First-order penalty methods for bilevel optimization

Zhaosong Lu * Sanyou Mei *

January 4, 2023 (Revised: April 18, 2023; September 27, 2023; December 6, 2023)

Abstract

In this paper we study a class of unconstrained and constrained bilevel optimization problems in which the lower level is a possibly nonsmooth convex optimization problem, while the upper level is a possibly nonconvex optimization problem. We introduce a notion of ε-KKT solution for them and show that an ε-KKT solution leads to an $\mathcal{O}(\sqrt{\varepsilon})$- or $\mathcal{O}(\varepsilon)$-hypergradient based stionary point under suitable assumptions. We also propose first-order penalty methods for finding an ε-KKT solution of them, whose subproblems turn out to be a structured minimax problem and can be suitably solved by a first-order method recently developed by the authors. Under suitable assumptions, an *operation complexity* of $\mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^{-7}\log\varepsilon^{-1})$, measured by their fundamental operations, is established for the proposed penalty methods for finding an ε-KKT solution of the unconstrained and constrained bilevel optimization problems, respectively. Preliminary numerical results are presented to illustrate the performance of our proposed methods. To the best of our knowledge, this paper is the first work to demonstrate that bilevel optimization can be approximately solved as minimax optimization, and moreover, it provides the first implementable method with complexity guarantees for such sophisticated bilevel optimization.

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

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

1 Introduction

Bilevel optimization is a two-level hierarchical optimization in which the decision variables in the upper level are also involved in the lower level. Generically, it can be written in the following form:

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

Bilevel optimization has found a variety of important applications, including adversarial training [45, 46, 57], continual learning [40], hyperparameter tuning [3, 17], image reconstruction [9], meta-learning [4, 28, 52], neural architecture search [15, 38], reinforcement learning [23, 31], and Stackelberg games [59]. More applications about it can be found in [2, 8, 10, 11, 12, 54] and the references therein. Theoretical properties including optimality conditions of (1) have been extensively studied in the literature (e.g., see [12, 13, 43, 58, 62]).

Numerous methods have been developed for solving some special cases of (1). For example, constraint-based methods [22, 53], deterministic gradient-based methods [16, 17, 20, 24, 44, 51, 52], and stochastic gradient-based methods [6, 18, 21, 23, 25, 26, 29, 30, 34, 35, 61] were proposed for solving (1) with $g\equiv 0$, $\tilde{g}\equiv 0$, 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 [36, 37, 56] 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, 42, 50]). Recently, difference-of-convex (DC) algorithms were developed in [63] for solving (1) with $g\equiv 0$, f being a DC function, and $\tilde{f}$, $\tilde{g}$ being convex

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

[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.

functions. In addition, a double penalty method [27] was proposed for (1), which solves a sequence of bilevel optimization problems of the form

$$\begin{array}{ll} \min\limits_{x,y} & f(x,y)+\rho_k\Psi(x,y) \\ \text{s.t.} & y\in \operatorname{argmin}\limits_z \tilde{f}(x,z)+\rho_k\tilde{\Psi}(x,z), \end{array} \tag{2}$$

where $\{\rho_k\}$ is a sequence of penalty parameters, and Ψ and $\tilde{\Psi}$ are a penalty function associated with the sets $\{(x,y)|g(x,y)\le 0\}$ and $\{(x,z)|\tilde{g}(x,z)\le 0\}$, respectively. Though problem (2) appears to be simpler than (1), there is no method available for finding an approximate solution of (2) in general. Consequently, the double penalty method [27] is typically not implementable. More discussion on algorithmic development for bilevel optimization can be found in [2, 8, 12, 39, 55, 58]) and the references therein.

It has long been known that the notorious challenge of bilevel optimization (1) mainly comes from the lower level part, which requires that the variable y be a solution of another optimization problem. Due to this, for the sake of simplicity, we only consider a subclass of bilevel optimization with the constraint $g(x,y)\le 0$ being excluded, namely,

$$\begin{array}{ll} \min\limits_{x,y} & f(x,y) \\ \text{s.t.} & y\in \operatorname{argmin}\limits_z\{\tilde{f}(x,z)|\tilde{g}(x,z)\le 0\}. \end{array} \tag{3}$$

Nevertheless, the results in this paper can be possibly extended to problem (1).

The main goal of this paper is to develop an implementable first-order method with complexity guarantees for solving problem (3). Our key insights for this development are: (i) problem (3) can be approximately solved as a structured minimax problem that results from a novel penalty approach; (ii) the resulting structured minimax problem can be suitably solved by a first-order method proposed in [41, Algorithm 2]. As a result, these lead to development of a novel first-order penalty method for solving (3), which enjoys the following appealing features.

- It uses only the first-order information of the problem. Specifically, its fundamental operations consist only of gradient evaluation of $\tilde{g}$ and the smooth component of f and $\tilde{f}$ and also proximal operator evaluation of the nonsmooth component of f and $\tilde{f}$. Thus, it is suitable for solving large-scale problems (see Sections 2 and 3).

- It has theoretical guarantees on operation complexity, which is measured by the aforementioned fundamental operations, for finding an ε-KKT solution of (3). Specifically, when $\tilde{g}\equiv 0$, it enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$. Otherwise, it enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-7}\log\varepsilon^{-1})$ (see Theorems 3 and 5).

- It is applicable to a broader class of problems than existing methods. For example, it can be applied to (3) with f, $\tilde{f}$ being nonsmooth and $\tilde{f}$, $\tilde{g}$ being nonconvex with respect to x, which is however not suitable for existing methods.

To the best of our knowledge, this paper is the first work to demonstrate that bilevel optimization can be approximately solved as minimax optimization, and moreover, it provides the first implementable method with complexity guarantees for the sophisticated bilevel optimization problem (3).

The rest of this paper is organized as follows. In Subsection 1.1 we introduce some notation and terminology. In Sections 2 and 3, we propose first-order penalty methods for unconstrained and constrained bilevel optimization and study their complexity, respectively. Preliminary numerical results and the proofs of the main results are respectively presented in Sections 4 and 5. Finally, we make some concluding remarks in Section 6.

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 respectively denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, unless stated otherwise. 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 $\operatorname{dist}(x,S)=\min_{x'\in S}\|x'-x\|$ 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 $x, x' \in S$.[2] For a closed convex function $p : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$,[3] the *proximal operator* associated with p is denoted by prox_p, that is,

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

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

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

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

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

$$\partial\phi(x) = \{s \in \mathbb{R}^n \big| s^T d \le \phi'(x; d) \ \ \forall d \in \mathbb{R}^n\}.$$

We use $\partial_{x_i}\phi(x)$ to denote the subdifferential with respect to x_i. In addition, for an upper semicontinuous function ϕ, its subdifferential is defined as $\partial\phi = -\partial(-\phi)$. If ϕ is locally Lipschitz continuous, the above definition of subdifferential coincides with the Clarke subdifferential. Besides, if ϕ is convex, it coincides with the ordinary subdifferential for convex functions. Also, if ϕ is continuously differentiable at x , we simply have $\partial\phi(x) = \{\nabla\phi(x)\}$, where $\nabla\phi(x)$ is the gradient of ϕ at x. In addition, it is not hard to verify that $\partial(\phi_1 + \phi_2)(x) = \nabla\phi_1(x) + \partial\phi_2(x)$ if ϕ_1 is continuously differentiable at x and ϕ_2 is lower or upper semicontinuous at x. See [7, 60] for more details.

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

$$\Psi^* = \min_x \max_y \Psi(x, y), \tag{4}$$

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 $(x_\varepsilon, y_\varepsilon)$ is called an ϵ-optimal solution of the minimax problem (4) if*

$$\max_y \Psi(x_\epsilon, y) - \Psi(x_\varepsilon, y_\varepsilon) \le \epsilon, \quad \Psi(x_\varepsilon, y_\varepsilon) - \Psi^* \le \epsilon.$$

Definition 2. *A point (x, y) is called a stationary point of the minimax problem (4) if*

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

In addition, for any $\epsilon > 0$, a point $(x_\varepsilon, y_\varepsilon)$ is called an ϵ-stationary point of the minimax problem (4) if

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

2 Unconstrained bilevel optimization

In this section, we consider an unconstrained bilevel optimization problem[4]

$$\begin{aligned} f^* = \min \quad & f(x, y) \\ \text{s.t.} \quad & y \in \underset{z}{\text{argmin}}\, \tilde{f}(x, z). \end{aligned} \tag{5}$$

Assume that problem (5) has at least one optimal solution. In addition, f and $\tilde{f}$ satisfy the following assumptions.

[2] When ϕ is a mapping, the norm used in $\|\nabla\phi(x) - \nabla\phi(x')\|$ is the Frobenius norm.

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

[4] For convenience, problem (5) is referred to as an unconstrained bilevel optimization problem since its lower level part does not have an explicit constraint. Strictly speaking, it can be a constrained bilevel optimization problem. For example, when part of f and/or $\tilde{f}$ is the indicator function of a closed convex set, (5) is essentially a constrained bilevel optimization problem.

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

For notational convenience, we define

$$D_{\mathbf{x}} := \max\{\|u - v\| \big| u, v \in \mathcal{X}\}, \quad D_{\mathbf{y}} := \max\{\|u - v\| \big| u, v \in \mathcal{Y}\}, \tag{6}$$

$$\tilde{f}_{\rm hi} := \max\{\tilde{f}(x,y) | (x,y) \in \mathcal{X} \times \mathcal{Y}\}, \quad \tilde{f}_{\rm low} := \min\{\tilde{f}(x,y) | (x,y) \in \mathcal{X} \times \mathcal{Y}\}, \tag{7}$$

$$f_{\rm low} := \min\{f(x,y) | (x,y) \in \mathcal{X} \times \mathcal{Y}\}. \tag{8}$$

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.

The goal of this section is to propose first-order penalty methods for solving problem (5). To this end, we first observe that problem (5) can be viewed as

$$\min_{x,y}\{f(x,y) | \tilde{f}(x,y) \le \min_z \tilde{f}(x,z)\}. \tag{9}$$

Notice that $\tilde{f}(x,y) - \min_z \tilde{f}(x,z) \ge 0$ for all x, y. Consequently, a natural *penalty problem* associated with (9) is

$$\min_{x,y} f(x,y) + \rho(\tilde{f}(x,y) - \min_z \tilde{f}(x,z)), \tag{10}$$

where $\rho > 0$ is a penalty parameter. We further observe that (10) is equivalent to the *minimax problem*

$$\min_{x,y} \max_z P_\rho(x,y,z), \quad \text{where} \quad P_\rho(x,y,z) := f(x,y) + \rho(\tilde{f}(x,y) - \tilde{f}(x,z)). \tag{11}$$

In view of Assumption 1(i), P_ρ can be rewritten as

$$P_\rho(x,y,z) = \big(f_1(x,y) + \rho\tilde{f}_1(x,y) - \rho\tilde{f}_1(x,z)\big) + \big(f_2(x) + \rho\tilde{f}_2(y) - \rho\tilde{f}_2(z)\big). \tag{12}$$

By this and Assumption 1, one can observe that P_ρ enjoys the following nice properties.

- P_ρ is the sum of smooth function $f_1(x,y) + \rho\tilde{f}_1(x,y) - \rho\tilde{f}_1(x,z)$ with Lipschitz continuous gradient and possibly nonsmooth function $f_2(x) + \rho\tilde{f}_2(y) - \rho\tilde{f}_2(z)$ with exactly computable proximal operator.
- P_ρ is nonconvex in (x,y) but concave in z.

Thanks to this nice structure of P_ρ, an approximate stationary point of the minimax problem (11) can be found by a first-order method proposed in [41, Algorithm 2] (see Algorithm 6 in Appendix A).

Based on the above observations, we are now ready to propose penalty methods for the unconstrained bilevel optimization problem (5) by solving either a sequence of minimax problems or a single minimax problem in the form of (11). Specifically, we first propose an *ideal* penalty method for (5) by solving a sequence of minimax problems (see Algorithm 1). Then we propose a *practical* penalty method for (5) by finding an approximate stationary point of a single minimax problem (see Algorithm 2).

Algorithm 1 An ideal penalty method for problem (5)

Input: positive sequences $\{\rho_k\}$ and $\{\epsilon_k\}$ with $\lim_{k\to\infty}(\rho_k, \epsilon_k) = (\infty, 0)$.

1: **for** $k = 0, 1, 2, \ldots$ **do**

2: Find an ϵ_k-optimal solution (x^k, y^k, z^k) of problem (11) with $\rho = \rho_k$.

3: **end for**

The following theorem states a convergence result of Algorithm 1, whose proof is deferred to Section 5.1.

Theorem 1 (Convergence of Algorithm 1). *Suppose that Assumption 1 holds and that $\{(x^k, y^k, z^k)\}$ is generated by Algorithm 1. Then any accumulation point of $\{(x^k, y^k)\}$ is an optimal solution of problem (5).*

Notice that (11) is a *nonconvex*-concave minimax problem. It is typically hard to find an ϵ-optimal solution of (11) for an arbitrary $\epsilon > 0$. Consequently, Algorithm 1 is *not implementable* in general. We next propose a *practical* penalty method for problem (5) by applying Algorithm 6 (see Appendix A) to find an approximate stationary point of a single minimax problem (11) with a suitable choice of ρ.

Algorithm 2 A practical penalty method for problem (5)

Input: $\varepsilon \in (0, 1/4]$, $\rho = \varepsilon^{-1}$, $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$ with $\tilde{f}(x^0, y^0) \le \min_y \tilde{f}(x^0, y) + \varepsilon$.

1: Call Algorithm 6 in Appendix A with $\epsilon \leftarrow \varepsilon$, $\epsilon_0 \leftarrow \varepsilon^{3/2}$, $\hat{x}^0 \leftarrow (x^0, y^0)$, $\hat{y}^0 \leftarrow y^0$, and $L_{\nabla h} \leftarrow L_{\nabla f_1} + 2\varepsilon^{-1} L_{\nabla \tilde{f}_1}$ to find an ε-stationary point $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ of problem (11) with $\rho = \varepsilon^{-1}$.

2: **Output**: $(x_\varepsilon, y_\varepsilon)$.

Remark 1. *(i) The initial point (x^0, y^0) of Algorithm 2 can be found by an additional procedure. Indeed, one can first choose any $x^0 \in \mathcal{X}$ and then apply accelerated proximal gradient method [47] to the problem $\min_y \tilde{f}(x^0, y)$ for finding $y^0 \in \mathcal{Y}$ such that $\tilde{f}(x^0, y^0) \le \min_y \tilde{f}(x^0, y) + \varepsilon$; (ii) As seen from Theorem 6 (see Appendix A), an ϵ-stationary point of* (11) *can be successfully found in step 1 of Algorithm 2 by applying Algorithm 6 to* (11)*; (iii) For the sake of simplicity, a single subproblem of the form* (11) *with static penalty and tolerance parameters is solved in Algorithm 2. Nevertheless, Algorithm 2 can be modified into a perhaps practically more efficient algorithm by solving a sequence of subproblems of the form* (11) *with dynamic penalty and tolerance parameters instead.*

In order to characterize the approximate solution found by Algorithm 2, we next introduce a notion of ε-KKT solution of problem (5).

Recall that problem (5) can be viewed as problem (9), which is a constrained optimization problem. In the spirit of classical constrained optimization, one would naturally be interested in a KKT solution (x, y) of (9) or equivalently (5), namely, (x, y) satisfies $\tilde{f}(x, y) \le \min_z \tilde{f}(x, z)$ and moreover (x, y) is a stationary point of the problem

$$\min_{x', y'} f(x', y') + \rho\big(\tilde{f}(x', y') - \min_{z'} \tilde{f}(x', z')\big) \tag{13}$$

for some $\rho \ge 0$.[5] Yet, due to the sophisticated problem structure, characterizing a stationary point of (13) is generally difficult. On another hand, notice that problem (13) is equivalent to the minimax problem

$$\min_{x', y'} \max_{z'} f(x', y') + \rho(\tilde{f}(x', y') - \tilde{f}(x', z')),$$

whose stationary point (x, y, z), according to Definition 2 and Assumption 1, satisfies

$$0 \in \partial f(x, y) + \rho \partial \tilde{f}(x, y) - (\rho \nabla_x \tilde{f}(x, z); 0), \quad 0 \in \rho \partial_z \tilde{f}(x, z). \tag{14}$$

Based on this observation, we are instead interested in a (weak) KKT solution of problem (5) and its inexact counterpart that are defined below.

Definition 3. *The pair (x, y) is said to be a KKT solution of problem* (5) *if there exists $(z, \rho) \in \mathbb{R}^m \times \mathbb{R}_+$ such that* (14) *and $\tilde{f}(x, y) \le \min_{z'} \tilde{f}(x, z')$ hold. In addition, for any $\varepsilon > 0$, (x, y) is said to be an ε-KKT solution of problem* (5) *if there exists $(z, \rho) \in \mathbb{R}^m \times \mathbb{R}_+$ such that*

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

Recently, a hypergradient-based stationary point has been considered in the literature (e.g., [18, 51]) for problem (5) under the assumption that f and $\tilde{f}$ are twice continuously differentiable in $\mathbb{R}^n \times \mathbb{R}^m$ and $\tilde{f}(x, \cdot)$ is strongly convex for any $x \in \mathbb{R}^n$. Under this assumption, the hyper-objective function Φ of (5), defined as

$$\Phi(x) := f(x, y^*(x)), \text{ where } y^*(x) = \operatorname*{argmin}_z \tilde{f}(x, z), \tag{15}$$

is continuously differentiable. Moreover, following from [18, Equation (2.8)], the hypergradient of (5), i.e., the gradient of Φ, is given by

$$\nabla \Phi(x) = \nabla_x f(x, y^*(x)) - \nabla^2_{xy} \tilde{f}(x, y^*(x)) [\nabla^2_{yy} \tilde{f}(x, y^*(x))]^{-1} \nabla_y f(x, y^*(x)) \quad \forall x \in \mathbb{R}^n. \tag{16}$$

[5] The relation $\tilde{f}(x, y) \le \min_z \tilde{f}(x, z)$ implies that $\tilde{f}(x, y) = \min_z \tilde{f}(x, z)$ and hence the complementary slackness condition $\rho(\tilde{f}(x, y) - \min_z \tilde{f}(x, z)) = 0$ holds.

In addition, it is not hard to observe that problem (5) is equivalent to

$$\min_x \Phi(x). \tag{17}$$

In view of this, hypergradient based stationary point and its approximate counterpart are introduced in the literature (e.g., [18, 51]) for problem (5), based on the classical stationary point and its approximate counterpart of problem (17). More specifically, $x \in \mathbb{R}^n$ is called a *hypergradient-based stationary point* of problem (5) if $\nabla\Phi(x) = 0$, and it is called an *ε-hypergradient-based stationary point* of (5) if $\|\nabla\Phi(x)\| \le \varepsilon$ for any $\varepsilon > 0$.

We now study the relationship between an ε-KKT solution and an approximate hypergradient based stationary point of problem (5). Specifically, under some suitable assumptions, the following theorem shows that if (x, y) is an ε-KKT solution of problem (5), then x is an $\mathcal{O}(\sqrt{\varepsilon})$- or $\mathcal{O}(\varepsilon)$-hypergradient-based stationary point of it. The proof of this theorem is deferred to Subsection 5.1.

Theorem 2. *Let $\varepsilon_0, \rho_0 > 0$ be given and $\Omega \subset \mathbb{R}^n$ be a nonempty compact set. Assume that f and $\tilde{f}$ are continuously differentiable and twice continuously differentiable in $\mathbb{R}^n \times \mathbb{R}^m$ respectively, $\tilde{f}(x', \cdot)$ is strongly convex with modulus $\sigma > 0$ for all x' in an open set $\mathcal{N}$ containing Ω, $\nabla f(x', \cdot)$ is L_1-Lipschitz continuous for all $x' \in \Omega$, and that $\nabla^2 \tilde{f}(x', \cdot)$ is L_2-Lipschitz continuous for all $x' \in \Omega$. Suppose that $(x, y) \in \Omega \times \mathbb{R}^m$ is an ε-KKT solution of problem (5) with its associated $\rho \ge \rho_0$ for some $0 < \varepsilon \le \varepsilon_0$. Let $y^*(x')$ be defined in (15) and*

$$\bar{C} = \max\big\{\|\nabla_y f(x', y')\| : x' \in \Omega, \|y' - y^*(x')\| \le \sqrt{2\sigma^{-1}\varepsilon}\big\}, \tag{18}$$

$$\theta = \min\big\{(\rho\sigma)^{-1}(\varepsilon + \bar{C}),\ \sqrt{2\sigma^{-1}\varepsilon}\big\}, \quad C = \max_{x' \in \Omega} \|\nabla^2_{xy} \tilde{f}(x', y^*(x'))[\nabla^2_{yy} \tilde{f}(x', y^*(x'))]^{-1}\|. \tag{19}$$

Then we have

$$\|\nabla\Phi(x)\| \le (2C+1)\varepsilon + (C+1)\Big(L_1\theta + \frac{L_2\rho\theta^2}{2} + \frac{L_2\varepsilon^2}{2\rho\sigma^2}\Big) \tag{20}$$

$$\le (2C+1)\varepsilon + (C+1)\Big(L_1\sqrt{2\sigma^{-1}} + \frac{L_2\sigma^{-3/2}(\varepsilon_0 + \bar{C})}{\sqrt{2}} + \frac{\varepsilon_0^{3/2} L_2}{2\rho_0\sigma^2}\Big)\sqrt{\varepsilon}. \tag{21}$$

Remark 2. *Based on the assumptions in Theorem 2, it is not hard to observe that $\bar{C}$, C and θ are finite. It then follows from (21) that $\nabla\Phi(x) = \mathcal{O}(\sqrt{\varepsilon})$ holds in general. Nevertheless, this result is improved to $\nabla\Phi(x) = \mathcal{O}(\varepsilon)$ when ρ is at least order of ε^{-1}, i.e., $\rho \ge c\varepsilon^{-1}$ for some constant $c > 0$ independent on ε, which can be observed from (20). Consequently, under the assumptions stated in Theorem 2, if (x, y) is an ε-KKT solution of problem (5), then x is an $\mathcal{O}(\sqrt{\varepsilon})$- or $\mathcal{O}(\varepsilon)$-hypergradient-based stationary point of it.*

We next present a theorem regarding *operation complexity* of Algorithm 2, 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 (5), whose proof is deferred to Subsection 5.1.

Theorem 3 (Complexity of Algorithm 2). *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 (5), (6), (7) and (8), $L_{\nabla f_1}$ and $L_{\nabla \tilde{f}_1}$ be given in Assumption 1, ε, ρ, x^0, y^0 and z_ε be given in Algorithm 2, and*

$$\widehat{L} = L_{\nabla f_1} + 2\varepsilon^{-1} L_{\nabla \tilde{f}_1},\ \hat{\alpha} = \min\left\{1, \sqrt{4\varepsilon/(D_{\mathbf{y}}\widehat{L})}\right\}, \tag{22}$$

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

$$\widehat{C} = \frac{4\max\left\{\frac{1}{2\widehat{L}}, \min\left\{\frac{D_{\mathbf{y}}}{\varepsilon}, \frac{4}{\hat{\alpha}\widehat{L}}\right\}\right\}\left[\hat{\delta} + 2\hat{\alpha}^{-1}(f^* - f_{\rm low} + \varepsilon^{-1}(\tilde{f}_{\rm hi} - \tilde{f}_{\rm low}) + \varepsilon D_{\mathbf{y}}/4 + \widehat{L}(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2))\right]}{\left[(3\widehat{L} + \varepsilon/(2D_{\mathbf{y}}))^2/\min\{\widehat{L}, \varepsilon/(2D_{\mathbf{y}})\} + 3\widehat{L} + \varepsilon/(2D_{\mathbf{y}})\right]^{-2}\varepsilon^3},$$

$$\widehat{K} = \Big\lceil 16(1 + f(x^0, y^0) - f_{\rm low} + \varepsilon D_{\mathbf{y}}/4)\widehat{L}\varepsilon^{-2} + 32(1 + 4D_{\mathbf{y}}^2\widehat{L}^2\varepsilon^{-2})\varepsilon - 1\Big\rceil_+,$$

$$\widehat{N} = \Big(\Big\lceil 96\sqrt{2}(1 + (24\widehat{L} + 4\varepsilon/D_{\mathbf{y}})\widehat{L}^{-1})\Big\rceil + 2\Big)\max\left\{2, \sqrt{D_{\mathbf{y}}\widehat{L}\varepsilon^{-1}}\right\}$$
$$\times ((\widehat{K} + 1)(\log\widehat{C})_+ + \widehat{K} + 1 + 2\widehat{K}\log(\widehat{K} + 1)).$$

Then Algorithm 2 outputs an approximate solution $(x_\varepsilon, y_\varepsilon)$ of (5) satisfying

$$\operatorname{dist}\Big(0, \partial f(x_\varepsilon, y_\varepsilon) + \rho\partial\tilde{f}(x_\varepsilon, y_\varepsilon) - (\rho\nabla_x\tilde{f}(x_\varepsilon, z_\varepsilon); 0)\Big) \le \varepsilon, \quad \operatorname{dist}\Big(0, \rho\partial_z\tilde{f}(x_\varepsilon, z_\varepsilon)\Big) \le \varepsilon, \tag{23}$$

$$\tilde{f}(x_\varepsilon, y_\varepsilon) \le \min_z \tilde{f}(x_\varepsilon, z) + \varepsilon\left(1 + f(x^0, y^0) - f_{\rm low} + 2\varepsilon^3(\widehat{L}^{-1} + 4D_{\mathbf{y}}^2\widehat{L}\varepsilon^{-2}) + D_{\mathbf{y}}\varepsilon/4\right), \tag{24}$$

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 3. *One can observe from Theorem 3 that $\widehat{L} = \mathcal{O}(\varepsilon^{-1})$, $\hat{\alpha} = \mathcal{O}(\varepsilon)$, $\hat{\delta} = \mathcal{O}(\varepsilon^{-2})$, $\widehat{C} = \mathcal{O}(\varepsilon^{-11})$, $\widehat{K} = \mathcal{O}(\varepsilon^{-3})$, and $\widehat{N} = \mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$. As a result, Algorithm 2 enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$, 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 $(x_\varepsilon, y_\varepsilon)$ of (5) satisfying*

$$\operatorname{dist}\Big(0, \partial f(x_\varepsilon, y_\varepsilon) + \rho\partial\tilde{f}(x_\varepsilon, y_\varepsilon) - (\rho\nabla_x\tilde{f}(x_\varepsilon, z_\varepsilon); 0)\Big) \le \varepsilon, \quad \operatorname{dist}\Big(0, \rho\partial_z\tilde{f}(x_\varepsilon, z_\varepsilon)\Big) \le \varepsilon,$$
$$\tilde{f}(x_\varepsilon, y_\varepsilon) - \min_z \tilde{f}(x_\varepsilon, z) = \mathcal{O}(\varepsilon),$$

where z_ε is given in Algorithm 2 and $\rho = \varepsilon^{-1}$.

3 Constrained bilevel optimization

In this section, we consider a constrained bilevel optimization problem[6]

$$\begin{aligned} f^* = \min \quad & f(x, y) \\ \text{s.t.} \quad & y \in \operatorname*{argmin}_z\{\tilde{f}(x, z) | \tilde{g}(x, z) \le 0\}, \end{aligned} \tag{25}$$

where f and $\tilde{f}$ satisfy 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 (25).

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

(ii) $\tilde{g}: \mathbb{R}^n\times\mathbb{R}^m \to \mathbb{R}^l$ is $L_{\nabla\tilde{g}}$-smooth and $L_{\tilde{g}}$-Lipschitz continuous on $\mathcal{X}\times\mathcal{Y}$.

(iii) $\tilde{g}_i(x, \cdot)$ 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 $G := \min\{-\tilde{g}_i(x, \hat{z}_x) | x \in \mathcal{X},\ i = 1, \ldots, l\} > 0$.[7]

For notational convenience, we define

$$\tilde{f}^*(x) := \min_z\{\tilde{f}(x, z) | \tilde{g}(x, z) \le 0\}, \tag{26}$$

$$\tilde{f}^*_{\rm hi} := \sup\{\tilde{f}^*(x) | x \in \mathcal{X}\}, \tag{27}$$

$$\tilde{g}_{\rm hi} := \max\{\|\tilde{g}(x, y)\| \big| (x, y) \in \mathcal{X}\times\mathcal{Y}\}. \tag{28}$$

It then follows from Assumption 2(ii) that

$$\|\nabla\tilde{g}(x, y)\| \le L_{\tilde{g}} \qquad \forall (x, y) \in \mathcal{X}\times\mathcal{Y}. \tag{29}$$

In addition, by Assumptions 1 and 2 and the compactness of $\mathcal{X}$ and $\mathcal{Y}$, one can observe that $\tilde{g}_{\rm hi}$ and G are finite. Besides, as will be shown in Lemma 3(ii), $\tilde{f}^*_{\rm hi}$ is finite.

The goal of this section is to propose first-order penalty methods for solving problem (25). To this end, let us first introduce a *penalty function* for the lower level optimization problem $y \in \operatorname*{argmin}_z\{\tilde{f}(x, z) | \tilde{g}(x, z) \le 0\}$ of (25), which is given by

$$\widetilde{P}_\mu(x, z) = \tilde{f}(x, z) + \mu\left\|[\tilde{g}(x, z)]_+\right\|^2 \tag{30}$$

for a penalty parameter $\mu > 0$. Observe that problem (25) can be approximately solved as the *unconstrained bilevel optimization* problem

$$f^*_\mu = \min_{x,y}\left\{f(x, y) | y \in \operatorname*{argmin}_z \widetilde{P}_\mu(x, z)\right\}. \tag{31}$$

[6] For convenience, problem (25) is referred to as a constrained bilevel optimization problem since its lower level part has at least one explicit constraint.

[7] 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 (25), 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 (25) with $\tilde{g}(x, z)$ being replaced by $\tilde{g}(x, z) - \epsilon$ for some suitable $\epsilon > 0$ instead, which satisfies Assumption 2(iii).

Further, by the study in Section 2, problem (31) can be approximately solved as the *penalty problem*

$$\min_{x,y} f(x,y) + \rho\Big(\widetilde{P}_\mu(x,y) - \min_z \widetilde{P}_\mu(x,z)\Big) \tag{32}$$

for some suitable $\rho > 0$. One can also observe that problem (32) is equivalent to the *minimax problem*

$$\min_{x,y}\max_z P_{\rho,\mu}(x,y,z), \quad \text{where} \quad P_{\rho,\mu}(x,y,z) \coloneqq f(x,y) + \rho(\widetilde{P}_\mu(x,y) - \widetilde{P}_\mu(x,z)). \tag{33}$$

In view of (30), (33) and Assumption 1(i), $P_{\rho,\mu}$ can be rewritten as

$$\begin{aligned} P_{\rho,\mu}(x,y,z) = \big(&f_1(x,y) + \rho\tilde{f}_1(x,y) + \rho\mu\,\|[\tilde{g}(x,y)]_+\|^2 - \rho\tilde{f}_1(x,z) - \rho\mu\,\|[\tilde{g}(x,z)]_+\|^2\big) \\ &+ \big(f_2(x) + \rho\tilde{f}_2(y) - \rho\tilde{f}_2(z)\big). \end{aligned} \tag{34}$$

By this and Assumptions 1 and 2, one can observe that $P_{\rho,\mu}$ enjoys the following nice properties.

- $P_{\rho,\mu}$ is the sum of smooth function $f_1(x,y)+\rho\tilde{f}_1(x,y)+\rho\mu\,\|[\tilde{g}(x,y)]_+\|^2-\rho\tilde{f}_1(x,z)-\rho\mu\,\|[\tilde{g}(x,z)]_+\|^2$ with Lipschitz continuous gradient and possibly nonsmooth function $f_2(x) + \rho\tilde{f}_2(y) - \rho\tilde{f}_2(z)$ with exactly computable proximal operator;
- $P_{\rho,\mu}$ is nonconvex in (x,y) but concave in z.

Due to this nice structure of $P_{\rho,\mu}$, an approximate stationary point of the minimax problem (33) can be found by a first-order method proposed in [41, Algorithm 2] (see Algorithm 6 in Appendix A).

Based on the above observations, we are now ready to propose penalty methods for the constrained bilevel optimization problem (25) by solving a sequence of minimax problems or a single minimax problem of the form (33). Specifically, we first propose an *ideal* penalty method for (25) by solving a sequence of minimax problems (see Algorithm 3). Then we propose a *practical* penalty method for (25) by finding an approximate stationary point of a single minimax problem (see Algorithm 4).

Algorithm 3 An ideal penalty method for problem (25)

Input: positive sequences $\{\rho_k\}$, $\{\mu_k\}$ and $\{\epsilon_k\}$ with $\lim_{k\to\infty}(\rho_k,\mu_k,\epsilon_k) = (\infty,\infty,0)$.

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: Find an ϵ_k-optimal solution (x^k, y^k, z^k) of problem (33) with $(\rho,\mu) = (\rho_k,\mu_k)$.
3: **end for**

To study convergence of Algorithm 3, we make the following error bound assumption on the solution set of the lower level optimization problem of (25). This type of error bounds has been considered in the context of set-value mappings in the literature (e.g., see [14]).

Assumption 3. *There exist $\bar{\theta} > 0$ and a non-decreasing function $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{\theta\downarrow 0}\omega(\theta) = 0$ such that $\mathrm{dist}(z, \mathcal{S}_\theta(x)) \le \omega(\theta)$ for all $x \in \mathcal{X}$, $z \in \mathcal{S}_0(x)$ and $\theta \in [0,\bar{\theta}]$, where*

$$\mathcal{S}_\theta(x) \coloneqq \operatorname*{argmin}_z\{\tilde{f}(x,z) : \|[\tilde{g}(x,z)]_+\| \le \theta\}.$$

We are now ready to state a convergence result of Algorithm 3, whose proof is deferred to Section 5.2.

Theorem 4 (Convergence of Algorithm 3). *Suppose that Assumptions 1-3 hold and that $\{(x^k, y^k, z^k)\}$ is generated by Algorithm 3. Then any accumulation point of $\{(x^k, y^k)\}$ is an optimal solution of problem (25).*

Notice that (33) is a *nonconvex*-concave minimax problem. It is typically hard to find an ϵ-optimal solution of (33) for an arbitrary $\epsilon > 0$. As a result, Algorithm 3 is generally *not implementable*. We next propose a *practical* penalty method for problem (25) by applying Algorithm 6 (see Appendix A) to find an approximate stationary point of (33) with a suitable choice of ρ and μ.

Algorithm 4 A practical penalty method for problem (25)

Input: $\varepsilon \in (0, 1/4]$, $\rho = \varepsilon^{-1}$, $\mu = \varepsilon^{-2}$, $(x^0, y^0) \in \mathcal{X}\times\mathcal{Y}$ with $\widetilde{P}_\mu(x^0,y^0) \le \min_y \widetilde{P}_\mu(x^0,y) + \varepsilon$.

1: Call Algorithm 6 in Appendix A with $\epsilon \leftarrow \varepsilon$, $\epsilon_0 \leftarrow \varepsilon^{5/2}$, $\hat{x}^0 \leftarrow (x^0,y^0)$, $\hat{y}^0 \leftarrow y^0$, and $L_{\nabla h} \leftarrow L_{\nabla f_1} + 2\rho L_{\nabla\tilde{f}_1} + 4\rho\mu(\tilde{g}_{\rm hi}L_{\nabla\tilde{g}} + L_{\tilde{g}}^2)$ to find an ε-stationary point $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ of problem (33) with $\rho = \varepsilon^{-1}$ and $\mu = \varepsilon^{-2}$.
2: **Output**: $(x_\varepsilon, y_\varepsilon)$.

Remark 4. *(i) The initial point (x^0, y^0) of Algorithm 4 can be found by a similar procedure as described in Remark 1 with $\tilde{f}$ being replaced by $\widetilde{P}_\mu$; (ii) The choice of $\rho = \varepsilon^{-1}$ and $\mu = \varepsilon^{-2}$ is crucial in terms of the order of ε for Algorithm 4 to find an $\mathcal{O}(\varepsilon)$-KKT solution of problem* (25) *with a best operation complexity among all possible choices of ρ and μ, which can be observed from the proof of Theorem 5. Intuitively, the lower level penalty parameter μ has to be larger than the upper level parameter ρ in terms of the order of ε so that the incurred error from the lower level constraint violation is not magnified (see* (92) *in Lemma 8). (iii) As seen from Theorem 6 (see Appendix A), an ϵ-stationary point of* (33) *can be successfully found in step 1 of Algorithm 4 by applying Algorithm 6 to* (33)*; (iv) For the sake of simplicity, a single subproblem of the form* (33) *with static penalty and tolerance parameters is solved in Algorithm 4. Nevertheless, Algorithm 4 can be modified into a perhaps practically more efficient algorithm by solving a sequence of subproblems of the form* (33) *with dynamic penalty and tolerance parameters instead.*

In order to characterize the approximate solution found by Algorithm 4, we next introduce a notion of ε-KKT solution of problem (25).

By the definition of $\tilde{f}^*$ in (26), problem (25) can be viewed as

$$\min_{x,y}\{f(x,y)|\tilde{f}(x,y) \le \tilde{f}^*(x),\ \tilde{g}(x,y) \le 0\}. \tag{35}$$

Its associated Lagrangian function is given by

$$\mathcal{L}(x,y,\rho,\lambda) = f(x,y) + \rho(\tilde{f}(x,y) - \tilde{f}^*(x)) + \langle\lambda, \tilde{g}(x,y)\rangle. \tag{36}$$

In the spirit of classical constrained optimization, one would naturally be interested in a KKT solution (x,y) of (35) or equivalently (25), namely, (x,y) satisfies

$$\tilde{f}(x,y) \le \tilde{f}^*(x), \quad \tilde{g}(x,y) \le 0, \quad \rho(\tilde{f}(x,y) - \tilde{f}^*(x)) = 0, \quad \langle\lambda, \tilde{g}(x,y)\rangle = 0, \tag{37}$$

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

$$\min_{x',y'} \mathcal{L}(x',y',\rho,\lambda) \tag{38}$$

for some $\rho \ge 0$ and $\lambda \in \mathbb{R}^l_+$. Yet, due to the sophisticated problem structure, characterizing a stationary point of (38) is generally difficult. On another hand, notice from Lemma 3 and (36) that problem (38) is equivalent to the minimax problem

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

whose stationary point $(x,y,\tilde{\lambda},z)$, according to Definition 2 and Assumptions 1 and 2, satisfies

$$0 \in \partial f(x,y) + \rho\partial\tilde{f}(x,y) - \rho(\nabla_x\tilde{f}(x,z) + \nabla_x\tilde{g}(x,z)\tilde{\lambda}; 0) + \nabla\tilde{g}(x,y)\lambda, \tag{39}$$

$$0 \in \rho(\partial_z\tilde{f}(x,z) + \nabla_z\tilde{g}(x,z)\tilde{\lambda}), \tag{40}$$

$$\tilde{\lambda} \in \mathbb{R}^l_+, \quad \tilde{g}(x,z) \le 0, \quad \langle\tilde{\lambda}, \tilde{g}(x,z)\rangle = 0.^8 \tag{41}$$

Based on this observation and also the fact that (37) is equivalent to

$$\tilde{f}(x,y) = \tilde{f}^*(x), \quad \tilde{g}(x,y) \le 0, \quad \langle\lambda, \tilde{g}(x,y)\rangle = 0, \tag{42}$$

we are instead interested in a (weak) KKT solution of problem (25) and its inexact counterpart that are defined below.

Definition 4. *The pair (x,y) is said to be a KKT solution of problem* (25) *if there exists $(z,\rho,\lambda,\tilde{\lambda}) \in \mathbb{R}^m \times \mathbb{R}_+ \times \mathbb{R}^l_+ \times \mathbb{R}^l_+$ such that* (39)-(42) *hold. In addition, for any $\varepsilon > 0$, (x,y) is said to be an ε-KKT solution of problem* (25) *if there exists $(z,\rho,\lambda,\tilde{\lambda}) \in \mathbb{R}^m \times \mathbb{R}_+ \times \mathbb{R}^l_+ \times \mathbb{R}^l_+$ such that*

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

where $\tilde{f}^*$ is defined in (26).

[8] The relations in (41) are equivalent to $0 \in -\tilde{g}(x,z) + \partial\mathscr{I}_{\mathbb{R}^l_+}(\tilde{\lambda})$.

We are now ready to present an *operation complexity* of Algorithm 4, 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 (25), whose proof is deferred to Subsection 5.2.

Theorem 5 (Complexity of Algorithm 4). *Suppose that Assumptions 1 and 2 hold. Let f^*, f, $\tilde{f}$, $\tilde{g}$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $\tilde{f}_{\rm hi}$, $\tilde{f}_{\rm low}$, $f_{\rm low}$, $\tilde{f}^*$, $\tilde{f}^*_{\rm hi}$, and $\tilde{g}_{\rm hi}$ be defined in (5), (6), (7), (8), (26), (27) and (28), $L_{\nabla f_1}$, $L_{\nabla \tilde{f}_1}$, $L_{\tilde{f}}$, $L_{\nabla \tilde{g}}$, $L_{\tilde{g}}$ and G be given in Assumptions 1 and 2, ε, ρ, μ, x^0, y^0 and z_ε be given in Algorithm 4, and*

$$\tilde{\lambda} = 2\varepsilon^{-1}[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+, \quad \hat{\lambda} = 2\varepsilon^{-3}[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+, \tag{43}$$

$$\widetilde{L} = L_{\nabla f_1} + 2\varepsilon^{-1} L_{\nabla \tilde{f}_1} + 4\varepsilon^{-3}(\tilde{g}_{\rm hi} L_{\nabla \tilde{g}} + L_{\tilde{g}}^2), \tag{44}$$

$$\tilde{\alpha} = \min\left\{1, \sqrt{4\varepsilon/(D_{\mathbf{y}}\widetilde{L})}\right\},\ \tilde{\delta} = (2+\tilde{\alpha}^{-1})(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\widetilde{L} + \max\left\{\varepsilon/D_{\mathbf{y}}, \tilde{\alpha}\widetilde{L}/4\right\} D_{\mathbf{y}}^2,$$

$$\widetilde{C} = \frac{4\max\{1/(2\widetilde{L}), \min\{D_{\mathbf{y}}\varepsilon^{-1}, 4/(\tilde{\alpha}\widetilde{L})\}\}}{[(3\widetilde{L}+\varepsilon/(2D_{\mathbf{y}}))^2/\min\{\widetilde{L}, \varepsilon/(2D_{\mathbf{y}})\} + 3\widetilde{L} + \varepsilon/(2D_{\mathbf{y}})]^{-2}\varepsilon^5}$$
$$\times \left(\tilde{\delta} + 2\tilde{\alpha}^{-1}[f^* - f_{\rm low} + 2\varepsilon^{-1}(\tilde{f}_{\rm hi} - \tilde{f}_{\rm low}) + \varepsilon^{-3}\tilde{g}_{\rm hi}^2 + \varepsilon D_{\mathbf{y}}/4 + \widetilde{L}(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)]\right),$$

$$\widetilde{K} = \left\lceil 32(1 + f(x^0, y^0) - f_{\rm low} + \varepsilon D_{\mathbf{y}}/4)\widetilde{L}\varepsilon^{-2} + 32\varepsilon^3\left(1 + 4D_{\mathbf{y}}^2\widetilde{L}^2\varepsilon^{-2}\right) - 1\right\rceil_+,$$

$$\widetilde{N} = \left(\left\lceil 96\sqrt{2}\left(1 + (24\widetilde{L} + 4\varepsilon/D_{\mathbf{y}})\widetilde{L}^{-1}\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}}\widetilde{L}\varepsilon^{-1}}\right\}$$
$$\times [(\widetilde{K}+1)(\log\widetilde{C})_+ + \widetilde{K} + 1 + 2\widetilde{K}\log(\widetilde{K}+1)].$$

Then Algorithm 4 outputs an approximate solution $(x_\varepsilon, y_\varepsilon)$ of (25) satisfying

$$\operatorname{dist}\left(0, \partial f(x_\varepsilon, y_\varepsilon) + \rho\partial\tilde{f}(x_\varepsilon, y_\varepsilon) - \rho(\nabla_x\tilde{f}(x_\varepsilon, z_\varepsilon) + \nabla_x\tilde{g}(x_\varepsilon, z_\varepsilon)\tilde{\lambda}; 0) + \nabla\tilde{g}(x_\varepsilon, y_\varepsilon)\hat{\lambda}\right) \le \varepsilon, \tag{45}$$

$$\operatorname{dist}\left(0, \rho(\partial_z\tilde{f}(x_\varepsilon, z_\varepsilon) + \nabla_z\tilde{g}(x_\varepsilon, z_\varepsilon)\tilde{\lambda})\right) \le \varepsilon, \tag{46}$$

$$\|[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+\| \le \varepsilon^2 G^{-1} D_{\mathbf{y}}(\varepsilon^2 + L_{\tilde{f}})/2, \tag{47}$$

$$|\langle\tilde{\lambda}, \tilde{g}(x_\varepsilon, z_\varepsilon)\rangle| \le \varepsilon^2 G^{-2} D_{\mathbf{y}}^2(\rho^{-1}\epsilon + L_{\tilde{f}})^2/2, \tag{48}$$

$$|\tilde{f}(x_\varepsilon, y_\varepsilon) - \tilde{f}^*(x_\varepsilon)| \le \max\left\{\varepsilon\left(1 + f(x^0, y^0) - f_{\rm low} + 2\varepsilon^5(\widetilde{L}^{-1} + 4D_{\mathbf{y}}^2\widetilde{L}\varepsilon^{-2}) + D_{\mathbf{y}}\varepsilon/4\right),\right.$$
$$\left.\varepsilon^2 G^{-2} D_{\mathbf{y}}^2 L_{\tilde{f}}(\varepsilon^2 + \varepsilon L_f + L_{\tilde{f}})/2\right\}, \tag{49}$$

$$\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| \le \varepsilon^2 G^{-1} D_{\mathbf{y}}(\varepsilon^2 + \varepsilon L_f + L_{\tilde{f}})/2, \tag{50}$$

$$|\langle\hat{\lambda}, \tilde{g}(x_\varepsilon, y_\varepsilon)\rangle| \le \varepsilon G^{-2} D_{\mathbf{y}}^2(\varepsilon^2 + \varepsilon L_f + L_{\tilde{f}})^2/2, \tag{51}$$

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

Remark 5. *One can observe from Theorem 5 that $\widetilde{L} = \mathcal{O}(\varepsilon^{-3})$, $\tilde{\alpha} = \mathcal{O}(\varepsilon^2)$, $\tilde{\delta} = \mathcal{O}(\varepsilon^{-5})$, $\widetilde{C} = \mathcal{O}(\varepsilon^{-23})$, $\widetilde{K} = \mathcal{O}(\varepsilon^{-5})$, and $\widetilde{N} = \mathcal{O}(\varepsilon^{-7}\log\varepsilon^{-1})$. As a result, Algorithm 4 enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-7}\log\varepsilon^{-1})$, measured by the amount of evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla \tilde{g}$ and proximal operator of f_2 and $\tilde{f}_2$, for finding an $\mathcal{O}(\varepsilon)$-KKT solution $(x_\varepsilon, y_\varepsilon)$ of (25) satisfying*

$$\operatorname{dist}\left(0, \partial f(x_\varepsilon, y_\varepsilon) + \rho\partial\tilde{f}(x_\varepsilon, y_\varepsilon) - \rho(\nabla_x\tilde{f}(x_\varepsilon, z_\varepsilon) + \nabla_x\tilde{g}(x_\varepsilon, z_\varepsilon)\tilde{\lambda}; 0) + \nabla\tilde{g}(x_\varepsilon, y_\varepsilon)\hat{\lambda}\right) \le \varepsilon,$$

$$\operatorname{dist}\left(0, \rho(\partial_z\tilde{f}(x_\varepsilon, z_\varepsilon) + \nabla_z\tilde{g}(x_\varepsilon, z_\varepsilon)\tilde{\lambda})\right) \le \varepsilon,$$

$$\|[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+\| = \mathcal{O}(\varepsilon^2), \quad |\langle\tilde{\lambda}, \tilde{g}(x_\varepsilon, z_\varepsilon)\rangle| = \mathcal{O}(\varepsilon^2),$$

$$|\tilde{f}(x_\varepsilon, y_\varepsilon) - \tilde{f}^*(x_\varepsilon)| = \mathcal{O}(\varepsilon), \quad \|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| = \mathcal{O}(\varepsilon^2), \quad |\langle\hat{\lambda}, \tilde{g}(x_\varepsilon, y_\varepsilon)\rangle| = \mathcal{O}(\varepsilon),$$

where $\tilde{f}^*$ is defined in (26), $\hat{\lambda}, \tilde{\lambda} \in \mathbb{R}^l_+$ are defined in (43), z_ε is given in Algorithm 4 and $\rho = \varepsilon^{-1}$.

4 Numerical results

In this section we conduct some preliminary experiments to test the performance of our proposed methods (Algorithms 2 and 4) with dynamic update on penalty and tolerance parameters. All the algorithms are coded in Matlab and all the computations are performed on a desktop with a 3.60 GHz Intel i7-12700K 12-core processor and 32 GB of RAM.

4.1 Unconstrained bilevel optimization with linear upper level and quadratic lower level

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

$$\begin{aligned} \min \quad & c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x) \\ \text{s.t.} \quad & y \in \operatorname*{argmin}_z x^T \widetilde{A} z + z^T \widetilde{B} z + \tilde{d}^T z + \mathscr{I}_{[-1,1]^m}(z), \end{aligned} \tag{52}$$

where $\widetilde{A} \in \mathbb{R}^{n\times m}$, $\widetilde{B} \in \mathbb{R}^{m\times m}$, $c \in \mathbb{R}^n$, $d, \tilde{d} \in \mathbb{R}^m$, and $\mathscr{I}_{[-1,1]^n}(\cdot)$ and $\mathscr{I}_{[-1,1]^m}(\cdot)$ are the indicator functions of $[-1,1]^n$ and $[-1,1]^m$ respectively.[9]

For each pair (n, m), we randomly generate 10 instances of problem (52). Specifically, we first randomly generate c, d with all the entries independently chosen from the standard normal distribution, and $\widetilde{A}$ with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.01. We then randomly generate an orthogonal matrix U by performing $U = \text{orth}(\text{randn}(m))$, an $m \times m$ diagonal matrix D with its diagonal entries independently chosen from a normal distribution with mean 0 and standard deviation 0.01 and then projected to $\mathbb{R}_+$, and set $\widetilde{B} = UDU^T$. In addition, we randomly generate $\hat{y} \in [-1,1]^m$ with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.1 and then projected to $[-1,1]^m$, and choose $\tilde{d}$ such that $\hat{y}$ is an optimal solution for the lower level optimization of (52) with $x = 0$.

Notice that (52) is a special case of (5) with $f(x,y) = c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x)$ and $\tilde{f}(x,z) = x^T \widetilde{A} z + z^T \widetilde{B} z + \tilde{d}^T z + \mathscr{I}_{[-1,1]^m}(z)$ and can be suitably solved by Algorithm 2. For the sake of efficiency, we implement a variant of Algorithm 2 with dynamic update on penalty and tolerance parameters. Specifically, we set $\rho_k = 5^{k-1}$, $\varepsilon_k = \rho_k^{-1}$ and $x_{\varepsilon_{-1}} = 0$. For each $k \geq 0$, we run Algorithm 2 with $(\varepsilon, \rho) = (\varepsilon_k, \rho_k)$ and $(x_{\varepsilon_{k-1}}, \tilde{y}_{\varepsilon_{k-1}})$ as the initial point to generate $(x_{\varepsilon_k}, y_{\varepsilon_k})$, where $\tilde{y}_{\varepsilon_{k-1}} \in \operatorname{argmin}_z \tilde{f}(x_{\varepsilon_{k-1}}, z)$ is found by CVX [19]. We terminate the process once $\varepsilon_{\bar{k}} \leq 10^{-4}$ and $(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})$ satisfies

$$\tilde{f}(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}}) - \min_z \tilde{f}(x_{\varepsilon_{\bar{k}}}, z) \leq 10^{-4}$$

for some $\bar{k}$ and output $(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})$ as an approximate solution of (52), where the value of $\min_z \tilde{f}(x_{\varepsilon_{\bar{k}}}, z)$ is computed by CVX.

The computational results of the aforementioned variant of Algorithm 2 for the instances randomly generated above are presented in Table 1. In detail, the values of n and m are listed in the first two columns. For each pair (n, m), the average initial objective value $f(x_{\varepsilon_{-1}}, \tilde{y}_{\varepsilon_{-1}})$ and the average final objective value $f(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})$ over 10 random instances are given in the rest of columns. One can observe that the approximate solution $(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})$ found by this method significantly reduces objective function value compared to the initial point $(x_{\varepsilon_{-1}}, \tilde{y}_{\varepsilon_{-1}})$.

n	m	Initial objective value	Final objective value
100	100	-0.35	-101.67
200	200	-0.53	-194.91
300	300	-0.48	-307.43
400	400	-0.44	-401.71
500	500	-0.05	-527.45
600	600	0.99	-644.53
700	700	0.49	-759.54
800	800	-1.23	-872.77
900	900	-2.07	-1004.27
1000	1000	-1.06	-1107.61

Table 1: Numerical results for problem (52)

4.2 Constrained bilevel linear optimization

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

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

[9]The notation $[-1,1]^n$ denotes the set $\{x \in \mathbb{R}^n | x_i \in [-1,1],\ i = 1, \ldots, n.\}$.

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

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

Notice that (53) is a special case of (25) with $f(x,y) = c^Tx + d^Ty + \mathscr{I}_{[-1,1]^n}(x)$, $\tilde{f}(x,z) = \tilde{d}^Tz + \mathscr{I}_{[-1,1]^m}(z)$ and $\tilde{g}(x,z) = \widetilde{A}x + \widetilde{B}z - \tilde{b}$ and can be suitably solved by Algorithm 4. For the sake of efficiency, we implement a variant of Algorithm 4 with dynamic update on penalty and tolerance parameters. Specifically, we set $\rho_k = 5^{k-1}$, $\mu_k = \rho_k^2$, $\varepsilon_k = \rho_k^{-1}$ and $x_{\varepsilon_{-1}} = 0$. For each $k \geq 0$, we run Algorithm 4 with $(\varepsilon,\rho,\mu) = (\varepsilon_k,\rho_k,\mu_k)$ and $(x_{\varepsilon_{k-1}}, \tilde{y}_{\varepsilon_{k-1}})$ as the initial point to generate $(x_{\varepsilon_k}, y_{\varepsilon_k})$, where $\tilde{y}_{\varepsilon_{k-1}}$ satisfies $\widetilde{P}_{\mu_k}(x_{\varepsilon_{k-1}}, \tilde{y}_{\varepsilon_{k-1}}) \leq \min_z \widetilde{P}_{\mu_k}(x_{\varepsilon_{k-1}}, z) + \varepsilon_k$ with $\widetilde{P}_{\mu_k}$ being given in (30), which can be found by the accelerated proximal gradient method [48]. We terminate the process once $\varepsilon_{\bar{k}} \leq 10^{-4}$ and $(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})$ satisfies

$$\|[\tilde{g}(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})]_+\| \leq 10^{-4}, \quad \tilde{f}(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}}) - \tilde{f}^*(x_{\varepsilon_{\bar{k}}}) \leq 10^{-4}$$

for some $\bar{k}$ and output $(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})$ as an approximate solution of (53), where $\tilde{f}^*$ is defined in (26) and the value $\tilde{f}^*(x_{\varepsilon_{\bar{k}}})$ is computed by CVX [19].

The computational results of the aforementioned variant of Algorithm 4 for the instances randomly generated above are presented in Table 2. In detail, the values of n, m and l are listed in the first three columns. For each triple (n,m,l), the average initial objective value $f(x_{\varepsilon_{-1}}, \hat{y})$ with $\hat{y}$ being generated above[10] and the average final objective value $f(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})$ over 10 random instances are given in the rest of the columns. One can observe that the approximate solution $(x_{\varepsilon_{\bar{k}}}, y_{\varepsilon_{\bar{k}}})$ found by this method significantly reduces objective function value compared to the initial point $(x_{\varepsilon_{-1}}, \hat{y})$.

n	m	l	Initial objective value	Final objective value
100	100	5	-0.51	-34.83
200	200	10	-0.15	-121.41
300	300	15	1.56	-208.44
400	400	20	-0.04	-298.25
500	500	25	1.45	-384.77
600	600	30	0.75	-470.31
700	700	35	0.09	-568.26
800	800	40	-0.98	-629.61
900	900	45	1.21	-689.00
1000	1000	50	1.44	-781.79

Table 2: Numerical results for problem (53)

5 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 1-5.

5.1 Proof of the main results in Section 2

In this subsection we prove Theorems 1, 2 and 3. We first establish a lemma below, which will be used to prove Theorem 1 subsequently.

Lemma 1. *Suppose that Assumption 1 holds and $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ is an ϵ-optimal solution of problem (11) for some $\epsilon > 0$. Let f, $\tilde{f}$, f^*, $f_{\rm low}$ and ρ be given in (5), (8) and (11), respectively. Then we have*

$$\tilde{f}(x_\varepsilon, y_\varepsilon) \leq \min_z \tilde{f}(x_\varepsilon, z) + \rho^{-1}(f^* - f_{\rm low} + 2\epsilon), \quad f(x_\varepsilon, y_\varepsilon) \leq f^* + 2\epsilon.$$

[10] Note that $(x_{\varepsilon_{-1}}, \tilde{y}_{\varepsilon_{-1}})$ may not be a feasible point of (53). Nevertheless, $(x_{\varepsilon_{-1}}, \hat{y})$ is a feasible point of (53) due to $x_{\varepsilon_{-1}} = 0$ and the particular way for generating instances of (53). Besides, (53) can be viewed as an implicit optimization problem in terms of the variable x. It is thus reasonable to use $f(x_{\varepsilon_{-1}}, \hat{y})$ as the initial objective value for the purpose of comparison.

Proof. Since $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ is an ϵ-optimal solution of (11), it follows from Definition 1 that

$$\max_z P_\rho(x_\varepsilon, y_\varepsilon, z) \le P_\rho(x_\varepsilon, y_\varepsilon, z_\varepsilon) + \epsilon, \quad P_\rho(x_\varepsilon, y_\varepsilon, z_\varepsilon) \le \min_{x,y} \max_z P_\rho(x, y, z) + \epsilon.$$

Summing up these inequalities yields

$$\max_z P_\rho(x_\varepsilon, y_\varepsilon, z) \le \min_{x,y} \max_z P_\rho(x, y, z) + 2\epsilon. \tag{54}$$

Let (x^*, y^*) be an optimal solution of (5). 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 (11), one has

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

which implies that

$$\min_{x,y} \max_z P_\rho(x, y, z) \le \max_z P_\rho(x^*, y^*, z) = f^*. \tag{55}$$

It then follows from (11), (54) and (55) that

$$f(x_\varepsilon, y_\varepsilon) + \rho(\tilde{f}(x_\varepsilon, y_\varepsilon) - \min_z \tilde{f}(x_\varepsilon, z)) \overset{(11)}{=} \max_z P_\rho(x_\varepsilon, y_\varepsilon, z) \overset{(54)(55)}{\le} f^* + 2\epsilon,$$

which together with $\tilde{f}(x_\varepsilon, y_\varepsilon) - \min_z \tilde{f}(x_\varepsilon, z) \ge 0$ implies that

$$f(x_\varepsilon, y_\varepsilon) \le f^* + 2\epsilon, \quad \tilde{f}(x_\varepsilon, y_\varepsilon) \le \min_z \tilde{f}(x_\varepsilon, z) + \rho^{-1}\left(f^* - f(x_\varepsilon, y_\varepsilon) + 2\epsilon\right).$$

The conclusion of this lemma directly follows from these and (8). □

We are now ready to prove Theorem 1.

Proof of Theorem 1. Let $\{(x^k, y^k, z^k)\}$ be generated by Algorithm 1 with $\lim_{k\to\infty}(\rho_k, \epsilon_k) = (\infty, 0)$. By considering a convergent subsequence if necessary, we assume without loss of generality that $\lim_{k\to\infty}(x^k, y^k) = (x^*, y^*)$. We now show that (x^*, y^*) is an optimal solution of problem (5). Indeed, since (x^k, y^k, z^k) is an ϵ_k-optimal solution of (11) with $\rho = \rho_k$, it follows from Lemma 1 with $(\rho, \epsilon) = (\rho_k, \epsilon_k)$ and $(x_\varepsilon, y_\varepsilon) = (x^k, y^k)$ that

$$\tilde{f}(x^k, y^k) \le \min_z \tilde{f}(x^k, z) + \rho_k^{-1}(f^* - f_{\text{low}} + 2\epsilon_k), \quad f(x^k, y^k) \le f^* + 2\epsilon_k.$$

By the continuity of f and $\tilde{f}$, $\lim_{k\to\infty}(x^k, y^k) = (x^*, y^*)$, $\lim_{k\to\infty}(\rho_k, \epsilon_k) = (\infty, 0)$, and taking limits as $k \to \infty$ on both sides of the above relations, we obtain that $\tilde{f}(x^*, y^*) \le \min_z \tilde{f}(x^*, z)$ and $f(x^*, y^*) \le f^*$, which clearly imply that $y^* \in \operatorname{argmin}_z \tilde{f}(x^*, z)$ and $f(x^*, y^*) = f^*$. Hence, (x^*, y^*) is an optimal solution of (5) as desired. □

We next prove Theorem 2.

Proof of Theorem 2. Since (x, y) is an ε-KKT solution of problem (5) with its associated $\rho \ge \rho_0$, it follows from Definition 3 that there exists $z \in \mathbb{R}^m$ such that

$$\|\nabla_x f(x, y) + \rho\nabla_x \tilde{f}(x, y) - \rho\nabla_x \tilde{f}(x, z)\| \le \varepsilon, \tag{56}$$

$$\|\nabla_y f(x, y) + \rho\nabla_y \tilde{f}(x, y)\| \le \varepsilon, \tag{57}$$

$$\rho\|\nabla_y \tilde{f}(x, z)\| \le \varepsilon, \quad \tilde{f}(x, y) - \min_{z'} \tilde{f}(x, z') \le \varepsilon. \tag{58}$$

Using (56), the triangle inequality, and the assumptions that $x \in \Omega$, $\nabla f(x', \cdot)$ is L_1-Lipschitz continuous and $\nabla^2 \tilde{f}(x', \cdot)$ is L_2-Lipschitz continuous for all $x' \in \Omega$, we have

$$\begin{aligned}
&\|\nabla_x f(x, y^*(x)) + \rho\nabla^2_{xy} \tilde{f}(x, y^*(x))(y - z)\| \\
&\le \|\nabla_x f(x, y) + \rho\nabla_x \tilde{f}(x, y) - \rho\nabla_x \tilde{f}(x, z)\| + \|\nabla_x f(x, y^*(x)) - \nabla_x f(x, y)\| \\
&\quad + \rho\|\nabla_x \tilde{f}(x, y^*(x)) + \nabla^2_{xy} \tilde{f}(x, y^*(x))(y - y^*(x)) - \nabla_x \tilde{f}(x, y)\| \\
&\quad + \rho\|\nabla_x \tilde{f}(x, z) - \nabla_x \tilde{f}(x, y^*(x)) - \nabla^2_{xy} \tilde{f}(x, y^*(x))(z - y^*(x))\| \\
&\le \varepsilon + L_1\|y - y^*(x)\| + \frac{\rho L_2}{2}\|y - y^*(x)\|^2 + \frac{\rho L_2}{2}\|z - y^*(x)\|^2.
\end{aligned} \tag{59}$$

By (57), (58) and a similar argument as for deriving (59), one has

$$\begin{aligned}
&\|\nabla_y f(x,y^*(x)) + \rho\nabla^2_{yy}\tilde{f}(x,y^*(x))(y-z)\| \\
&\le \|\nabla_y f(x,y^*(x)) - \nabla_y f(x,y)\| + \|\nabla_y f(x,y) + \rho\nabla_y\tilde{f}(x,y)\| + \rho\|\nabla_y\tilde{f}(x,z)\| \\
&\quad + \rho\|\nabla_y\tilde{f}(x,y^*(x)) + \nabla^2_{yy}\tilde{f}(x,y^*(x))(y-y^*(x)) - \nabla_y\tilde{f}(x,y)\| \\
&\quad + \rho\|\nabla_y\tilde{f}(x,z) - \nabla_y\tilde{f}(x,y^*(x)) - \nabla^2_{yy}\tilde{f}(x,y^*(x))(z-y^*(x))\| \\
&\le L_1\|y-y^*(x)\| + 2\varepsilon + \frac{\rho L_2}{2}\|y-y^*(x)\|^2 + \frac{\rho L_2}{2}\|z-y^*(x)\|^2.
\end{aligned}$$

Using this inequality, (16), (19) and (59), we obtain that

$$\begin{aligned}
&\|\nabla\Phi(x)\| = \|\nabla_x f(x,y^*(x)) - \nabla^2_{xy}\tilde{f}(x,y^*(x))[\nabla^2_{yy}\tilde{f}(x,y^*(x))]^{-1}\nabla_y f(x,y^*(x))\| \\
&= \|\nabla_x f(x,y^*(x)) + \rho\nabla^2_{xy}\tilde{f}(x,y^*(x))(y-z) \\
&\quad - \nabla^2_{xy}\tilde{f}(x,y^*(x))[\nabla^2_{yy}\tilde{f}(x,y^*(x))]^{-1}[\nabla_y f(x,y^*(x)) + \rho\nabla^2_{yy}\tilde{f}(x,y^*(x))(y-z)]\| \\
&\le \|\nabla_x f(x,y^*(x)) + \rho\nabla^2_{xy}\tilde{f}(x,y^*(x))(y-z)\| \\
&\quad + \|\nabla^2_{xy}\tilde{f}(x,y^*(x))[\nabla^2_{yy}\tilde{f}(x,y^*(x))]^{-1}\| \cdot \|\nabla_y f(x,y^*(x)) + \rho\nabla^2_{yy}\tilde{f}(x,y^*(x))(y-z)\| \\
&\le (2C+1)\varepsilon + (C+1)\Big(L_1\|y-y^*(x)\| + \frac{\rho L_2}{2}\|y-y^*(x)\|^2 + \frac{\rho L_2}{2}\|z-y^*(x)\|^2\Big). \qquad (60)
\end{aligned}$$

Recall from the assumption that $x \in \Omega \subset \mathcal{N}$ and $\tilde{f}(x',\cdot)$ is strongly convex with modulus $\sigma > 0$ for all $x' \in \mathcal{N}$. It follows from these, (58) and the definition of $y^*(x)$ in (15) that

$$\|y-y^*(x)\|^2 \le 2\sigma^{-1}\big(\tilde{f}(x,y) - \min_{z'}\tilde{f}(x,z')\big) \le 2\sigma^{-1}\varepsilon, \qquad (61)$$

which together with $x \in \Omega$ and (18) implies that $\|\nabla_y f(x,y)\| \le \bar{C}$. Using this, (57), (58), $x \in \Omega \subset \mathcal{N}$, $\nabla_y\tilde{f}(x,y^*(x)) = 0$, and the assumption that $\tilde{f}(x',\cdot)$ is strongly convex with modulus $\sigma > 0$ for all $x' \in \mathcal{N}$, we have

$$\begin{aligned}
\|y-y^*(x)\| &\le \sigma^{-1}\|\nabla_y\tilde{f}(x,y) - \nabla_y\tilde{f}(x,y^*(x))\| = \sigma^{-1}\|\nabla_y\tilde{f}(x,y)\| \\
&\le (\rho\sigma)^{-1}(\|\nabla_y f(x,y) + \rho\nabla_y\tilde{f}(x,y)\| + \|\nabla_y f(x,y)\|) \le (\rho\sigma)^{-1}(\varepsilon + \bar{C}), \qquad (62) \\
\|z-y^*(x)\| &\le \sigma^{-1}\|\nabla_y\tilde{f}(x,z) - \nabla_y\tilde{f}(x,y^*(x))\| = \sigma^{-1}\|\nabla_y\tilde{f}(x,z)\| \le (\rho\sigma)^{-1}\varepsilon. \qquad (63)
\end{aligned}$$

It then follows from (61), (62) and the definition of θ in (19) that $\|y-y^*(x)\| \le \theta$. By this, (60) and (63), one can conclude that (20) holds. In addition, in view of (19), one has $\theta \le \sqrt{2\sigma^{-1}\varepsilon}$ and

$$\rho\theta^2 = \min\big\{\rho^{-1}\sigma^{-2}(\varepsilon+\bar{C})^2,\ 2\rho\sigma^{-1}\varepsilon\big\} \le \min\big\{\rho^{-1}\sigma^{-2}(\varepsilon_0+\bar{C})^2,\ 2\rho\sigma^{-1}\varepsilon\big\} \le \sqrt{2}\sigma^{-3/2}(\varepsilon_0+\bar{C})\sqrt{\varepsilon}.$$

Using these inequalities, (20), $\varepsilon \le \varepsilon_0$ and $\rho \ge \rho_0$, we see that (21) holds. □

We next prove Theorem 3. Before proceeding, we establish a lemma below, which will be used to prove Theorem 3 subsequently.

Lemma 2. *Suppose that Assumption 1 holds and $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ is an ϵ-stationary point of (11). Let $D_{\mathbf{y}}$, $f_{\rm low}$, $\tilde{f}$, ρ, and P_ρ be given in (6), (8) and (11), respectively. Then we have*

$$\begin{aligned}
&\mathrm{dist}\Big(0, \partial f(x_\varepsilon,y_\varepsilon) + \rho\partial\tilde{f}(x_\varepsilon,y_\varepsilon) - (\rho\nabla_x\tilde{f}(x_\varepsilon,z_\varepsilon);0)\Big) \le \epsilon, \quad \mathrm{dist}\big(0,\rho\partial_z\tilde{f}(x_\varepsilon,z_\varepsilon)\big) \le \epsilon, \\
&\tilde{f}(x_\varepsilon,y_\varepsilon) \le \min_z \tilde{f}(x_\varepsilon,z) + \rho^{-1}(\max_z P_\rho(x_\varepsilon,y_\varepsilon,z) - f_{\rm low}).
\end{aligned}$$

Proof. Since $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ is an ϵ-stationary point of (11), it follows from Definition 2 that

$$\mathrm{dist}\big(0,\partial_{x,y}P_\rho(x_\varepsilon,y_\varepsilon,z_\varepsilon)\big) \le \epsilon, \quad \mathrm{dist}\big(0,\partial_z P_\rho(x_\varepsilon,y_\varepsilon,z_\varepsilon)\big) \le \epsilon.$$

Using these and the definition of P_ρ in (11), we have

$$\mathrm{dist}\Big(0, \partial f(x_\varepsilon,y_\varepsilon) + \rho\partial\tilde{f}(x_\varepsilon,y_\varepsilon) - (\rho\nabla_x\tilde{f}(x_\varepsilon,z_\varepsilon);0)\Big) \le \epsilon, \quad \mathrm{dist}\big(0,\rho\partial_z\tilde{f}(x_\varepsilon,z_\varepsilon)\big) \le \varepsilon.$$

In addition, by (11), we have

$$f(x_\varepsilon,y_\varepsilon) + \rho(\tilde{f}(x_\varepsilon,y_\varepsilon) - \min_z\tilde{f}(x_\varepsilon,z)) = \max_z P_\rho(x_\varepsilon,y_\varepsilon,z),$$

which along with (8) implies that

$$\tilde{f}(x_\varepsilon, y_\varepsilon) - \min_z \tilde{f}(x_\varepsilon, z) \le \rho^{-1}(\max_z P_\rho(x_\varepsilon, y_\varepsilon, z) - f_{\rm low}).$$

This completes the proof of this lemma. □

We are now ready to prove Theorem 3.

Proof of Theorem 3. Observe from (12) that problem (11) can be viewed as

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

where $h(x, y, z) = f_1(x, y) + \rho\tilde{f}_1(x, y) - \rho\tilde{f}_1(x, z)$, $p(x, y) = f_2(x) + \rho\tilde{f}_2(y)$, and $q(z) = \rho\tilde{f}_2(z)$. Hence, problem (11) is in the form of (99) with $H = P_\rho$. By Assumption 1 and $\rho = \varepsilon^{-1}$, one can see that h is $\widehat{L}$-smooth on its domain, where $\widehat{L}$ is given in (22). Also, notice from Algorithm 2 that $\epsilon_0 = \varepsilon^{3/2} \le \varepsilon/2$ due to $\varepsilon \in (0, 1/4]$. Consequently, Algorithm 6 can be suitably applied to problem (11) with $\rho = \varepsilon^{-1}$ for finding an ϵ-stationary point $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ of it.

In addition, notice from Algorithm 2 that $\tilde{f}(x^0, y^0) \le \min_y \tilde{f}(x^0, y) + \varepsilon$. Using this, (11) and $\rho = \varepsilon^{-1}$, we obtain

$$\max_z P_\rho(x^0, y^0, z) = f(x^0, y^0) + \rho(\tilde{f}(x^0, y^0) - \min_z \tilde{f}(x^0, z)) \le f(x^0, y^0) + \rho\varepsilon = f(x^0, y^0) + 1. \tag{64}$$

By this and (104) with $H = P_\rho$, $\epsilon = \varepsilon$, $\epsilon_0 = \varepsilon^{3/2}$, $\hat{x}^0 = (x^0, y^0)$, $D_q = D_{\mathbf{y}}$, and $L_{\nabla h} = \widehat{L}$, one has

$$\begin{aligned}\max_z P_\rho(x_\varepsilon, y_\varepsilon, z) \le & \max_z P_\rho(x^0, y^0, z) + \varepsilon D_{\mathbf{y}}/4 + 2\varepsilon^3(\widehat{L}^{-1} + 4D_{\mathbf{y}}^2\widehat{L}\varepsilon^{-2}) \\ \overset{(64)}{\le} & 1 + f(x^0, y^0) + \varepsilon D_{\mathbf{y}}/4 + 2\varepsilon^3(\widehat{L}^{-1} + 4D_{\mathbf{y}}^2\widehat{L}\varepsilon^{-2}).\end{aligned}$$

It then follows from this and Lemma 2 with $\epsilon = \varepsilon$ and $\rho = \varepsilon^{-1}$ that $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ satisfies (23) and (24).

We next show that at most $\widehat{N}$ evaluations of ∇f_1, $\nabla\tilde{f}_1$, and proximal operator of f_2 and $\tilde{f}_2$ are respectively performed in Algorithm 2. Indeed, by (7), (8) and (11), one has

$$\min_{x,y} \max_z P_\rho(x, y, z) \overset{(11)}{=} \min_{x,y}\{f(x, y) + \rho(\tilde{f}(x, y) - \min_z \tilde{f}(x, z))\} \ge \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x, y) \overset{(8)}{=} f_{\rm low}, \tag{65}$$

$$\min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} P_\rho(x, y, z) \overset{(11)}{=} \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}}\{f(x, y) + \rho(\tilde{f}(x, y) - \tilde{f}(x, z))\} \overset{(7)(8)}{\ge} f_{\rm low} + \rho(\tilde{f}_{\rm low} - \tilde{f}_{\rm hi}). \tag{66}$$

For convenience of the rest proof, let

$$H = P_\rho, \quad H^* = \min_{x,y}\max_z P_\rho(x, y, z), \quad H_{\rm low} = \min\{P_\rho(x, y, z) | (x, y, z) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\}. \tag{67}$$

In view of these, (55), (64), (65), (66), and $\rho = \varepsilon^{-1}$, we obtain that

$$\begin{aligned}&\max_z H(x^0, y^0, z) \overset{(64)}{\le} f(x^0, y^0) + 1, \qquad f_{\rm low} \overset{(65)}{\le} H^* \overset{(55)}{\le} f^*, \\ &H_{\rm low} \overset{(66)}{\ge} f_{\rm low} + \rho(\tilde{f}_{\rm low} - \tilde{f}_{\rm hi}) = f_{\rm low} + \varepsilon^{-1}(\tilde{f}_{\rm low} - \tilde{f}_{\rm hi}).\end{aligned}$$

Using these and Theorem 6 with $\epsilon = \varepsilon$, $\hat{x}^0 = (x^0, y^0)$, $D_p = \sqrt{D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2}$, $D_q = D_{\mathbf{y}}$, $\epsilon_0 = \varepsilon^{3/2}$, $L_{\nabla h} = \widehat{L}$, $\alpha = \hat{\alpha}$, $\delta = \hat{\delta}$, and H, H^*, $H_{\rm low}$ given in (67), we can conclude that Algorithm 2 performs at most $\widehat{N}$ evaluations of ∇f_1, $\nabla\tilde{f}_1$ and proximal operator of f_2 and $\tilde{f}_2$ respectively for finding an approximate solution $(x_\varepsilon, y_\varepsilon)$ of problem (5) satisfying (23) and (24). □

5.2 Proof of the main results in Section 3

In this subsection we prove Theorems 4 and 5. Before proceeding, we define

$$r = G^{-1}D_{\mathbf{y}}(\rho^{-1}\epsilon + L_{\tilde{f}}), \quad \mathbb{B}_r^+ = \{\lambda \in \mathbb{R}_+^l : \|\lambda\| \le r\}, \tag{68}$$

where $D_{\mathbf{y}}$ is defined in (6), G is given in Assumption 2(iii), and ϵ and ρ are given in Algorithm 4. In addition, one can observe from (26) and (30) that

$$\min_z \widetilde{P}_\mu(x, z) \le \tilde{f}^*(x) \qquad \forall x \in \mathcal{X}, \tag{69}$$

which will be frequently used later.

We next establish several technical lemmas that will be used to prove Theorem 4 subsequently.

Lemma 3. *Suppose that Assumptions 1 and 2 hold. Let $D_{\mathbf{y}}$, $L_{\tilde{f}}$, G, $\tilde{f}^*$, $\tilde{f}^*_{\rm hi}$ and $\mathbb{B}^+_r$ be given in* (6), (26), (27), (68) *and Assumption 2, respectively. Then the following statements hold.*

(i) $\|\lambda^*\| \le G^{-1}L_{\tilde{f}}D_{\mathbf{y}}$ *and* $\lambda^* \in \mathbb{B}^+_r$ *for all* $\lambda^* \in \Lambda^*(x)$ *and* $x \in \mathcal{X}$, *where* $\Lambda^*(x)$ *denotes the set of optimal Lagrangian multipliers of problem* (26) *for any* $x \in \mathcal{X}$.

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

(iii) It holds that

$$\tilde{f}^*(x) = \max_{\lambda} \min_{z} \tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle - \mathscr{I}_{\mathbb{R}^l_+}(\lambda) \qquad \forall x \in \mathcal{X}, \tag{70}$$

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

Proof. (i) Let $x \in \mathcal{X}$ and $\lambda^* \in \Lambda^*(x)$ be arbitrarily chosen, and let $z^* \in \mathcal{Y}$ be such that (z^*, λ^*) is a pair of primal-dual optimal solutions of (26). It then follows that

$$z^* \in \operatorname*{argmin}_z \tilde{f}(x,z) + \langle \lambda^*, \tilde{g}(x,z)\rangle, \quad \langle \lambda^*, \tilde{g}(x,z^*)\rangle = 0, \quad \tilde{g}(x,z^*) \le 0, \quad \lambda^* \ge 0.$$

The first relation above yields

$$\tilde{f}(x,z^*) + \langle \lambda^*, \tilde{g}(x,z^*)\rangle \le \tilde{f}(x,\hat{z}_x) + \langle \lambda^*, \tilde{g}(x,\hat{z}_x)\rangle,$$

where $\hat{z}_x$ is given in Assumption 2(iii). By this and $\langle \lambda^*, \tilde{g}(x,z^*)\rangle = 0$, one has

$$\langle \lambda^*, -\tilde{g}(x,\hat{z}_x)\rangle \le \tilde{f}(x,\hat{z}_x) - \tilde{f}(x,z^*),$$

which together with $\lambda^* \ge 0$, (6) and Assumption 2 implies that

$$G\sum_{i=1}^l \lambda^*_i \le \langle \lambda^*, -\tilde{g}(x,\hat{z}_x)\rangle \le \tilde{f}(x,\hat{z}_x) - \tilde{f}(x,z^*) \le L_{\tilde{f}}\|\hat{z}_x - z^*\| \le L_{\tilde{f}}D_{\mathbf{y}}, \tag{71}$$

where the first inequality is due to Assumption 2(iii), and the third inequality follows from (6) and $L_{\tilde{f}}$-Lipschitz continuity of $\tilde{f}$ (see Assumption 2(i)). By (68), (71) and $\lambda^* \ge 0$, we have $\|\lambda^*\| \le \sum_{i=1}^l \lambda^*_i \le G^{-1}L_{\tilde{f}}D_{\mathbf{y}}$ and $\lambda^* \in \mathbb{B}^+_r$.

(ii) Recall from Assumptions 1(i) and 2(iii) that $\tilde{f}(x,\cdot)$ and $\tilde{g}_i(x,\cdot)$, $i = 1,\ldots,l$, are convex for any given $x \in \mathcal{X}$. Using this, (26) and the first statement of this lemma, we observe that

$$\tilde{f}^*(x) = \min_z \max_{\lambda \in \mathbb{B}^+_r} \tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle \qquad \forall x \in \mathcal{X}. \tag{72}$$

Notice from Assumption 2 that $\tilde{f}$ and $\tilde{g}$ are Lipschitz continuous on their domain. Then it is not hard to observe that $\max\{\tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle | \lambda \in \mathbb{B}^+_r\}$ is a Lipschitz continuous function of (x,z) on its domain. By this and (72), one can easily verify that $\tilde{f}^*$ is Lipschitz continuous on $\mathcal{X}$. In addition, the finiteness of $\tilde{f}^*_{\rm hi}$ follows from (27), the continuity of $\tilde{f}^*$, and the compactness of $\mathcal{X}$.

(iii) One can observe from (26) that for all $x \in \mathcal{X}$,

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

where the inequality follows from the weak duality. In addition, it follows from Assumption 1 that the domain of $\tilde{f}(x,\cdot)$ is compact for all $x \in \mathcal{X}$. By this, (72) and the strong duality, one has

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

which together with the above inequality implies that (70) holds. □

Lemma 4. *Suppose that Assumptions 1 and 2 hold and that* $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ *is an* ϵ*-optimal solution of problem* (33) *for some* $\epsilon > 0$. *Let* $f_{\rm low}$, f, $\widetilde{P}_\mu$, f^*_μ, ρ *and* μ *be given in* (8), (25), (30), (31) *and* (33), *respectively. Then we have*

$$\widetilde{P}_\mu(x_\varepsilon, y_\varepsilon) \le \min_z \widetilde{P}_\mu(x_\varepsilon, z) + \rho^{-1}(f^*_\mu - f_{\rm low} + 2\epsilon), \qquad f(x_\varepsilon, y_\varepsilon) \le f^*_\mu + 2\epsilon. \tag{73}$$

Proof. The proof follows from the same argument as the one for Lemma 1 with f^* and $\tilde{f}$ being replaced by f^*_μ and $\widetilde{P}_\mu$, respectively. □

Lemma 5. *Suppose that Assumptions 1-3 hold. Let $\tilde{f}_{\rm low}$, f^*, $\tilde{f}^*_{\rm hi}$, f^*_μ be defined in* (7), (25), (27) *and* (31)*, and L_f, ω and $\bar{\theta}$ be given in Assumptions 2 and 3. Suppose that $\mu \geq (\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})/\bar{\theta}^2$. Then we have*

$$f^*_\mu \leq f^* + L_f \omega\Big(\sqrt{\mu^{-1}(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})}\Big). \tag{74}$$

Proof. Let $x \in \mathcal{X}$, $y \in \operatorname{argmin}_z\{\tilde{f}(x,z)|\tilde{g}(x,z) \leq 0\}$ and $z^* \in \operatorname{argmin}_z \widetilde{P}_\mu(x,z)$ be arbitrarily chosen. One can easily see from (30) and (69) that $\tilde{f}(x,z^*) + \mu \|[\tilde{g}(x,z^*)]_+\|^2 \leq \tilde{f}^*(x)$, which together with (7) and (27) implies that

$$\|[\tilde{g}(x,z^*)]_+\|^2 \leq \mu^{-1}(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}). \tag{75}$$

Since $\mu \geq (\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})/\bar{\theta}^2$, it follows from (75) that $\|[\tilde{g}(x,z^*)]_+\| \leq \bar{\theta}$. By this relation, $y \in \operatorname{argmin}_z\{\tilde{f}(x,z)|\tilde{g}(x,z) \leq 0\}$ and Assumption 3, there exists some $\hat{z}^*$ such that

$$\|y - \hat{z}^*\| \leq \omega(\|[\tilde{g}(x,z^*)]_+\|), \qquad \hat{z}^* \in \operatorname*{argmin}_z \Big\{\tilde{f}(x,z)\big|\, \|[\tilde{g}(x,z)]_+\| \leq \|[\tilde{g}(x,z^*)]_+\|\Big\}. \tag{76}$$

In view of (30), $z^* \in \operatorname{argmin}_z \widetilde{P}_\mu(x,z)$ and the second relation in (76), one can observe that $\hat{z}^* \in \operatorname{argmin}_z \widetilde{P}_\mu(x,z)$, which along with (31) yields $f(x,\hat{z}^*) \geq f^*_\mu$. Also, using (76) and L_f-Lipschitz continuity of f (see Assumption 2), we have

$$f(x,y) - f(x,\hat{z}^*) \geq -L_f\|y - \hat{z}^*\| \overset{(76)}{\geq} -L_f\omega(\|[\tilde{g}(x,z^*)]_+\|).$$

Taking minimum over $x \in \mathcal{X}$ and $y \in \operatorname{argmin}_z\{\tilde{f}(x,z)|\tilde{g}(x,z) \leq 0\}$ on both sides of this relation, and using (25), (75), $f(x,\hat{z}^*) \geq f^*_\mu$ and the monotonicity of ω, we can conclude that (74) holds. □

Lemma 6. *Suppose that Assumptions 1-3 hold. Let $\tilde{f}_{\rm low}$, $f_{\rm low}$, f, $\tilde{f}$, f^*, $\tilde{f}^*$, $\tilde{f}^*_{\rm hi}$, ρ and μ be given in* (7), (8), (25), (26), (27) *and* (33)*, and L_f, ω and $\bar{\theta}$ be given in Assumptions 2 and 3, respectively. Suppose that $\mu \geq (\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})/\bar{\theta}^2$ and $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ is an ϵ-optimal solution of problem* (33) *for some $\epsilon > 0$. Then we have*

$$\begin{aligned}
&f(x_\varepsilon, y_\varepsilon) \leq f^* + L_f\omega\Big(\sqrt{\mu^{-1}(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})}\Big) + 2\epsilon,\\
&\tilde{f}(x_\varepsilon, y_\varepsilon) \leq \tilde{f}^*(x_\varepsilon) + \rho^{-1}\Big(f^* - f_{\rm low} + L_f\omega\Big(\sqrt{\mu^{-1}(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})}\Big) + 2\epsilon\Big),\\
&\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\|^2 \leq \mu^{-1}\Big(\tilde{f}^*(x_\varepsilon) - \tilde{f}_{\rm low} + \rho^{-1}\Big(f^* - f_{\rm low} + L_f\omega\Big(\sqrt{\mu^{-1}(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})}\Big) + 2\epsilon\Big)\Big).
\end{aligned}$$

Proof. By (30), (69), and the first relation in (73), one has

$$\tilde{f}(x_\varepsilon, y_\varepsilon) + \mu\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\|^2 \overset{(30)}{=} \widetilde{P}_\mu(x_\varepsilon, y_\varepsilon) \overset{(69)(73)}{\leq} \tilde{f}^*(x_\varepsilon) + \rho^{-1}(f^*_\mu - f_{\rm low} + 2\epsilon).$$

It then follows from this and (7) that

$$\tilde{f}(x_\varepsilon, y_\varepsilon) \leq \tilde{f}^*(x_\varepsilon) + \rho^{-1}(f^*_\mu - f_{\rm low} + 2\epsilon), \quad \|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\|^2 \leq \mu^{-1}(\tilde{f}^*(x_\varepsilon) - \tilde{f}_{\rm low} + \rho^{-1}(f^*_\mu - f_{\rm low} + 2\epsilon)).$$

In addition, recall from (73) that $f(x_\varepsilon, y_\varepsilon) \leq f^*_\mu + 2\epsilon$. The conclusion of this lemma then follows from these three relations and (74). □

We are now ready to prove Theorem 4.

Proof of Theorem 4. Let $\{(x^k, y^k, z^k)\}$ be generated by Algorithm 3 with $\lim_{k\to\infty}(\rho_k, \mu_k, \epsilon_k) = (\infty, \infty, 0)$. By considering a convergent subsequence if necessary, we assume without loss of generality that $\lim_{k\to\infty}(x^k, y^k) = (x^*, y^*)$. We now show that (x^*, y^*) is an optimal solution of problem (25). Indeed, since (x^k, y^k, z^k) is an ϵ_k-optimal solution of (33) with $(\rho, \mu) = (\rho_k, \mu_k)$ and $\lim_{k\to\infty}\mu_k = \infty$, it follows from Lemma 6 with $(\rho, \mu, \epsilon) = (\rho_k, \mu_k, \epsilon_k)$ and $(x_\varepsilon, y_\varepsilon) = (x^k, y^k)$ that for all sufficiently large k, one has

$$\begin{aligned}
&f(x^k, y^k) \leq f^* + L_f\omega\Big(\sqrt{\mu_k^{-1}(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})}\Big) + 2\epsilon_k,\\
&\tilde{f}(x^k, y^k) \leq \tilde{f}^*(x^k) + \rho_k^{-1}\Big(f^* - f_{\rm low} + L_f\omega\Big(\sqrt{\mu_k^{-1}(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})}\Big) + 2\epsilon_k\Big),\\
&\big\|[\tilde{g}(x^k, y^k)]_+\big\|^2 \leq \mu_k^{-1}\Big(\tilde{f}^*(x^k) - \tilde{f}_{\rm low} + \rho_k^{-1}\Big(f^* - f_{\rm low} + L_f\omega\Big(\sqrt{\mu_k^{-1}(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})}\Big) + 2\epsilon_k\Big)\Big).
\end{aligned}$$

By the continuity of f, $\tilde{f}$ and $\tilde{f}^*$ (see Assumption 1(i) and Lemma 3(ii)), $\lim_{k\to\infty}(x^k, y^k) = (x^*, y^*)$, $\lim_{k\to\infty}(\rho_k, \mu_k, \epsilon_k) = (\infty, \infty, 0)$, $\lim_{\theta\downarrow 0}\omega(\theta) = 0$, and taking limits as $k \to \infty$ on both sides of the above relations, we obtain that $f(x^*, y^*) \le f^*$, $\tilde{f}(x^*, y^*) \le \tilde{f}^*(x^*)$ and $[\tilde{g}(x^*, y^*)]_+ = 0$, which along with (25) and (26) imply that $f(x^*, y^*) = f^*$ and $y^* \in \mathrm{argmin}_z\{\tilde{f}(x^*, z)|\tilde{g}(x^*, z) \le 0\}$. Hence, (x^*, y^*) is an optimal solution of (25) as desired. □

We next prove Theorem 5. Before proceeding, we establish several technical lemmas below, which will be used to prove Theorem 5 subsequently.

Lemma 7. *Suppose that Assumptions 1 and 2 hold and that $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ is an ϵ-stationary point of problem (33) for some $\epsilon > 0$. Let $D_{\mathbf{y}}$, $\tilde{g}$, ρ, μ, L_f, $L_{\tilde{f}}$ and G be given in (6), (25), (33) and Assumption 2, respectively. Then we have*

$$\|[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+\| \le (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + L_{\tilde{f}}), \tag{77}$$

$$\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| \le (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}}). \tag{78}$$

Proof. We first prove (77). Since $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ is an ϵ-stationary point of (33), it follows from Definition 2 that $\mathrm{dist}(0, \partial_z P_{\rho,\mu}(x_\varepsilon, y_\varepsilon, z_\varepsilon)) \le \epsilon$. Also, by (30) and (33), one has

$$P_{\rho,\mu}(x, y, z) = f(x, y) + \rho(\tilde{f}(x, y) + \mu\,\|[\tilde{g}(x, y)]_+\|^2) - \rho(\tilde{f}(x, z) + \mu\,\|[\tilde{g}(x, z)]_+\|^2). \tag{79}$$

Using these relations, we have

$$\mathrm{dist}\Big(0, \partial_z \tilde{f}(x_\varepsilon, z_\varepsilon) + 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \nabla_z \tilde{g}_i(x_\varepsilon, z_\varepsilon)\Big) \le \rho^{-1}\epsilon.$$

Hence, there exists $s \in \partial_z \tilde{f}(x_\varepsilon, z_\varepsilon)$ such that

$$\Big\|s + 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \nabla_z \tilde{g}_i(x_\varepsilon, z_\varepsilon)\Big\| \le \rho^{-1}\epsilon. \tag{80}$$

Let $\hat{z}_{x_\varepsilon}$ and G be given in Assumption 2(iii). It then follows that $\hat{z}_{x_\varepsilon} \in \mathcal{Y}$ and $-\tilde{g}_i(x_\varepsilon, \hat{z}_{x_\varepsilon}) \ge G > 0$ for all i. Notice that $[\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \tilde{g}_i(x_\varepsilon, z_\varepsilon) \ge 0$ for all i and $\|z_\varepsilon - \hat{z}_{x_\varepsilon}\| \le D_{\mathbf{y}}$ due to (6). Using these, (80), and the convexity of $\tilde{f}(x_\varepsilon, \cdot)$ and $\tilde{g}_i(x_\varepsilon, \cdot)$ for all i, we have

$$\begin{aligned}
&\tilde{f}(x_\varepsilon, z_\varepsilon) - \tilde{f}(x_\varepsilon, \hat{z}_{x_\varepsilon}) + 2\mu G \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \le \tilde{f}(x_\varepsilon, z_\varepsilon) - \tilde{f}(x_\varepsilon, \hat{z}_{x_\varepsilon}) - 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \tilde{g}_i(x_\varepsilon, \hat{z}_{x_\varepsilon}) \\
&\le \tilde{f}(x_\varepsilon, z_\varepsilon) - \tilde{f}(x_\varepsilon, \hat{z}_{x_\varepsilon}) + 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ (\tilde{g}_i(x_\varepsilon, z_\varepsilon) - \tilde{g}_i(x_\varepsilon, \hat{z}_{x_\varepsilon})) \\
&\le \langle s, z_\varepsilon - \hat{z}_{x_\varepsilon}\rangle + 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \langle \nabla_z \tilde{g}_i(x_\varepsilon, z_\varepsilon), z_\varepsilon - \hat{z}_{x_\varepsilon}\rangle \\
&= \langle s + 2\mu \sum_{i=1}^{l} [\tilde{g}(x_\varepsilon, z_\varepsilon)]_+ \nabla_z \tilde{g}_i(x_\varepsilon, z_\varepsilon), z_\varepsilon - \hat{z}_{x_\varepsilon}\rangle \le \rho^{-1} D_{\mathbf{y}} \epsilon,
\end{aligned} \tag{81}$$

where the first inequality is due to $-\tilde{g}_i(x_\varepsilon, \hat{z}_{x_\varepsilon}) \ge G$ for all i, the second inequality follows from $[\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ \tilde{g}_i(x_\varepsilon, z_\varepsilon) \ge 0$ for all i, the third inequality is due to $s \in \partial_z \tilde{f}(x_\varepsilon, z_\varepsilon)$ and the convexity of $\tilde{f}(x_\varepsilon, \cdot)$ and $\tilde{g}_i(x_\varepsilon, \cdot)$ for all i, and the last inequality follows from (6) and (80). In view of (6), (81), and $L_{\tilde{f}}$-Lipschitz continuity of $\tilde{f}(x, y)$ (see Assumption 2), one has

$$\begin{aligned}
\|[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+\| \le \sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, z_\varepsilon)]_+ &\overset{(81)}{\le} (2\mu G)^{-1}(\rho^{-1} D_{\mathbf{y}} \epsilon + \tilde{f}(x_\varepsilon, \hat{z}_{x_\varepsilon}) - \tilde{f}(x_\varepsilon, z_\varepsilon)) \\
&\le (2\mu G)^{-1}(\rho^{-1} D_{\mathbf{y}} \epsilon + L_{\tilde{f}} \|\hat{z}_{x_\varepsilon} - z_\varepsilon\|) \overset{(6)}{\le} (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + L_{\tilde{f}}).
\end{aligned}$$

Hence, (77) holds.

We next prove (78). Since $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ is an ϵ-stationary point of (33), it follows from Definition 2 that $\mathrm{dist}(0, \partial_{x,y} P_{\rho,\mu}(x_\varepsilon, y_\varepsilon, z_\varepsilon)) \le \epsilon$. In addition, notice from (34) that $P_{\rho,\mu}$ is the sum of a smooth

function and a possibly nonsmooth function that is separable with respect to x, y and z. Consequently, $\partial_{x,y}P_{\rho,\mu} = \partial_x P_{\rho,\mu} \times \partial_y P_{\rho,\mu}$, which together with $\mathrm{dist}(0, \partial_{x,y}P_{\rho,\mu}(x_\varepsilon, y_\varepsilon, z_\varepsilon)) \le \epsilon$ implies that $\mathrm{dist}(0, \partial_y P_{\rho,\mu}(x_\varepsilon, y_\varepsilon, z_\varepsilon)) \le \epsilon$. By this relation and (34), one has

$$\mathrm{dist}\big(0,\ \partial_y f(x_\varepsilon, y_\varepsilon) + \rho\partial_y \tilde{f}(x_\varepsilon, y_\varepsilon) + 2\rho\mu\nabla_y \tilde{g}(x_\varepsilon, y_\varepsilon)[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\big) \le \epsilon.$$

Hence, there exists $s \in \partial_y f(x_\varepsilon, y_\varepsilon)$ and $\tilde{s} \in \partial_y \tilde{f}(x_\varepsilon, y_\varepsilon)$ such that

$$\|s + \rho\tilde{s} + 2\rho\mu\nabla_y \tilde{g}(x_\varepsilon, y_\varepsilon)[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| \le \epsilon. \tag{82}$$

Let $\bar{\mathcal{A}}(x_\varepsilon, y_\varepsilon) = \{i | \tilde{g}_i(x_\varepsilon, y_\varepsilon) > 0, 1 \le i \le l\}$, $\hat{z}_{x_\varepsilon}$ and G be given in Assumption 2(iii). It then follows that $\hat{z}_{x_\varepsilon} \in \mathcal{Y}$ and $-\tilde{g}_i(x_\varepsilon, \hat{z}_{x_\varepsilon}) \ge G > 0$ for all i. Using these and the convexity of $\tilde{g}_i(x_\varepsilon, \cdot)$ for all i, we have

$$\begin{aligned}
&\langle \nabla_y \tilde{g}(x_\varepsilon, y_\varepsilon)[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+, y_\varepsilon - \hat{z}_{x_\varepsilon}\rangle = \sum_{i \in \bar{\mathcal{A}}(x_\varepsilon, y_\varepsilon)} \langle \nabla_y \tilde{g}_i(x_\varepsilon, y_\varepsilon), y_\varepsilon - \hat{z}_{x_\varepsilon}\rangle [g_i(x_\varepsilon, y_\varepsilon)]_+ \\
&\ge \sum_{i \in \bar{\mathcal{A}}(x_\varepsilon, y_\varepsilon)} (\tilde{g}_i(x_\varepsilon, y_\varepsilon) - \tilde{g}_i(x_\varepsilon, \hat{z}_{x_\varepsilon}))[\tilde{g}_i(x_\varepsilon, y_\varepsilon)]_+ \\
&\ge \sum_{i \in \bar{\mathcal{A}}(x_\varepsilon, y_\varepsilon)} G[\tilde{g}_i(x_\varepsilon, y_\varepsilon)]_+ = G\sum_{i=1}^{l} [\tilde{g}_i(x_\varepsilon, y_\varepsilon)]_+ \ge G\,\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\|,
\end{aligned} \tag{83}$$

where the first inequality follows from the convexity of $\tilde{g}(x_\varepsilon, \cdot)$ and the second inequality is due to $-\tilde{g}_i(x_\varepsilon, \hat{z}_{x_\varepsilon}) \ge G$. It then follows from this, (82) and (83) that

$$\begin{aligned}
D_{\mathbf{y}}\epsilon &\ge \|s + \rho\tilde{s} + 2\rho\mu\nabla_y \tilde{g}(x_\varepsilon, y_\varepsilon)[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| \cdot \|y_\varepsilon - \hat{z}_{x_\varepsilon}\| \\
&\ge \langle s + \rho\tilde{s} + 2\rho\mu\nabla_y \tilde{g}(x_\varepsilon, y_\varepsilon)[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+, y_\varepsilon - \hat{z}_{x_\varepsilon}\rangle \\
&= \langle s + \rho\tilde{s}, y_\varepsilon - \hat{z}_{x_\varepsilon}\rangle + 2\rho\mu\langle \nabla_y \tilde{g}(x_\varepsilon, y_\varepsilon)[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+, y_\varepsilon - \hat{z}_{x_\varepsilon}\rangle \\
&\overset{(83)}{\ge} -(\|s\| + \rho\|\tilde{s}\|)\,\|y_\varepsilon - \hat{z}_{x_\varepsilon}\| + 2\rho\mu G\,\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| \\
&\ge -(L_f + \rho L_{\tilde{f}})D_{\mathbf{y}} + 2\rho\mu G\,\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\|,
\end{aligned} \tag{84}$$

where the last inequality follows from $\|y_\varepsilon - \hat{z}_{x_\varepsilon}\| \le D_{\mathbf{y}}$ and the fact that $\|s\| \le L_f$ and $\|\tilde{s}\| \le L_{\tilde{f}}$, which are due to (6), $s \in \partial_y f(x_\varepsilon, y_\varepsilon)$, $\tilde{s} \in \partial_y \tilde{f}(x_\varepsilon, y_\varepsilon)$ and Assumption 2(i). By (84), one can immediately see that (78) holds. □

Lemma 8. *Suppose that Assumptions 1 and 2 hold. Let f, $\tilde{f}$, $\tilde{g}$, $D_{\mathbf{y}}$, f_{low}, $\tilde{f}^*$ and $P_{\rho,\mu}$ be given in* (5), (6), (8), (26) *and* (33), *L_f, $L_{\tilde{f}}$ and G be given in Assumptions 1 and 2, $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ be an ϵ-stationary point of* (33) *for some $\epsilon > 0$, and*

$$\tilde{\lambda} = 2\mu[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+, \quad \hat{\lambda} = 2\rho\mu[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+. \tag{85}$$

Then we have

$$\mathrm{dist}\left(\partial f(x_\varepsilon, y_\varepsilon) + \rho\partial\tilde{f}(x_\varepsilon, y_\varepsilon) - \rho(\nabla_x \tilde{f}(x_\varepsilon, z_\varepsilon) + \nabla_x \tilde{g}(x_\varepsilon, z_\varepsilon)\tilde{\lambda}; 0) + \nabla\tilde{g}(x_\varepsilon, y_\varepsilon)\hat{\lambda}\right) \le \epsilon, \tag{86}$$

$$\mathrm{dist}\left(0, \rho(\partial_z \tilde{f}(x_\varepsilon, z_\varepsilon) + \nabla_z \tilde{g}(x_\varepsilon, z_\varepsilon)\tilde{\lambda})\right) \le \epsilon, \tag{87}$$

$$\|[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+\| \le (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + L_{\tilde{f}}), \tag{88}$$

$$|\langle \tilde{\lambda}, \tilde{g}(x_\varepsilon, z_\varepsilon)\rangle| \le (2\mu)^{-1} G^{-2} D_{\mathbf{y}}^2 (\rho^{-1}\epsilon + L_{\tilde{f}})^2, \tag{89}$$

$$|\tilde{f}(x_\varepsilon, y_\varepsilon) - \tilde{f}^*(x_\varepsilon)| \le \max\left\{\rho^{-1}(\max_z P_{\rho,\mu}(x_\varepsilon, y_\varepsilon, z) - f_{\mathrm{low}}), (2\mu)^{-1} G^{-2} D_{\mathbf{y}}^2 L_{\tilde{f}}(\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}})\right\}, \tag{90}$$

$$\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| \le (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}}), \tag{91}$$

$$|\langle \hat{\lambda}, \tilde{g}(x_\varepsilon, y_\varepsilon)\rangle| \le (2\mu)^{-1}\rho G^{-2} D_{\mathbf{y}}^2 (\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}})^2. \tag{92}$$

Proof. Since $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ is an ϵ-stationary point of (33), it easily follows from (79), (85) and Definition 2 that (86) and (87) hold. Also, it follows from (77) and (78) that (88) and (91) hold. In addition, in view of (85), (88) and (91), one has

$$\begin{aligned}
|\langle \tilde{\lambda}, \tilde{g}(x_\varepsilon, z_\varepsilon)\rangle| &\overset{(85)}{=} 2\mu\,\|[\tilde{g}(x_\varepsilon, z_\varepsilon)]_+\|^2 \overset{(88)}{\le} (2\mu)^{-1} G^{-2} D_{\mathbf{y}}^2 (\rho^{-1}\epsilon + L_{\tilde{f}})^2, \\
|\langle \hat{\lambda}, \tilde{g}(x_\varepsilon, y_\varepsilon)\rangle| &\overset{(85)}{=} 2\rho\mu\,\|[\tilde{g}(x_\varepsilon, y_\varepsilon)]\|_+\|^2 \overset{(91)}{\le} (2\mu)^{-1}\rho G^{-2} D_{\mathbf{y}}^2 (\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}})^2,
\end{aligned}$$

and hence (89) and (92) hold. Also, observe from the definition of $P_{\rho,\mu}$ in (33) that

$$\widetilde{P}_\mu(x_\varepsilon, y_\varepsilon) - \min_z \widetilde{P}_\mu(x_\varepsilon, z) = \rho^{-1}(\max_z P_{\rho,\mu}(x_\varepsilon, y_\varepsilon, z) - f(x_\varepsilon, y_\varepsilon)).$$

Using this, (8), (30) and (69), we obtain that

$$\begin{aligned}\tilde{f}(x_\varepsilon, y_\varepsilon) + \mu \|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\|^2 \overset{(30)}{=} \widetilde{P}_\mu(x_\varepsilon, y_\varepsilon) = &\min_z \widetilde{P}_\mu(x_\varepsilon, z) + \rho^{-1}(\max_z P_{\rho,\mu}(x_\varepsilon, y_\varepsilon, z) - f(x_\varepsilon, y_\varepsilon)) \\ &\overset{(8)(69)}{\leq} \tilde{f}^*(x_\varepsilon) + \rho^{-1}(\max_z P_{\rho,\mu}(x_\varepsilon, y_\varepsilon, z) - f_{\rm low}). \end{aligned} \tag{93}$$

On the other hand, let $\lambda^* \in \mathbb{R}^l_+$ be an optimal Lagrangian multiplier of problem (26) with $x = x_\varepsilon$. It then follows from Lemma 3(i) that $\|\lambda^*\| \leq G^{-1} L_{\tilde{f}} D_{\mathbf{y}}$. Using these and (91), we have

$$\begin{aligned}\tilde{f}^*(x_\varepsilon) = &\min_y \left\{ \tilde{f}(x_\varepsilon, y) + \langle \lambda^*, \tilde{g}(x_\varepsilon, y)\rangle \right\} \leq \tilde{f}(x_\varepsilon, y_\varepsilon) + \langle \lambda^*, \tilde{g}(x_\varepsilon, y_\varepsilon)\rangle \\ &\leq \tilde{f}(x_\varepsilon, y_\varepsilon) + \|\lambda^*\| \|[\tilde{g}(x_\varepsilon, y_\varepsilon)]_+\| \leq \tilde{f}(x_\varepsilon, y_\varepsilon) + (2\mu)^{-1} G^{-2} D_{\mathbf{y}}^2 L_{\tilde{f}} (\rho^{-1}\epsilon + \rho^{-1} L_f + L_{\tilde{f}}).\end{aligned}$$

By this and (93), one can see that (90) holds. □

We are now ready to prove Theorem 5.

Proof of Theorem 5. Observe from (34) that problem (33) can be viewed as

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

where $h(x, y, z) = f_1(x, y) + \rho \tilde{f}_1(x, y) + \rho\mu \|[\tilde{g}(x, y)]_+\|^2 - \rho \tilde{f}_1(x, z) - \rho\mu \|[\tilde{g}(x, z)]_+\|^2$, $p(x, y) = f_2(x) + \rho \tilde{f}_2(y)$ and $q(z) = \rho \tilde{f}_2(z)$. Hence, problem (33) is in the form of (99) with $H = P_{\rho,\mu}$. By Assumption 1, (28), (29), $\rho = \varepsilon^{-1}$ and $\mu = \varepsilon^{-2}$, one can see that h is $\widetilde{L}$-smooth on its domain, where $\widetilde{L}$ is given in (44). Also, notice from Algorithm 4 that $\epsilon_0 = \varepsilon^{5/2} \leq \varepsilon/2 = \epsilon/2$ due to $\varepsilon \in (0, 1/4]$. Consequently, Algorithm 6 can be suitably applied to problem (33) with $\rho = \varepsilon^{-1}$ and $\mu = \varepsilon^{-2}$ for finding an ϵ-stationary point $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ of it.

In addition, notice from Algorithm 4 that $\widetilde{P}_\mu(x^0, y^0) \leq \min_y \widetilde{P}_\mu(x^0, y) + \varepsilon$. Using this, (33) and $\rho = \varepsilon^{-1}$, we obtain

$$\max_z P_{\rho,\mu}(x^0, y^0, z) \overset{(33)}{=} f(x^0, y^0) + \rho(\widetilde{P}_\mu(x^0, y^0) - \min_z \widetilde{P}_\mu(x^0, z)) \leq f(x^0, y^0) + \rho\varepsilon = f(x^0, y^0) + 1. \tag{94}$$

By this and (104) with $H = P_{\rho,\mu}$, $\epsilon = \varepsilon$, $\epsilon_0 = \varepsilon^{5/2}$, $\hat{x}^0 = (x^0, y^0)$, $D_q = D_{\mathbf{y}}$ and $L_{\nabla h} = \widetilde{L}$, one has

$$\begin{aligned}\max_z P_{\rho,\mu}(x_\varepsilon, y_\varepsilon, z) \leq &\max_z P_{\rho,\mu}(x^0, y^0, z) + \varepsilon D_{\mathbf{y}}/4 + 2\varepsilon^5(\widetilde{L}^{-1} + 4D_{\mathbf{y}}^2 \widetilde{L}\varepsilon^{-2}) \\ &\overset{(94)}{\leq} 1 + f(x^0, y^0) + \varepsilon D_{\mathbf{y}}/4 + 2\varepsilon^5(\widetilde{L}^{-1} + 4D_{\mathbf{y}}^2 \widetilde{L}\varepsilon^{-2}).\end{aligned}$$

It then follows from this and Lemma 8 with $\epsilon = \varepsilon$, $\rho = \varepsilon^{-1}$ and $\mu = \varepsilon^{-2}$ that $(x_\varepsilon, y_\varepsilon, z_\varepsilon)$ satisfies the relations (45)-(51).

We next show that at most $\widetilde{N}$ evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla \tilde{g}$ and proximal operator of f_2 and $\tilde{f}_2$ are respectively performed in Algorithm 4. Indeed, by (7), (8), (28), (30) and (33), one has

$$\min_{x,y} \max_z P_{\rho,\mu}(x, y, z) \overset{(33)}{=} \min_{x,y}\{f(x, y) + \rho(\widetilde{P}_\mu(x, y) - \min_z \widetilde{P}_\mu(x, z))\} \geq \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x, y) \overset{(8)}{=} f_{\rm low}, \tag{95}$$

$$\begin{aligned}&\min\{P_{\rho,\mu}(x, y, z) | (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}\} \overset{(33)}{=} \min\{f(x, y) + \rho(\widetilde{P}_\mu(x, y) - \widetilde{P}_\mu(x, z)) | (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}\} \\ &\overset{(30)}{=} \min\{f(x, y) + \rho(\tilde{f}(x, y) + \mu\|[\tilde{g}(x, y)]_+\|^2 - \tilde{f}(x, z) - \mu\|[\tilde{g}(x, z)]_+\|^2) | (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}\} \\ &\geq f_{\rm low} + \rho(\tilde{f}_{\rm low} - \tilde{f}_{\rm hi}) - \rho\mu\tilde{g}_{\rm hi}^2,\end{aligned} \tag{96}$$

where the last inequality follows from (7), (8) and (28). In addition, let (x^*, y^*) be an optimal solution of (25). It then follows that $f(x^*, y^*) = f^*$ and $[\tilde{g}(x^*, y^*)]_+ = 0$. By these, (7), (30) and (33), one has

$$\begin{aligned}\min_{x,y} \max_z P_{\rho,\mu}(x, y, z) \leq &\max_z P_{\rho,\mu}(x^*, y^*, z) \overset{(33)}{=} f(x^*, y^*) + \rho\left(\widetilde{P}_\mu(x^*, y^*) - \min_z \widetilde{P}_\mu(x^*, z)\right) \\ &\overset{(30)}{=} f(x^*, y^*) + \rho(\tilde{f}(x^*, y^*) + \mu\|[\tilde{g}(x^*, y^*)]_+\|^2 - \min_z\{\tilde{f}(x^*, z) + \mu\|[\tilde{g}(x^*, z)]_+\|^2\}) \\ &\overset{(7)}{\leq} f^* + \rho(\tilde{f}_{\rm hi} - \tilde{f}_{\rm low}).\end{aligned} \tag{97}$$

For convenience of the rest proof, let

$$H = P_{\rho,\mu}, \quad H^* = \min_{x,y} \max_z P_{\rho,\mu}(x,y,z), \quad H_{\rm low} = \min\{P_{\rho,\mu}(x,y,z) | (x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}\}. \tag{98}$$

In view of these, (94), (95), (96), (97), $\rho = \varepsilon^{-1}$ and $\mu = \varepsilon^{-2}$, we obtain that

$$\max_z H(x^0, y^0, z) \overset{(94)}{\leq} f(x^0, y^0) + 1, \quad f_{\rm low} \overset{(95)}{\leq} H^* \overset{(97)}{\leq} f^* + \rho(\tilde{f}_{\rm hi} - \tilde{f}_{\rm low}) = f^* + \varepsilon^{-1}(\tilde{f}_{\rm hi} - \tilde{f}_{\rm low}),$$
$$H_{\rm low} \overset{(96)}{\geq} f_{\rm low} + \rho(\tilde{f}_{\rm low} - \tilde{f}_{\rm hi}) - \rho\mu\tilde{g}_{\rm hi}^2 = f_{\rm low} + \varepsilon^{-1}(\tilde{f}_{\rm low} - \tilde{f}_{\rm hi}) - \varepsilon^{-3}\tilde{g}_{\rm hi}^2.$$

Using these and Theorem 6 with $\epsilon = \varepsilon$, $\hat{x}^0 = (x^0, y^0)$, $D_p = \sqrt{D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2}$, $D_q = D_{\mathbf{y}}$, $\epsilon_0 = \varepsilon^{5/2}$, $L_{\nabla h} = \widetilde{L}$, $\alpha = \tilde{\alpha}$, $\delta = \tilde{\delta}$, and H, H^*, $H_{\rm low}$ given in (98), we can conclude that Algorithm 4 performs at most $\widetilde{N}$ evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla \tilde{g}$ and proximal operator of f_2 and $\tilde{f}_2$ for finding an approximate solution $(x_\varepsilon, y_\varepsilon)$ of problem (25) satisfying (45)-(51). □

6 Concluding remarks

For the sake of simplicity, first-order penalty methods are proposed only for problem (3) in this paper. It would be interesting to extend them to problem (1) by using a standard technique (e.g., see [49]) for handling the constraint $g(x,y) \leq 0$. This will be left for the future research.

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

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

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

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

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

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

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

[14] A. L. Dontchev and R. T. Rockafellar. *Implicit functions and solution mappings*, volume 543. Springer, 2009.

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

[16] 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.

[17] 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.

[18] S. Ghadimi and M. Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

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

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

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

[28] 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.

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

[30] 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.

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

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

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

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

[35] 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*, pages 7426–7434, 2022.

[36] 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.

[37] 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.

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

[39] 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.

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

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

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

[43] 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*.

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

[45] 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.

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

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

[48] Y. E. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140:125–161, 2013.

[49] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 1999.

[50] 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.

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

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

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

[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.

[64] R. Zhao. A primal-dual smoothing framework for max-structured non-convex optimization. *Mathematics of Operations Research*, 2023.

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

In this part, we aim to find an ϵ-stationary point of the nonconvex-concave minimax problem

$$H^* = \min_x \max_y \{H(x,y) \coloneqq h(x,y) + p(x) - q(y)\}, \tag{99}$$

which has at least one optimal solution and satisfies the following assumptions.

Assumption 4. *(i) $p : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $q : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ are proper convex functions and continuous on $\operatorname{dom} p$ and $\operatorname{dom} q$, respectively, and moreover, $\operatorname{dom} p$ and $\operatorname{dom} q$ are compact.*

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

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

Recently, an accelerated inexact proximal point smoothing (AIPP-S) scheme was proposed in [32] for finding an approximate stationary point of a class of minimax composite nonconvex optimization problems, which includes (99) as a special case. When applied to (99), AIPP-S requires the exact solution of $\max_y \left\{h(x',y) - q(y) - \frac{1}{2\lambda}\|y-y'\|^2\right\}$ for any $\lambda > 0$, $x' \in \mathbb{R}^n$, and $y' \in \mathbb{R}^m$. However, h is typically sophisticated and the *exact* solution of such problem usually cannot be found. Consequently, AIPP-S is generally not implementable for (99). In addition, a first-order method was proposed in [64] which enjoys a first-order oracle complexity of $\mathcal{O}(\varepsilon^{-3}\log\varepsilon^{-1})$ for finding an ϵ-primal stationary point x' of (99) that satisfies

$$\left\|\lambda^{-1}(x' - \operatorname*{argmin}_x \left\{\max_y H(x,y) + \frac{1}{2\lambda}\|x-x'\|^2\right\}\right\| \le \epsilon$$

for some $0 < \lambda < L_{\nabla h}^{-1}$. Yet, this method does not suit our needs since our aim is to find an ϵ-stationary point of (99) introduced in Definition 2. In what follows, we present a first-order method proposed in [41, Algorithm 2] for finding such an ϵ-stationary point of (99).

For ease of presentation, we define

$$D_p = \max\{\|u-v\| \big| u, v \in \operatorname{dom} p\}, \quad D_q = \max\{\|u-v\| \big| u, v \in \operatorname{dom} q\}, \tag{100}$$

$$H_{\text{low}} = \min\{H(x,y) | (x,y) \in \operatorname{dom} p \times \operatorname{dom} q\}. \tag{101}$$

Given an iterate (x^k, y^k), the first-order method [41, Algorithm 2] finds the next iterate (x^{k+1}, y^{k+1}) by applying [41, Algorithm 1], which is a slight modification of a novel optimal first-order method [33, Algorithm 4], to the strongly-convex-strongly-concave minimax problem

$$\min_x \max_y \left\{h_k(x,y) = h(x,y) - \epsilon\|y-y^0\|^2/(4D_q) + L_{\nabla h}\|x-x^k\|^2\right\}. \tag{102}$$

For ease of reference, we next present a modified optimal first-order method [41, Algorithm 1] in Algorithm 5 below for solving the strongly-convex-strongly-concave minimax problem

$$\min_x \max_y \left\{\bar{h}(x,y) + p(x) - q(y)\right\}, \tag{103}$$

where $\bar{h}(x,y)$ is σ_x-strongly-convex-σ_y-strongly-concave and $L_{\nabla\bar{h}}$-smooth on $\operatorname{dom} p \times \operatorname{dom} q$ for some $\sigma_x, \sigma_y > 0$. In Algorithm 5, the functions $\hat{h}$, a_x^k and a_y^k are defined as follows:

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

where y_g^k and z_g^k are generated at iteration k of Algorithm 5 below.

Algorithm 5 A modified optimal first-order method for problem (103)

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

for $k = 0, 1, 2, \ldots$ **do**
  $(z_g^k, y_g^k) = \bar{\alpha}(z^k, y^k) + (1-\bar{\alpha})(z_f^k, y_f^k)$.
  $(x^{k,-1}, y^{k,-1}) = (-\sigma_x^{-1}z_g^k, y_g^k)$.
  $x^{k,0} = \operatorname{prox}_{\zeta\gamma_x p}(x^{k,-1} - \zeta\gamma_x a_x^k(x^{k,-1}, y^{k,-1}))$.
  $y^{k,0} = \operatorname{prox}_{\zeta\gamma_y q}(y^{k,-1} - \zeta\gamma_y a_y^k(x^{k,-1}, y^{k,-1}))$.
  $b_x^{k,0} = \frac{1}{\zeta\gamma_x}(x^{k,-1} - \zeta\gamma_x a_x^k(x^{k,-1}, y^{k,-1}) - x^{k,0})$.
  $b_y^{k,0} = \frac{1}{\zeta\gamma_y}(y^{k,-1} - \zeta\gamma_y a_y^k(x^{k,-1}, y^{k,-1}) - y^{k,0})$.
  $t = 0$.
  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**
    $x^{k,t+1/2} = x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x(a_x^k(x^{k,t}, y^{k,t}) + b_x^{k,t})$.
    $y^{k,t+1/2} = y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y(a_y^k(x^{k,t}, y^{k,t}) + b_y^{k,t})$.
    $x^{k,t+1} = \operatorname{prox}_{\zeta\gamma_x p}(x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x a_x^k(x^{k,t+1/2}, y^{k,t+1/2}))$.
    $y^{k,t+1} = \operatorname{prox}_{\zeta\gamma_y q}(y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y a_y^k(x^{k,t+1/2}, y^{k,t+1/2}))$.
    $b_x^{k,t+1} = \frac{1}{\zeta\gamma_x}(x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x a_x^k(x^{k,t+1/2}, y^{k,t+1/2}) - x^{k,t+1})$.
    $b_y^{k,t+1} = \frac{1}{\zeta\gamma_y}(y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y a_y^k(x^{k,t+1/2}, y^{k,t+1/2}) - y^{k,t+1})$.
    $t \leftarrow t + 1$.
  end while
  $(x_f^{k+1}, y_f^{k+1}) = (x^{k,t}, y^{k,t})$.
  $(z_f^{k+1}, w_f^{k+1}) = (\nabla_x\hat{h}(x_f^{k+1}, y_f^{k+1}) + b_x^{k,t}, -\nabla_y\hat{h}(x_f^{k+1}, y_f^{k+1}) + b_y^{k,t})$.
  $z^{k+1} = z^k + \eta_z\sigma_x^{-1}(z_f^{k+1} - z^k) - \eta_z(x_f^{k+1} + \sigma_x^{-1}z_f^{k+1})$.
  $y^{k+1} = y^k + \eta_y\sigma_y(y_f^{k+1} - y^k) - \eta_y(w_f^{k+1} + \sigma_y y_f^{k+1})$.
  $x^{k+1} = -\sigma_x^{-1}z^{k+1}$.
  $\hat{x}^{k+1} = \operatorname{prox}_{\hat{\zeta}p}(x^{k+1} - \hat{\zeta}\nabla_x\bar{h}(x^{k+1}, y^{k+1}))$.
  $\hat{y}^{k+1} = \operatorname{prox}_{\hat{\zeta}q}(y^{k+1} + \hat{\zeta}\nabla_y\bar{h}(x^{k+1}, y^{k+1}))$.
  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 [41, Algorithm 2] for finding an ϵ-stationary point of (99) in Algorithm 6 below.

[11] For convenience, $-\sigma_x \operatorname{dom} p$ stands for the set $\{-\sigma_x u | u \in \operatorname{dom} p\}$.

Algorithm 6 A first-order method for problem (99)

Input: $\epsilon > 0$, $\epsilon_0 \in (0, \epsilon/2]$, $(\hat{x}^0, \hat{y}^0) \in \operatorname{dom} p \times \operatorname{dom} q$, $(x^0, y^0) = (\hat{x}^0, \hat{y}^0)$, and $\epsilon_k = \epsilon_0/(k+1)$.

1: **for** $k = 0, 1, 2, \ldots$ **do**

2: Call Algorithm 5 with $\bar{h} \leftarrow h_k$, $\tau \leftarrow \epsilon_k$, $\sigma_x \leftarrow L_{\nabla h}$, $\sigma_y \leftarrow \epsilon/(2D_q)$, $L_{\nabla \bar{h}} \leftarrow 3L_{\nabla h} + \epsilon/(2D_q)$, $\bar{z}^0 = z_f^0 \leftarrow -\sigma_x x^k$, $\bar{y}^0 = y_f^0 \leftarrow y^k$, and denote its output by (x^{k+1}, y^{k+1}), where h_k is given in (102).

3: Terminate the algorithm and output $(x_\varepsilon, y_\varepsilon) = (x^{k+1}, y^{k+1})$ if

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

4: **end for**

The following theorem presents the iteration complexity of Algorithm 6, whose proof is given in [41, Theorem 2].

Theorem 6 (**Complexity of Algorithm 6**). *Suppose that Assumption 4 holds. Let H^*, H D_p, D_q, and H_{low} be defined in* (99), (100) *and* (101), *$L_{\nabla h}$ be given in Assumption 4, ϵ, ϵ_0 and x^0 be given in Algorithm 6, and*

$$\begin{aligned}
\alpha =& \ \min\left\{1, \sqrt{4\epsilon/(D_q L_{\nabla h})}\right\},\\
\delta =& \ (2+\alpha^{-1}) L_{\nabla h} D_p^2 + \max\left\{\epsilon/D_q, \alpha L_{\nabla h}/4\right\} D_q^2,\\
K =& \ \left\lceil 16(\max_y H(x^0, y) - H^* + \epsilon D_q/4) L_{\nabla h} \epsilon^{-2} + 32\epsilon_0^2 (1 + 4D_q^2 L_{\nabla h}^2 \epsilon^{-2})\epsilon^{-2} - 1 \right\rceil_+,\\
N =& \ \left(\left\lceil 96\sqrt{2}\left(1 + (24L_{\nabla h} + 4\epsilon/D_q) L_{\nabla h}^{-1}\right)\right\rceil + 2\right)\left\{2, \sqrt{D_q L_{\nabla h} \epsilon^{-1}}\right\}\\
&\times \left((K+1)\left(\log \frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{D_q}{\epsilon}, \frac{4}{\alpha L_{\nabla h}}\right\}\right\}\left(\delta + 2\alpha^{-1}(H^* - H_{\text{low}} + \epsilon D_q/4 + L_{\nabla h} D_p^2)\right)}{\left[(3L_{\nabla h} + \epsilon/(2D_q))^2/\min\{L_{\nabla h}, \epsilon/(2D_q)\} + 3L_{\nabla h} + \epsilon/(2D_q)\right]^{-2}\epsilon_0^2} \right)_+ \right.\\
&\left. + K + 1 + 2K\log(K+1)\right).
\end{aligned}$$

Then Algorithm 6 terminates and outputs an ϵ-stationary point $(x_\varepsilon, y_\varepsilon)$ of (99) *in at most $K+1$ outer iterations that satisfies*

$$\max_y H(x_\varepsilon, y) \le \max_y H(\hat{x}^0, y) + \epsilon D_q/4 + 2\epsilon_0^2\left(L_{\nabla h}^{-1} + 4D_q^2 L_{\nabla h}\epsilon^{-2}\right). \tag{104}$$

Moreover, the total number of evaluations of ∇h and proximal operator of p and q performed in Algorithm 6 is no more than N, respectively.