \leq \tilde{f}^*(x) + \frac{\|\lambda\|^2}{2\rho\mu} \qquad \forall x \in \mathcal{X}, \lambda \in \mathbb{R}^l_+, \rho, \mu > 0, \tag{61}$$

which will be frequently used later.

We next establish several technical lemmas that will be used to prove Theorem 2 subsequently.

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 (5), (24), (25), and Assumptions 1 and 2, respectively. Then the following statements hold.*

(i) $\lambda^* \geq 0$ and $\|\lambda^*\| \leq G^{-1}L_{\tilde{f}}D_{\mathbf{y}}$ for all $\lambda^* \in \Lambda^*(x)$ and $x \in \mathcal{X}$, where $\Lambda^*(x)$ denotes the set of optimal Lagrangian multipliers of problem (24) for any $x \in \mathcal{X}$.

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

Proof. Its proof is similar to that of [39, Lemma 1] and thus omitted. □

Lemma 4. *Suppose that Assumptions 1 and 2 hold. Let $\mathbb{K}$ and ϑ be defined in (27) and (29), μ_k and ρ_k be given in Algorithm 2, and $\{\lambda^k\}_{k\in\mathbb{K}}$ be generated by Algorithm 2. Then we have*

$$\|\lambda^k\|^2 \leq 2\rho_k\mu_k\vartheta \qquad \forall 0 \leq k \in \mathbb{K}-1. \tag{62}$$

Proof. Its proof is similar to that of [39, Lemma 2] and thus omitted. □

Lemma 5. *Suppose that Assumptions 1 and 2 hold. Let $D_{\mathbf{y}}$, $\mathbb{K}$ and ϑ be defined in (5), (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 2. Suppose that $(x^{k+1}, y^{k+1}, z^{k+1}, \lambda^{k+1})$ is generated by Algorithm 2 for some $0 \leq k \in \mathbb{K}-1$ with*

$$\rho_k^{-1}\mu_k \geq 8G^{-2}\vartheta.$$

Then we have

$$\begin{aligned}
&\|[\tilde{g}(x^{k+1}, z^{k+1})]_+\| \leq \mu_k^{-1}\|\lambda^{k+1}\| \leq 2\mu_k^{-1}G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}},\\
&\|[\tilde{g}(x^{k+1}, y^{k+1})]_+\| \leq \mu_k^{-1}\|[\lambda^k + \mu_k\tilde{g}(x^{k+1}, y^{k+1})]_+\| \leq 2\mu_k^{-1}G^{-1}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})D_{\mathbf{y}}.
\end{aligned}$$

Proof. Its proof is similar to that of [39, Lemma 3] and thus omitted. □

Lemma 6. *Suppose that Assumptions 1 and 2 hold. Let $D_{\mathbf{y}}$, $\mathbb{K}$ and ϑ be defined in (5), (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 2. Suppose that $(x^{k+1}, y^{k+1}, z^{k+1}, \lambda^{k+1})$ is generated by Algorithm 2 for some $0 \leq k \in \mathbb{K}-1$ with*

$$\rho_k^{-1}\mu_k \geq 8\tau^{-2}G^{-2}\vartheta. \tag{63}$$

Let

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

Then we have

$$\begin{aligned}
&\mathrm{dist}\Big(0, \partial f(x^{k+1}, y^{k+1}) + \rho_k\partial\tilde{f}(x^{k+1}, y^{k+1}) - \rho_k\big(\nabla_x\tilde{f}(x^{k+1}, z^{k+1}) + \nabla_x\tilde{g}(x^{k+1}, z^{k+1})\lambda^{k+1}_{\mathbf{z}}; 0\big)\\
&\qquad + \nabla\tilde{g}(x^{k+1}, y^{k+1})\lambda^{k+1}_{\mathbf{y}}\Big) \leq \epsilon_k,\\
&\mathrm{dist}\left(0, \rho_k(\partial_z\tilde{f}(x^{k+1}, z^{k+1}) + \nabla_z\tilde{g}(x^{k+1}, z^{k+1})\lambda^{k+1}_{\mathbf{z}})\right) \leq \epsilon_k,\\
&\|[\tilde{g}(x^{k+1}, z^{k+1})]_+\| \leq 2\mu_k^{-1}G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}},\\
&|\langle\lambda^{k+1}_{\mathbf{z}}, \tilde{g}(x^{k+1}, z^{k+1})\rangle| \leq 2\rho_k^{-1}\mu_k^{-1}G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\},\\
&\|[\tilde{g}(x^{k+1}, y^{k+1})]_+\| \leq 2\mu_k^{-1}G^{-1}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})D_{\mathbf{y}},\\
&|\langle\lambda^{k+1}_{\mathbf{y}}, \tilde{g}(x^{k+1}, y^{k+1})\rangle| \leq 2\mu_k^{-1}G^{-1}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\}.
\end{aligned} \tag{64}$$

Proof. Its proof is similar to that of [39, Lemma 4] and thus omitted. □

Lemma 7. *Suppose that Assumptions 1 and 2 hold. Let* $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $\tilde{f}_{\rm low}$, $f_{\rm low}$, f^*, L_k, $\tilde{f}^*_{\rm hi}$, $f_{\rm hi}$, $\tilde{g}_{\rm hi}$, $\mathbb{K}$ *and* ϑ *be defined in* (5), (6), (7), (1), (23), (25), (26), (27) *and* (29), σ, $L_{\tilde{f}}$ *be given in Assumption 1,* ϵ_k, ρ_k, μ_k *and* η_k *be given in Algorithm 2, and*

$$\alpha_k = \min\left\{1, \sqrt{8\sigma\rho_k/L_k}\right\}, \tag{65}$$

$$\delta_k = (2+\alpha_k^{-1})(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)L_k + \max\left\{2\sigma\rho_k, \alpha_k L_k/4\right\} D_{\mathbf{y}}^2, \tag{66}$$

$$\begin{aligned} M_k = {} & \frac{16\max\left\{1/(2L_k), \min\left\{1/(2\sigma\rho_k), 4/(\alpha_k L_k)\right\}\right\}}{\left[9L_k^2/\min\{L_k, \sigma\rho_k\} + 3L_k\right]^{-2}\epsilon_k^2} \\ & \times \left(\delta_k + 2\alpha_k^{-1}\left(f^* - f_{\rm low} + \rho_k(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}) + \rho_k L_{\tilde{f}} D_{\mathbf{y}} + 3\rho_k\vartheta + \mu_k\tilde{g}_{\rm hi}^2 + L_k(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\right)\right), \end{aligned} \tag{67}$$

$$T_k = \left\lceil 16\left(f_{\rm hi} - f_{\rm low} + \rho_k\eta_k\right)L_k\epsilon_k^{-2} + 8\sigma^{-2}\rho_k^{-2}L_k^2 + 7\right\rceil_+, \tag{68}$$

$$N_k = 3397\max\left\{2, \sqrt{L_k/(2\sigma\rho_k)}\right\}\left((T_k+1)(\log M_k)_+ + T_k + 1 + 2T_k\log(T_k+1)\right). \tag{69}$$

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* (22) *is successfully found at step 3 of Algorithm 2 that satisfies*

$$\max_z \mathcal{L}(x^{k+1}, y^{k+1}, z, \lambda^k; \rho_k, \mu_k) \le f_{\rm hi} + \rho_k\eta_k + \frac{1}{2\epsilon_k^2}\left(L_k^{-1} + \sigma^{-2}\rho_k^{-2}L_k\right). \tag{70}$$

Moreover, the total number of evaluations of ∇f_1, $\nabla\tilde{f}_1$, $\nabla\tilde{g}$ *and proximal operator of* f_2 *and* $\tilde{f}_2$ *performed at step 3 in iteration* k *of Algorithm 2 is no more than* N_k, *respectively.*

Proof. Observe from (18) and Assumption 1 that problem (22) can be viewed as

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

with

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

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 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-convex 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 (23). Consequently, it follows from Theorem 4 (see Appendix B) that an ϵ_k-primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of problem (22) is successfully found by Algorithm 5 at step 3 of Algorithm 2.

In addition, by (7), (18), (19), one has

$$\begin{aligned} \min_{x,y}\max_z \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k) & \overset{(18)(19)}{=} \min_{x,y}\left\{f(x,y) + \rho_k\widetilde{L}(x,y,\lambda^k;\rho_k,\mu_k) - \min_z \rho_k\widetilde{L}(x,z,\lambda^k;\rho_k,\mu_k)\right\} \\ & \ge \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x,y) \overset{(7)}{=} f_{\rm low}. \end{aligned} \tag{71}$$

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 (24), respectively. Using these, (6), (18), (19), (25) and (62), we obtain that

$$\begin{aligned} & \min_{x,y}\max_z \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k) \le \max_z \mathcal{L}(x^*,y^*,z,\lambda^k;\rho_k,\mu_k) \\ & \overset{(18)(19)}{=} f(x^*,y^*) + \rho_k\tilde{f}(x^*,y^*) + \frac{1}{2\mu_k}\|[\lambda^k + \mu_k\tilde{g}(x^*,y^*)]_+\|^2 - \min_z \rho_k\widetilde{L}(x^*,z,\lambda^k;\rho_k,\mu_k) \\ & \le f^* + \rho_k\tilde{f}^*(x^*) + \frac{1}{2\mu_k}\|\lambda^k\|^2 - \min_z\left\{\rho_k\tilde{f}(x^*,z) + \frac{1}{2\mu_k}\|[\lambda^k + \mu_k\tilde{g}(x^*,z)]_+\|^2\right\} \\ & \overset{(6)(25)}{\le} f^* + \rho_k(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}) + \frac{1}{2\mu_k}\|\lambda^k\|^2 \overset{(62)}{\le} f^* + \rho_k(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}) + \rho_k\vartheta, \end{aligned} \tag{72}$$

where the second inequality is due to $\tilde{f}(x^*, y^*) = \tilde{f}^*(x^*)$, $\tilde{g}(x^*, y^*) \le 0$ and (19). Also, by (5), (6), (18), (26) and (62), one has

$$\begin{aligned}
&\min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \mathcal{L}(x, y, z, \lambda^k; \rho_k, \mu_k) \\
&\overset{(18)}{\ge} \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \left\{ f(x, y) + \rho_k(\tilde{f}(x, y) - \tilde{f}(x, z)) - \frac{1}{2\mu_k}\|[\lambda^k + \mu_k\tilde{g}(x, z)]_+\|^2 \right\} \\
&\ge \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \left\{ f(x, y) - \rho_k L_{\tilde{f}}\|y - z\| - \frac{1}{2\mu_k}\left(\|\lambda^k\| + \mu_k\|[\tilde{g}(x, z)]_+\|\right)^2 \right\} \\
&\ge \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \left\{ f(x, y) - \rho_k L_{\tilde{f}}\|y - z\| - \frac{1}{\mu_k}\|\lambda^k\|^2 - \mu_k\|[\tilde{g}(x, z)]_+\|^2 \right\} \\
&\ge f_{\rm low} - \rho_k L_{\tilde{f}} D_{\mathbf{y}} - 2\rho_k\vartheta - \mu_k\tilde{g}_{\rm hi}^2,
\end{aligned} \tag{73}$$

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 (5), (6), (26) and (62). Notice from step 2 of Algorithm 2 that $y^k_{\rm init}$ is an approximate solution of $\min_z \widetilde{L}(x^k, z, \lambda^k; \rho_k, \mu_k)$ satisfying (20). It then follows from (18), (19), (20) and (26) that

$$\begin{aligned}
\max_z \mathcal{L}(x^k, y^k_{\rm init}, z, \lambda^k; \rho_k, \mu_k) &\overset{(18)(19)}{=} f(x^k, y^k_{\rm init}) + \rho_k\left(\widetilde{\mathcal{L}}(x^k, y^k_{\rm init}, \lambda^k; \rho_k, \mu_k) - \min_z \widetilde{\mathcal{L}}(x^k, z, \lambda^k; \rho_k, \mu_k)\right) \\
&\overset{(20)}{\le} f(x^k, y^k_{\rm init}) + \rho_k\eta_k \overset{(26)}{\le} f_{\rm hi} + \rho_k\eta_k.
\end{aligned} \tag{74}$$

To complete the rest of the proof, let

$$H(x, y, z) = \mathcal{L}(x, y, z, \lambda^k; \rho_k, \mu_k), \quad H^* = \min_{x,y}\max_z \mathcal{L}(x, y, z, \lambda^k; \rho_k, \mu_k), \tag{75}$$

$$H_{\rm low} = \min\left\{\mathcal{L}(x, y, z, \lambda^k; \rho_k, \mu_k) | (x, y, z) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\right\}. \tag{76}$$

In view of these, (71), (72), (73) and (74), we obtain that

$$\begin{aligned}
&\max_z H(x^k, y^k_{\rm init}, z) \overset{(74)}{\le} f_{\rm hi} + \rho_k\eta_k, \\
&f_{\rm low} \overset{(71)}{\le} H^* \overset{(72)}{\le} f^* + \rho_k(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}) + \rho_k\vartheta, \\
&H_{\rm low} \overset{(73)}{\ge} f_{\rm low} - \rho_k L_{\tilde{f}} D_{\mathbf{y}} - 2\rho_k\vartheta - \mu_k\tilde{g}_{\rm hi}^2.
\end{aligned}$$

Using these and Theorem 4 (see Appendix B) with $\tilde{x}^0 = (x^k, y^k_{\rm init})$, $\epsilon = \epsilon_k$, $\tilde{\epsilon}_0 = \epsilon_k/2$, $\sigma_y = \sigma\rho_k$, $L_{\nabla h} = L_k$, $\tilde{\alpha} = \alpha_k$, $\tilde{\delta} = \delta_k$, $D_p = \sqrt{D^2_{\mathbf{x}} + D^2_{\mathbf{y}}}$, $D_q = D_{\mathbf{y}}$, and H, H^*, $H_{\rm low}$ given in (75) and (76), we can conclude that the ϵ_k-primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of problem (22) found at step 3 of Algorithm 2 satisfies (70). Moreover, the total number of evaluations of ∇f_1, $\nabla\tilde{f}_1$, $\nabla\tilde{g}$ and proximal operator of f_2 and $\tilde{f}_2$ performed by Algorithm 5 at step 3 of Algorithm 2 is no more than N_k, respectively. □

Lemma 8. *Suppose that Assumptions 1 and 2 hold. Let $D_{\mathbf{y}}$, $f_{\rm low}$, $\tilde{f}^*$, L_k, $f_{\rm hi}$ and $\mathbb{K}$ be defined in (5), (7), (24), (23) and (27), σ, L_f, $L_{\tilde{f}}$ and G be given in Assumptions 1 and 2, and ϵ_k, ρ_k, μ_k, η_k and λ^0 be given in Algorithm 2. Suppose that $(x^{k+1}, y^{k+1}, \lambda^{k+1})$ is generated by Algorithm 2 for some $0 \le k \in \mathbb{K} - 1$ satisfying (63). Then we have*

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

Proof. We first claim that

$$\|\lambda^k\| \le \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\}. \tag{77}$$

Indeed, (77) clearly holds if $k = 0$. We now assume that $k > 0$. Notice from Algorithm 2 that $\mu_{k-1} = \tau^3\mu_k$ and $\rho_{k-1} = \tau\rho_k$, which along with (63) imply that $\rho_{k-1}^{-1}\mu_{k-1} \geq 8G^{-2}\vartheta$. By this and Lemma 4 with k replaced by $k-1$, one can conclude that $\|\lambda^k\| \leq 2G^{-1}(\epsilon_0 + \rho_{k-1}L_{\tilde{f}})D_{\mathbf{y}}$. This together with $\rho_{k-1} < \rho_k$ implies that (77) holds as desired.

Using this, (7), (18), (19), (61) and (77), we have

$$\begin{aligned}
&\max_z \mathcal{L}(x^{k+1}, y^{k+1}, z, \lambda^k; \rho_k, \mu_k)\\
&\overset{(18)(19)}{=} f(x^{k+1}, y^{k+1}) + \rho_k\tilde{f}(x^{k+1}, y^{k+1}) + \frac{1}{2\mu_k}\|[\lambda^k + \mu_k\tilde{g}(x^{k+1}, y^{k+1})]_+\|^2 - \min_z \rho_k\widetilde{\mathcal{L}}(x^{k+1}, z, \lambda^k; \rho_k, \mu_k)\\
&\geq f(x^{k+1}, y^{k+1}) + \rho_k\tilde{f}(x^{k+1}, y^{k+1}) - \min_z \rho_k\widetilde{\mathcal{L}}(x^{k+1}, z, \lambda^k; \rho_k, \mu_k)\\
&\overset{(7)(61)}{\geq} f_{\rm low} + \rho_k\big(\tilde{f}(x^{k+1}, y^{k+1}) - \tilde{f}^*(x^{k+1})\big) - \frac{1}{2\mu_k}\|\lambda^k\|^2\\
&\overset{(77)}{\geq} f_{\rm low} + \rho_k\big(\tilde{f}(x^{k+1}, y^{k+1}) - \tilde{f}^*(x^{k+1})\big) - \mu_k^{-1}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\}/2.
\end{aligned}$$

This together with (70) implies that

$$\begin{aligned}
\tilde{f}(x^{k+1}, y^{k+1}) - \tilde{f}^*(x^{k+1}) \ \leq\ & \rho_k^{-1}\left(f_{\rm hi} - f_{\rm low} + \rho_k\eta_k + \frac{1}{2\epsilon_k^2}\left(L_k^{-1} + \sigma^{-2}\rho_k^{-2}L_k\right)\right)\\
&+ \rho_k^{-1}\mu_k^{-1}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\}/2. \qquad (78)
\end{aligned}$$

On the other hand, let $\lambda^* \in \mathbb{R}^l_+$ be an optimal Lagrangian multiplier of problem (24) with $x = x^{k+1}$. It then follows from Lemma 3(i) that $\|\lambda^*\| \leq G^{-1}L_{\tilde{f}}D_{\mathbf{y}}$. Using these, Lemma 2, (24) and (64), we have

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

The conclusion of this lemma then follows from this and (78). □

Lemma 9. *Suppose that Assumptions 1 and 2 hold. Let $D_{\mathbf{y}}$, $\widetilde{L}_k$ and $\mathbb{K}$ be defined in* (5), (21) *and* (27), *σ be given in Assumption 1, η_k be given in Algorithm 2, and*

$$\widetilde{N}_k = 2\left\lceil\sqrt{\widetilde{L}_k\sigma^{-1}}\right\rceil\max\left\{1, \left\lceil 2\log(2\eta_k^{-1}\widetilde{L}_k D_{\mathbf{y}}^2)\right\rceil\right\} + 1. \qquad (79)$$

Then for all $0 \leq k \in \mathbb{K} - 1$, $y_{\rm init}^k$ satisfying (20) *is found at step 3 of Algorithm 2 by Algorithm 3 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 (19) and Algorithm 2 that $y_{\rm init}^k$ satisfying (20) is found by Algorithm 3 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 and (28), one can see that ϕ is σ-strongly-convex and $\widetilde{L}_k$-smooth on $\operatorname{dom} P$, where $\widetilde{L}_k$ is given in (21). It then follows from this and Theorem 3 (see Appendix A) and (90) with $\tilde{\epsilon} = \eta_k/D_{\mathbf{y}}$, $D_P = D_{\mathbf{y}}$, $\sigma_\phi = \sigma$ and $L_{\nabla\phi} = \widetilde{L}_k$ that Algorithm 3 finds $y_{\rm init}^k$ satisfying (20) in no more than $\widetilde{T}_k$ iterations, where

$$\widetilde{T}_k = \left\lceil\sqrt{\widetilde{L}_k\sigma^{-1}}\right\rceil\max\left\{1, \left\lceil 2\log(2\eta_k^{-1}\widetilde{L}_k D_{\mathbf{y}}^2)\right\rceil\right\}.$$

Notice that the first step of Algorithm 3 requires one evaluation of $\nabla\phi$ and the proximal operator of P, respectively, and each iteration of Algorithm 3 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) 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 2 terminates and outputs (x^{K+1}, y^{K+1}) after $K+1$ outer iterations. Also, one can see from Algorithm 2 that

$$\rho_K = \epsilon_K^{-1}, \quad \mu_K = \epsilon_K^{-3}, \quad \eta_K = \epsilon_K. \tag{80}$$

Moreover, notice from the assumption of Theorem 2 that $\varepsilon^{-2} - 8\tau^{-2}G^{-2}\vartheta \ge 0$. It then follows from this and (80) that

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

which implies that (63) holds for $k = K$. In addition, by (23), (30), (62) and $\mu_k \ge \rho_k \ge 1$, one has that for all $0 \le k \in \mathbb{K} - 1$,

$$\begin{aligned} 2\mu_k L_{\tilde g}^2 \le L_k &\overset{(23)}{=} L_{\nabla f_1} + 2\rho_k L_{\nabla \tilde f_1} + 2\mu_k L_{\tilde g}^2 + 2\mu_k \tilde g_{\rm hi} L_{\nabla \tilde g} + 2\|\lambda^k\| L_{\nabla \tilde g} \\ &\overset{(62)}{\le} L_{\nabla f_1} + 2\rho_k L_{\nabla \tilde f_1} + 2\mu_k L_{\tilde g}^2 + 2\mu_k \tilde g_{\rm hi} L_{\nabla \tilde g} + 2\sqrt{2\rho_k\mu_k\vartheta} L_{\nabla \tilde g} \le \mu_k L. \end{aligned} \tag{81}$$

It then follows from $\epsilon_K \le \varepsilon$, (80) and Lemmas 6 and 8 that (35)-(41) hold.

(ii) Let K and N be given in (27) and (42). Recall from Lemmas 7 and 9 that the number of evaluations of ∇f_1, $\nabla \tilde f_1$, $\nabla \tilde g$, proximal operator of f_2 and $\tilde f_2$ performed by Algorithms 3 and 5 at iteration k of Algorithm 2 is at most $N_k + \widetilde N_k$, where N_k and $\widetilde N_k$ are given in (69) and (79), 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 operator of f_2 and $\tilde f_2$ performed in Algorithm 2 is no more than $\sum_{k=0}^K (N_k + \widetilde N_k)$, respectively. As a result, to prove statement (ii) of this theorem, it suffices to show that $\sum_{k=0}^K (N_k + \widetilde N_k) \le N$.

To this end, using $\mu_k \ge \rho_k \ge 1 \ge \epsilon_k$, (31), (32), (33), (65), (66), (67), (68) and (81), we obtain that

$$1 \ge \alpha_k \ge \min\left\{1, \sqrt{8\sigma\rho_k/(\mu_k L)}\right\} \ge \rho_k^{1/2}\mu_k^{-1/2}\alpha, \tag{82}$$

$$\delta_k \le (2 + \rho_k^{-1/2}\mu_k^{1/2}\alpha^{-1})(D_{\mathbf x}^2 + D_{\mathbf y}^2)\mu_k L + \max\{2\sigma\rho_k, \mu_k L/4\} D_{\mathbf y}^2 \le \rho_k^{-1/2}\mu_k^{3/2}\delta, \tag{83}$$

$$\begin{aligned} M_k \le{}& \frac{16\max\left\{1/(4\mu_k L_{\tilde g}^2), 2/(\rho_k^{1/2}\mu_k^{-1/2}\alpha\mu_k L_{\tilde g}^2)\right\}}{\left[9\mu_k^2 L^2/\min\{2\mu_k L_{\tilde g}^2, \sigma\rho_k\} + 3\mu_k L\right]^{-2}\epsilon_k^2} \times \Bigg(\rho_k^{-1/2}\mu_k^{3/2}\delta \\ &+ 2\rho_k^{-1/2}\mu_k^{1/2}\alpha^{-1}\Big(f^* - f_{\rm low} + \rho_k(\tilde f_{\rm hi}^* - \tilde f_{\rm low}) + \rho_k L_{\tilde f} D_{\mathbf y} + 3\rho_k\vartheta + \mu_k \tilde g_{\rm hi}^2 + \mu_k L(D_{\mathbf x}^2 + D_{\mathbf y}^2)\Big)\Bigg) \end{aligned} \tag{84}$$

$$\begin{aligned} \le{}& \frac{16\rho_k^{-1/2}\mu_k^{-1/2}\max\left\{1/(4L_{\tilde g}^2), 2/(\alpha L_{\tilde g}^2)\right\}}{\rho_k^2\mu_k^{-4}\left[9L^2/\min\{2L_{\tilde g}^2, \sigma\} + 3L\right]^{-2}\epsilon_k^2} \times \rho_k^{-1/2}\mu_k^{3/2} \\ &\times \left(\delta + 2\alpha^{-1}\Big(f^* - f_{\rm low} + \tilde f_{\rm hi}^* - \tilde f_{\rm low} + L_{\tilde f} D_{\mathbf y} + 3\vartheta + \tilde g_{\rm hi}^2 + L(D_{\mathbf x}^2 + D_{\mathbf y}^2)\Big)\right) = \epsilon_k^{-2}\rho_k^{-3}\mu_k^5 M, \end{aligned} \tag{85}$$

$$T_k \le \left\lceil 16\left(f_{\rm hi} - f_{\rm low} + \rho_k\eta_k\right)\epsilon_k^{-2}\mu_k L + 8\sigma^{-2}\rho_k^{-2}\mu_k^2 L^2) + 7\right\rceil_+ \le \epsilon_k^{-2}\mu_k T, \tag{86}$$

where (82) follows from (31), (65) and (81); (83) is due to (31), (66), (82) and $\mu_k \ge 1 \ge \epsilon_k$; (84) is due to (67), (81), (82), (83) and $\epsilon_k \in (0, 1]$; (85) follows from $\mu_k \ge \rho_k \ge 1 \ge \epsilon_k$ and (32); and (86) is due to (33), (81) and the fact that $\epsilon_k \in (0, 1]$ and $\rho_k\eta_k = 1$. By the above inequalities, (69), (81), $T > 1$ and $\mu_k \ge 1 \ge \epsilon_k$, one has

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

$$\leq \sum_{k=0}^{K} 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} \epsilon_k^{-2} \rho_k^{-1/2} \mu_k^{3/2} T \left(2(\log(\epsilon_k^{-2}\rho_k^{-3}\mu_k^5 M))_+ + 2 + 2\log(2\epsilon_k^{-2}\mu_k T)\right)$$
$$\leq \sum_{k=0}^{K} 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T \epsilon_k^{-2} \rho_k^{-1/2} \mu_k^{3/2} \left(12\log\mu_k - 6\log\rho_k - 8\log\epsilon_k + 2(\log M)_+ + 2 + 2\log(2T)\right), \tag{87}$$

where the first inequality follows from $\epsilon_k \in (0,1]$, (69), (81), (85) and (86), 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 2 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 (87) that

$$\begin{aligned}
\sum_{k=0}^{K} N_k &\leq \sum_{k=0}^{K} 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T \\
&\quad \times \epsilon_k^{-6} \left(38\log(1/\epsilon_k) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&= 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T \sum_{k=0}^{K} \epsilon_0^{-6}\tau^{-6k} \left(38k\log(1/\tau) + 38\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&\leq 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T \sum_{k=0}^{K} \epsilon_0^{-6}\tau^{-6k} \left(38K\log(1/\tau) + 38\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&\leq 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T \epsilon_0^{-6}\tau^{-6K}(1-\tau^6)^{-1} \left(38K\log(1/\tau) + 38\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&\leq 3397 \max\left\{2, \sqrt{L/(2\sigma)}\right\} T \epsilon_0^{-6}(1-\tau^6)^{-1} \\
&\quad \times (\tau\varepsilon/\epsilon_0)^{-6} \left(38K\log(1/\tau) + 38\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right),
\end{aligned} \tag{88}$$

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

In addition, observe from (21), (30), (62) and $\rho_k^{-1}\mu_k \geq 1$, one has that for all $0 \leq k \in \mathbb{K}-1$,

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

Using this, (79), $\epsilon_k = \epsilon_0\tau^k$, $\rho_k = \epsilon_k^{-1}$, $\mu_k = \epsilon_k^{-3}$ and $\eta_k = \epsilon_k$, we have

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

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 (42) and (88) implies that $\sum_{k=1}^{K}(N_k + \widetilde{N}_k) \leq N$. 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] L. Chen, J. Xu, and J. Zhang. On bilevel optimization without lower-level strong convexity. *arXiv preprint arXiv:2301.00712*, 2023.

[6] 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.

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

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

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

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

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

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

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] X. Hu, N. Xiao, X. Liu, and K.-C. Toh. An improved unconstrained approach for bilevel optimization. *arXiv preprint arXiv:2208.00732*, 2022.

[21] F. Huang and H. Huang. Biadam: Fast adaptive bilevel optimization methods. *arXiv preprint arXiv:2106.11396*, 2021.

[22] M. Huang, K. Ji, S. Ma, and L. Lai. Efficiently escaping saddle points in bilevel optimization. *arXiv preprint arXiv:2202.03684*, 2022.

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

[24] 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.

[25] K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pages 4882–4892, 2021.

[26] 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.

[27] V. Konda and J. Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

[28] J. Kwon, D. Kwon, S. Wright, and R. Nowak. A fully first-order method for stochastic bilevel optimization. *arXiv preprint arXiv:2301.10945*, 2023.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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*, 2021.

[35] D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

[36] Z. Lu and S. Mei. A first-order augmented Lagrangian method for constrained minimax optimization. *arXiv preprint arXiv:2301.02060*, 2023.

[37] Z. Lu and S. Mei. A first-order method for nonconvex-strongly-concave constrained minimax optimization. *Preprint*, 2023. `https://zhaosong-lu.github.io/ResearchPapers/strongly-cvx-minimax.pdf`.

[38] Z. Lu and S. Mei. First-order penalty methods for bilevel optimization. *arXiv preprint arXiv:2301.01716*, 2023.

[39] Z. Lu and S. Mei. Solving bilevel optimization via sequential minimax optimization. *Preprint*, 2023.

[40] 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.

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

[42] X. Ma, W. Yao, J. J. Ye, and J. Zhang. Combined approach with second-order optimality conditions for bilevel programming problems. 2023. To appear in *Journal of Convex Analysis*.

[43] D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122, 2015.

[44] 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.

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

[46] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

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

[48] 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.

[49] 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, 2013.

[50] F. Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pages 737–746, 2016.

[51] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

[52] H. Shen and T. Chen. On penalty-based bilevel gradient descent method. *arXiv preprint arXiv:2302.05185*, 2023.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

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

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

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

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

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

[63] J. J. Ye, X. Yuan, S. Zeng, and J. Zhang. Difference of convex algorithms for bilevel programs with applications in hyperparameter selection. *Mathematical Programming*, pages 1–34, 2022.

A An optimal first-order method for unconstrained strongly convex optimization problems

In this part we review an optimal first-order method for solving a class of strongly convex optimization problems in the form of

$$\Psi^* = \min_x \{\Psi(x) \coloneqq \phi(x) + P(x)\} \tag{89}$$

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

We next present an optimal first-order method [40, Algorithm 4] for solving problem (89), which is a slight variant of Nesterov's optimal first-order methods [31, 47].

Algorithm 3 An optimal first-order method for problem (89)

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

1: Compute

$$x^0 = \mathrm{prox}_{P/L_{\nabla\phi}}\left(\tilde{x}^0 - \frac{1}{L_{\nabla\phi}}\nabla\phi(\tilde{x}^0)\right).$$

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

$$z^{k+1} = \arg\min_z \left\{\ell(z; y^k) + \frac{\alpha L_{\nabla\phi}}{2}\|z - \alpha y^k - (1-\alpha)z^k\|^2\right\},$$

where

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

6: Set $x^{k+1} = (1 - \alpha)x^k + \alpha z^{k+1}$.

7: Compute

$$\tilde{x}^{k+1} = \mathrm{prox}_{P/L_{\nabla\phi}}\left(x^{k+1} - \frac{1}{L_{\nabla\phi}}\nabla\phi(x^{k+1})\right).$$

8: Terminate the algorithm and output $\tilde{x}^{k+1}$ if

$$\|\tilde{x}^{k+1} - x^k\| \le \frac{\tilde{\epsilon}}{2L_{\nabla\phi}}.$$

9: **end for**

The following result provides an *iteration-complexity* of Algorithm 3 for finding an approximate optimal solution of problem (89), which was established in [40, Proposition 4].

Theorem 3. *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 (89) satisfying $\mathrm{dist}(0, \partial\Psi(\tilde{x}^{k+1})) \le 2L_{\nabla\phi}\|\tilde{x}^{k+1} - x^{k+1}\| \le \tilde{\epsilon}$ is generated by running Algorithm 3 for at most $\widetilde{T}$ iterations, where*

$$\widetilde{T} = \left\lceil \sqrt{\frac{L_{\nabla\phi}}{\sigma_\phi}} \right\rceil \max\left\{1, \left\lceil 2\log\frac{2L_{\nabla\phi}D_P}{\tilde{\epsilon}} \right\rceil\right\}.$$

Remark 5. *By the convexity of Ψ, $D_P = \max_{x,y\in\mathrm{dom}\ P}\|x - y\|$, and Theorem 3, it is not hard to show that the output $\tilde{x}^{k+1}$ of Algorithm 3 satisfies*

$$\Psi(\tilde{x}^{k+1}) - \Psi^* \le \mathrm{dist}(0, \partial\Psi(\tilde{x}^{k+1}))D_P \le \tilde{\epsilon}D_P. \tag{90}$$

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

In this part we present a first-order method proposed in [36, Algorithm 2] for finding an ϵ-primal-dual stationary point of the nonconvex-concave minimax problem

$$H^* = \min_x \max_y \{H(x,y) \coloneqq h(x,y) + p(x) - q(y)\}, \tag{91}$$

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 operator 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 for any $x \in \operatorname{dom} p$.

For ease of presentation, we define

$$D_p = \max\{\|u-v\| \big| u, v \in \operatorname{dom} p\}, \quad D_q = \max\{\|u-v\| \big| u, v \in \operatorname{dom} q\}, \tag{92}$$

$$H_{\text{low}} = \min\{H(x,y) | (x,y) \in \operatorname{dom} p \times \operatorname{dom} q\}. \tag{93}$$

Given an iterate (x^k, y^k), the first-order method [36, Algorithm 2] finds the next iterate (x^{k+1}, y^{k+1}) by applying a modified optimal first-order method [36, Algorithm 1] to the strongly-convex-strongly-concave minimax problem

$$\min_x \max_y \left\{h_k(x,y) = h(x,y) + L_{\nabla h}\|x - x^k\|^2\right\}. \tag{94}$$

For ease reference, we next present the modified optimal first-order method [36, Algorithm 1] in Algorithm 4 below for solving the strongly-convex-strongly-concave minimax problem

$$\min_x \max_y \left\{\bar{h}(x,y) + p(x) - q(y)\right\}, \tag{95}$$

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 $\bar{\sigma}_x, \bar{\sigma}_y > 0$. In Algorithm 4, the functions $\hat{h}$, a_x^k and a_y^k are defined as follows:

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

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 (95)

Input: $\tau > 0$, $\bar{z}^0 = z_f^0 \in -\bar{\sigma}_x \mathrm{dom}\, p$,[4] $\bar{y}^0 = y_f^0 \in \mathrm{dom}\, q$, $(z^0, y^0) = (\bar{z}^0, \bar{y}^0)$, $\bar{\alpha} = \min\left\{1, \sqrt{8\bar{\sigma}_y/\bar{\sigma}_x}\right\}$, $\eta_z = \bar{\sigma}_x/2$, $\eta_y = \min\{1/(2\bar{\sigma}_y), 4/(\bar{\alpha}\bar{\sigma}_x)\}$, $\beta_t = 2/(t+3)$, $\zeta = \left(2\sqrt{5}(1 + 8L_{\nabla\bar{h}}/\bar{\sigma}_x)\right)^{-1}$, $\gamma_x = \gamma_y = 8\bar{\sigma}_x^{-1}$, 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 are now ready to present the first-order method [36, Algorithm 2] for finding an ϵ-primal-dual stationary point of (91) in Algorithm 5 below.

Algorithm 5 A first-order method for problem (91)

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

for $k = 0, 1, 2, \ldots$ **do**
- Call Algorithm 4 with $\bar{h} \leftarrow h_k$, $\tau \leftarrow \epsilon_k$, $\bar{\sigma}_x \leftarrow L_{\nabla h}$, $\bar{\sigma}_y \leftarrow \sigma_y$, $L_{\nabla\bar{h}} \leftarrow 3L_{\nabla h}$, $\bar{z}^0 = z_f^0 \leftarrow -\bar{\sigma}_x x^k$, $\bar{y}^0 = y_f^0 \leftarrow y^k$, and denote its output by (x^{k+1}, y^{k+1}), where h_k is given in (94).
- 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 the iteration complexity of Algorithm 5, whose proof is given in [36, Theorem 2].

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

Theorem 4 (Complexity of Algorithm 5). *Suppose that Assumption 3 holds. Let H^*, H D_p, D_q, and $H_{\rm low}$ be defined in (91), (92) and (93), $L_{\nabla h}$ be given in Assumption 3, ϵ, ϵ_0 and x^0 be given in Algorithm 5, and*

$$
\begin{aligned}
\tilde{\alpha} =& \ \min\left\{1, \sqrt{8\sigma_y/L_{\nabla h}}\right\},\\
\tilde{\delta} =& \ (2+\tilde{\alpha}^{-1})L_{\nabla h}D_p^2 + \max\left\{2\sigma_y, \tilde{\alpha}L_{\nabla h}/4\right\} D_q^2,\\
\widetilde{K} =& \ \left\lceil 16(\max_y H(\tilde{x}^0, y) - H^*)L_{\nabla h}\epsilon^{-2} + 32\tilde{\epsilon}_0^2(1+\sigma_y^{-2}L_{\nabla h}^2)\epsilon^{-2} - 1\right\rceil_+,\\
\widetilde{N} =& \ 3397\max\left\{2, \sqrt{L_{\nabla h}/(2\sigma_y)}\right\}\\
&\times\left((\widetilde{K}+1)\left(\log\frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{1}{2\sigma_y}, \frac{4}{\tilde{\alpha}L_{\nabla h}}\right\}\right\}\left(\tilde{\delta} + 2\tilde{\alpha}^{-1}(H^* - H_{\rm low} + L_{\nabla h}D_p^2)\right)}{\left[9L_{\nabla h}^2/\min\{L_{\nabla h}, \sigma_y\} + 3L_{\nabla h}\right]^{-2}\tilde{\epsilon}_0^2}\right)_+\right.\\
&\left. + \widetilde{K} + 1 + 2\widetilde{K}\log(\widetilde{K}+1)\right).
\end{aligned}
$$

Then Algorithm 5 terminates and outputs an ϵ-primal-dual stationary point (x_ϵ, y_ϵ) of (91) in at most $\widetilde{K}+1$ outer iterations that satisfies

$$\max_y H(x_\epsilon, y) \le \max_y H(\tilde{x}^0, y) + 2\tilde{\epsilon}_0^2\left(L_{\nabla h}^{-1} + \sigma_y^{-2}L_{\nabla h}\right). \tag{96}$$

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