First-order penalty methods for bilevel optimization

Zhaosong Lu * Sanyou Mei *

January 4, 2023

Abstract

In this paper we study a class of unconstrained and constrained bilevel optimization problems in which the lower-level part is a convex optimization problem, while the upper-level part is possibly a nonconvex optimization problem. In particular, we propose penalty methods for solving them, whose subproblems turn out to be a structured minimax problem and are suitably solved by a first-order method developed in this paper. Under some 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. To the best of our knowledge, the methodology and results in this paper are new.

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 partial or full decision variables in the upper level are also involved in the lower level. Generically, it can be written in the following form:

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

Bilevel optimization has found a variety of important applications, including adversarial training [36, 37, 46], continual learning [32], hyperparameter tuning [3, 17], image reconstruction [9], meta-learning [4, 23, 42], neural architecture search [15, 30], reinforcement learning [20, 27], and Stackelberg games [48]. More applications about it can be found in [2, 8, 10, 11, 12, 44] and the references therein. Theoretical properties including optimality conditions of (1) have been extensively studied in the literature (e.g., see [12, 13, 34, 47, 50]).

Numerous methods have been developed for solving some special cases of (1). For example, constraint-based methods [19, 43], deterministic gradient-based methods [16, 17, 21, 35, 41, 42], and stochastic gradient-based methods [6, 18, 20, 24, 26] 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. Besides, when all the functions involved in (1) are smooth and $\tilde{f}$, $\tilde{g}$ are convex with respect to y, gradient-type methods were proposed by solving the mathematical program with equilibrium constraints (MPEC) resulting from replacing the lower-level optimization problem of (1) by its first-order optimality conditions (e.g., see [1, 33, 40]). Recently, difference-of-convex (DC) algorithms were developed in [51] for solving (1) with $g\equiv 0$, f being a DC function, and $\tilde{f}$, $\tilde{g}$ being convex functions. In addition, a double penalty method [22] was proposed for (1), which solves a sequence of bilevel optimization problems of the form

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

*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.

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 [22] is typically not implementable. More discussion on algorithmic development for bilevel optimization can be found in [2, 8, 12, 31, 45, 47]) 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,

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

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

The main goal of this paper is to develop a first-order penalty method for solving problem (3). Our key observations toward this development are: (i) problem (3) can be approximately solved as a penalty problem (see (49)); (ii) such a penalty problem is equivalent to a structured minimax problem (see (50)), which can be suitably solved by a first-order method proposed in Section 2. As a result, these observations lead to development of a novel first-order penalty method for solving (3) (see Sections 3 and 4), which enjoys the following appealing features.

- It uses only the first-order information of the problem. Specifically, its fundamental operations consist only of evaluations of the gradient of $\tilde{g}$ and the smooth component of f and $\tilde{f}$ and also the proximal operator of the nonsmooth component of f and $\tilde{f}$. Thus, it is suitable for solving large-scale problems (see Sections 3 and 4).

- It has theoretical guarantees on operation complexity, which is measured by the aforementioned fundamental operations, for finding an ε-KKT solution of (3). In particular, 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 4 and 6).

- 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, the methodology and results in this paper are new.

The rest of this paper is organized as follows. In Subsection 1.1 we introduce some notation and terminology. In Section 2 we propose a first-order method for solving a nonconvex-concave minimax problem and study its complexity. In Sections 3 and 4, we propose first-order penalty methods for unconstrained and constrained bilevel optimization and study their complexity, respectively. In Section 5 we present the proofs of the main results. 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 denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively. For any $v \in \mathbb{R}^n$, let v_+ denote the nonnegative part of v, that is, $(v_+)_i = \max\{v_i, 0\}$ for all i. For any two vectors u and v, $(u;v)$ denotes the vector resulting from stacking v under u. Given a point x and a closed set S in $\mathbb{R}^n$, let $\operatorname{dist}(x,S) = \min_{x' \in S}\|x'-x\|$ 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 $x, x' \in S$. For a closed convex function $p : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$,[2] the *proximal operator* associated with p is denoted by prox_p, that is,

$$\operatorname{prox}_p(x) = \arg\min_{x' \in \mathbb{R}^n}\left\{\frac{1}{2}\|x'-x\|^2 + p(x')\right\} \quad \forall x \in \mathbb{R}^n. \tag{4}$$

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

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

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

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

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

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

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, 49] for more details.

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

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

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_ϵ, y_ϵ) is called an ϵ-optimal solution of the minimax problem* (5) *if*

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

Definition 2. *A point (x, y) is called a stationary point of the minimax problem* (5) *if*

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

In addition, for any $\epsilon > 0$, a point (x_ϵ, y_ϵ) is called an ϵ-stationary point of the minimax problem (5) *if*

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

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

In this section, we propose a first-order method for finding an approximate stationary point of a nonconvex-concave minimax problem, which will be used as a subproblem solver for the penalty methods proposed in Sections 3 and 4. In particular, we consider the minimax problem

$$H^* = \min_x \max_y \left\{ H(x, y) := h(x, y) + p(x) - q(y) \right\}. \tag{6}$$

Assume that problem (6) has at least one optimal solution. In addition, h, p and q satisfy the following assumptions.

Assumption 1. *(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 their domain, and moreover, their domain is compact.*

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

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

Recently, an accelerated inexact proximal point smoothing (AIPP-S) scheme was proposed in [28] for finding an approximate stationary point of a class of minimax composite nonconvex optimization problems, which includes (6) as a special case. When applied to (6), AIPP-S requires the availability of the oracle including exact evaluation of $\nabla_x h(x,y)$ and

$$\arg\min_x \left\{ p(x) + \frac{1}{2\lambda}\|x-x'\|^2 \right\}, \qquad \arg\max_y \left\{ h(x',y) - q(y) - \frac{1}{2\lambda}\|y-y'\|^2 \right\} \tag{7}$$

for any $\lambda > 0$, $x' \in \mathbb{R}^n$ and $y' \in \mathbb{R}^m$. Note that h is typically sophisticated and the *exact* solution of the second problem in (7) usually cannot be found. As a result, AIPP-S is generally *not implementable* for (6), though an oracle complexity of $\mathcal{O}(\epsilon^{-5/2})$ was established in [28] for it to find an ϵ-stationary point of (6).

In what follows, we first propose a modified optimal first-order method for solving a strongly-convex-strongly-concave minimax problem in Subsection 2.1. Using this method as a subproblem solver for an inexact proximal point scheme, we then propose a first-order method for (6) in Subsection 2.2, which enjoys an operation complexity of $\mathcal{O}(\epsilon^{-5/2}\log\epsilon^{-1})$, measured by the amount of evaluations of ∇h and proximal operator of p and q, for finding an ϵ-stationary point of (6).

2.1 A modified optimal first-order method for strongly-convex-strongly-concave minimax problem

In this subsection, we consider the strongly-convex-strongly-concave minimax problem

$$\bar{H}^* = \min_x \max_y \left\{ \bar{H}(x,y) := \bar{h}(x,y) + p(x) - q(y) \right\}, \tag{8}$$

where p and q satisfy Assumption 1 and $\bar{h}$ satisfies the following assumption.

Assumption 2. *$\bar{h}(x,y)$ 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 $\sigma_x, \sigma_y > 0$.*

The goal of this subsection is to propose a modified optimal first-order method for finding an approximate stationary point of problem (8) and study its complexity. Before proceeding, we introduce some more notation below, most of which is adopted from [29].

Let (x^*, y^*) denote the optimal solution of (8), $z^* = -\sigma_x x^*$, and

$$D_p = \max\{\|u-v\| \,|\, u,v \in \operatorname{dom} p\}, \quad D_q = \max\{\|u-v\| \,|\, u,v \in \operatorname{dom} q\}, \tag{9}$$

$$\bar{H}_{\rm low} = \min\left\{\bar{H}(x,y) | \big(x,y) \in \operatorname{dom} p \times \operatorname{dom} q\right\}, \tag{10}$$

$$\hat{h}(x,y) = \bar{h}(x,y) - \sigma_x\|x\|^2/2 + \sigma_y\|y\|^2/2, \tag{11}$$

$$\mathcal{G}(z,y) = \sup_x\{\langle x,z\rangle - p(x) - \hat{h}(x,y) + q(y)\}, \tag{12}$$

$$\mathcal{P}(z,y) = \sigma_x^{-1}\|z\|^2/2 + \sigma_y\|y\|^2/2 + \mathcal{G}(z,y), \tag{13}$$

$$\vartheta_k = \eta_z^{-1}\|z^k - z^*\|^2 + \eta_y^{-1}\|y^k - y^*\|^2 + 2\bar{\alpha}^{-1}(\mathcal{P}(z_f^k, y_f^k) - \mathcal{P}(z^*, y^*)), \tag{14}$$

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

where $\bar{\alpha} = \min\left\{1, \sqrt{8\sigma_y/\sigma_x}\right\}$, $\eta_z = \sigma_x/2$, $\eta_y = \min\{1/(2\sigma_y), 4/(\bar{\alpha}\sigma_x)\}$, and y^k, y_f^k, y_g^k, z^k, z_f^k and z_g^k are generated at iteration k of Algorithm 1 below. By Assumptions 1 and 2, one can observe that D_p, D_q and $\bar{H}_{\rm low}$ are finite.

We are now ready to present a modified optimal first-order method for solving (8) in Algorithm 1. It is a slight modification of the novel optimal first-order method [29, Algorithm 4] by incorporating a forward-backward splitting scheme and also a verifiable termination criterion (see steps 23-25 in Algorithm 1) in order to find a τ-stationary point of (8) (see Definition 2) for any prescribed tolerance $\tau > 0$.

Algorithm 1 A modified optimal first-order method for (8)

```
Input: $\tau > 0$, $\bar{z}^0 = z_f^0 \in -\sigma_x \mathrm{dom}\, p$,[3] $\bar{y}^0 = y_f^0 \in \mathrm{dom}\, q$, $(z^0, y^0) = (\bar{z}^0, \bar{y}^0)$, $\bar{\alpha} = \min\left\{1, \sqrt{8\sigma_y/\sigma_x}\right\}$,
    $\eta_z = \sigma_x/2$, $\eta_y = \min\{1/(2\sigma_y), 4/(\bar{\alpha}\sigma_x)\}$, $\beta_t = 2/(t+3)$, $\zeta = \left(2\sqrt{5}(1 + 8L_{\nabla \bar{h}}/\sigma_x)\right)^{-1}$, $\gamma_x = \gamma_y = 8\sigma_x^{-1}$, and $\hat{\zeta} = \min\{\sigma_x, \sigma_y\}/L^2_{\nabla \bar{h}}$.
for $k = 0, 1, 2, \ldots$ do
    $(z_g^k, y_g^k) = \bar{\alpha}(z^k, y^k) + (1 - \bar{\alpha})(z_f^k, y_f^k)$.
    $(x^{k,-1}, y^{k,-1}) = (-\sigma_x^{-1} z_g^k, y_g^k)$.
    $x^{k,0} = \mathrm{prox}_{\zeta\gamma_x p}(x^{k,-1} - \zeta\gamma_x a_x^k(x^{k,-1}, y^{k,-1}))$.
    $y^{k,0} = \mathrm{prox}_{\zeta\gamma_y q}(y^{k,-1} - \zeta\gamma_y a_y^k(x^{k,-1}, y^{k,-1}))$.
    $b_x^{k,0} = \frac{1}{\zeta\gamma_x}(x^{k,-1} - \zeta\gamma_x a_x^k(x^{k,-1}, y^{k,-1}) - x^{k,0})$.
    $b_y^{k,0} = \frac{1}{\zeta\gamma_y}(y^{k,-1} - \zeta\gamma_y a_y^k(x^{k,-1}, y^{k,-1}) - y^{k,0})$.
    $t = 0$.
    while
    $\gamma_x\|a_x^k(x^{k,t}, y^{k,t}) + b_x^{k,t}\|^2 + \gamma_y\|a_y^k(x^{k,t}, y^{k,t}) + b_y^{k,t}\|^2 > \gamma_x^{-1}\|x^{k,t} - x^{k,-1}\|^2 + \gamma_y^{-1}\|y^{k,t} - y^{k,-1}\|^2$
      do
        $x^{k,t+1/2} = x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x(a_x^k(x^{k,t}, y^{k,t}) + b_x^{k,t})$.
        $y^{k,t+1/2} = y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y(a_y^k(x^{k,t}, y^{k,t}) + b_y^{k,t})$.
        $x^{k,t+1} = \mathrm{prox}_{\zeta\gamma_x p}(x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x a_x^k(x^{k,t+1/2}, y^{k,t+1/2}))$.
        $y^{k,t+1} = \mathrm{prox}_{\zeta\gamma_y q}(y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y a_y^k(x^{k,t+1/2}, y^{k,t+1/2}))$.
        $b_x^{k,t+1} = \frac{1}{\zeta\gamma_x}(x^{k,t} + \beta_t(x^{k,0} - x^{k,t}) - \zeta\gamma_x a_x^k(x^{k,t+1/2}, y^{k,t+1/2}) - x^{k,t+1})$.
        $b_y^{k,t+1} = \frac{1}{\zeta\gamma_y}(y^{k,t} + \beta_t(y^{k,0} - y^{k,t}) - \zeta\gamma_y a_y^k(x^{k,t+1/2}, y^{k,t+1/2}) - y^{k,t+1})$.
        $t \leftarrow t + 1$.
    end while
    $(x_f^{k+1}, y_f^{k+1}) = (x^{k,t}, y^{k,t})$.
    $(z_f^{k+1}, w_f^{k+1}) = (\nabla_x \hat{h}(x_f^{k+1}, y_f^{k+1}) + b_x^{k,t}, -\nabla_y \hat{h}(x_f^{k+1}, y_f^{k+1}) + b_y^{k,t})$.
    $z^{k+1} = z^k + \eta_z\sigma_x^{-1}(z_f^{k+1} - z^k) - \eta_z(x_f^{k+1} + \sigma_x^{-1} z_f^{k+1})$.
    $y^{k+1} = y^k + \eta_y\sigma_y(y_f^{k+1} - y^k) - \eta_y(w_f^{k+1} + \sigma_y y_f^{k+1})$.
    $x^{k+1} = -\sigma_x^{-1} z^{k+1}$.
    $\hat{x}^{k+1} = \mathrm{prox}_{\hat{\zeta} p}(x^{k+1} - \hat{\zeta}\nabla_x \bar{h}(x^{k+1}, y^{k+1}))$.
    $\hat{y}^{k+1} = \mathrm{prox}_{\hat{\zeta} q}(y^{k+1} + \hat{\zeta}\nabla_y \bar{h}(x^{k+1}, y^{k+1}))$.
    Terminate the algorithm and output $(\hat{x}^{k+1}, \hat{y}^{k+1})$ if
        $$\|\hat{\zeta}^{-1}(x^{k+1} - \hat{x}^{k+1}, \hat{y}^{k+1} - y^{k+1}) - (\nabla \bar{h}(x^{k+1}, y^{k+1}) - \nabla \bar{h}(\hat{x}^{k+1}, \hat{y}^{k+1}))\| \le \tau. \quad (15)$$
end for
```

The following theorem presents *iteration and operation complexity* of Algorithm 1 for finding a τ-stationary point of problem (8), whose proof is deferred to Subsection 5.1.

Theorem 1 (Complexity of Algorithm 1). *Suppose that Assumptions 1 and 2 hold. Let $\bar{H}^*$, D_p, D_q, $\bar{H}_{\mathrm{low}}$, and ϑ_0 be defined in* (8), (9), (10) *and* (14)*, σ_x, σ_y and $L_{\nabla\bar{h}}$ be given in Assumption 2, $\bar{\alpha}$, η_y, η_z, τ, $\hat{\zeta}$ be given in Algorithm 1, and*

$$\bar{\delta} = (2 + \bar{\alpha}^{-1})\sigma_x D_p^2 + \max\{2\sigma_y, \bar{\alpha}\sigma_x/4\} D_q^2, \tag{16}$$

$$\bar{K} = \left\lceil \max\left\{\frac{2}{\bar{\alpha}}, \frac{\bar{\alpha}\sigma_x}{4\sigma_y}\right\} \log \frac{4\max\{\eta_z\sigma_x^{-2}, \eta_y\}\vartheta_0}{(\hat{\zeta}^{-1} + L_{\nabla\bar{h}})^{-2}\tau^2} \right\rceil_+, \tag{17}$$

$$\bar{N} = \left\lceil \max\left\{2, \sqrt{\frac{\sigma_x}{2\sigma_y}}\right\} \log \frac{4\max\{1/(2\sigma_x), \min\{1/(2\sigma_y), 4/(\bar{\alpha}\sigma_x)\}\}\left(\bar{\delta} + 2\bar{\alpha}^{-1}\left(\bar{H}^* - \bar{H}_{\mathrm{low}}\right)\right)}{(L^2_{\nabla\bar{h}}/\min\{\sigma_x, \sigma_y\} + L_{\nabla\bar{h}})^{-2}\tau^2} \right\rceil_+ \times \left(\left\lceil 96\sqrt{2}\left(1 + 8L_{\nabla\bar{h}}\sigma_x^{-1}\right)\right\rceil + 2\right). \tag{18}$$

Then Algorithm 1 outputs a τ-stationary point of (8) *in at most $\bar{K}$ iterations. Moreover, the total*

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

number of evaluations of $\nabla\bar{h}$ and proximal operator of p and q performed in Algorithm 1 is no more than $\bar{N}$, respectively.

Remark 1. *It can be observed from Theorem 1 that Algorithm 1 enjoys an operation complexity of $\mathcal{O}(\log\tau^{-1})$, measured by the amount of evaluations of $\nabla\bar{h}$ and proximal operator of p and q, for finding a τ-stationary point of the strongly-convex-strongly-concave minimax problem* (8)*.*

2.2 A first-order method for problem (6)

In this subsection, we propose a first-order method for finding an ϵ-stationary point of problem (6) (see Definition 2) for any prescribed tolerance $\epsilon > 0$. In particular, we first add a perturbation to the max part of (6) for obtaining an approximation of (6), which is given as follows:

$$\min_x \max_y \left\{ h(x,y) + p(x) - q(y) - \frac{\epsilon}{4D_q}\|y - \hat{y}^0\|^2 \right\} \tag{19}$$

for some $\hat{y}^0 \in \operatorname{dom} q$. We then apply an inexact proximal point method [25] to (19), which consists of approximately solving a sequence of subproblems

$$\min_x \max_y \left\{ H_k(x,y) := h_k(x,y) + p(x) - q(y) \right\}, \tag{20}$$

where

$$h_k(x,y) = h(x,y) - \epsilon\|y - \hat{y}^0\|^2/(4D_q) + L_{\nabla h}\|x - x^k\|^2. \tag{21}$$

By Assumption 1, one can observe that (i) h_k is $L_{\nabla h}$-strongly convex in x and $\epsilon/(2D_q)$-strongly concave in y on $\operatorname{dom} p \times \operatorname{dom} q$; (ii) h_k is $(3L_{\nabla h} + \epsilon/(2D_q))$-smooth on $\operatorname{dom} p \times \operatorname{dom} q$. Consequently, problem (20) is a special case of (8) and we can apply Algorithm 1 to solve (20). The resulting first-order method for (6) is presented in Algorithm 2.

Algorithm 2 A first-order method for problem (6)

```
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 1 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 (21).
3:     Terminate the algorithm and output $(x_\epsilon, y_\epsilon) = (x^{k+1}, y^{k+1})$ if
           $\|x^{k+1} - x^k\| \le \epsilon/(4L_{\nabla h}).$    (22)
4: end for
```

Remark 2. *It can be observed from step 2 of Algorithm 2 that (x^{k+1}, y^{k+1}) results from applying Algorithm 1 to the subproblem* (20)*. As will be shown in Lemma 2, (x^{k+1}, y^{k+1}) is an ϵ_k-stationary point of* (20)*.*

We next study complexity of Algorithm 2 for finding an ϵ-stationary point of problem (6). Before proceeding, we define

$$H_{\text{low}} := \min\left\{ H(x,y) | (x,y) \in \operatorname{dom} p \times \operatorname{dom} q \right\}. \tag{23}$$

By Assumption 1, one can observe that H_{low} is finite.

The following theorem presents *iteration and operation complexity* of Algorithm 2 for finding an ϵ-stationary point of problem (6), whose proof is deferred to Subsection 5.2.

Theorem 2 (Complexity of Algorithm 2). *Suppose that Assumption 1 holds. Let H^*, H D_p, D_q, and H_{low} be defined in* (6)*,* (9) *and* (23)*, $L_{\nabla h}$ be given in Assumption 1, ϵ, ϵ_0 and $\hat{x}^0$ be given in*

Algorithm 2, and

$$
\begin{aligned}
\alpha &= \min\left\{1, \sqrt{4\epsilon/(D_q L_{\nabla h})}\right\}, && (24)\\
\delta &= (2+\alpha^{-1})L_{\nabla h}D_p^2 + \max\left\{\epsilon/D_q, \alpha L_{\nabla h}/4\right\}D_q^2, && (25)\\
K &= \left\lceil 16(\max_y H(\hat{x}^0,y) - H^* + \epsilon D_q/4)L_{\nabla h}\epsilon^{-2} + 32\epsilon_0^2(1+4D_q^2 L_{\nabla h}^2\epsilon^{-2})\epsilon^{-2} - 1\right\rceil_+, && (26)\\
N &= \left(\left\lceil 96\sqrt{2}\left(1 + (24L_{\nabla h} + 4\epsilon/D_q)\,L_{\nabla h}^{-1}\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_q L_{\nabla h}\epsilon^{-1}}\right\}\\
&\quad \times \left((K+1)\left(\log \frac{4\max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{D_q}{\epsilon}, \frac{4}{\alpha L_{\nabla h}}\right\}\right\}\left(\delta + 2\alpha^{-1}(H^* - H_{\text{low}} + \epsilon D_q/4 + L_{\nabla h}D_p^2)\right)}{\left[(3L_{\nabla h} + \epsilon/(2D_q))^2/\min\{L_{\nabla h}, \epsilon/(2D_q)\} + 3L_{\nabla h} + \epsilon/(2D_q)\right]^{-2}\epsilon_0^2}\right)_+\right.\\
&\quad \left. + K + 1 + 2K\log(K+1)\right). && (27)
\end{aligned}
$$

Then Algorithm 2 terminates and outputs an ϵ-stationary point (x_ϵ, y_ϵ) of (6) in at most $K+1$ outer iterations that satisfies

$$\max_y H(x_\epsilon, y) \le \max_y H(\hat{x}^0, y) + \epsilon D_q/4 + 2\epsilon_0^2\left(L_{\nabla h}^{-1} + 4D_q^2 L_{\nabla h}\epsilon^{-2}\right). \tag{28}$$

Moreover, the total number of evaluations of ∇h and proximal operator of p and q performed in Algorithm 2 is no more than N, respectively.

Remark 3. *Since $\epsilon_0 \in (0, \epsilon/2]$, one can observe from Theorem 2 that $\alpha = \mathcal{O}(\epsilon^{1/2})$, $\delta = \mathcal{O}(\epsilon^{-1/2})$, $K = \mathcal{O}(\epsilon^{-2})$, and $N = \mathcal{O}(\epsilon^{-5/2}\log(\epsilon_0^{-1}\epsilon^{-1}))$. Consequently, Algorithm 2 enjoys an operation complexity of $\mathcal{O}(\epsilon^{-5/2}\log(\epsilon_0^{-1}\epsilon^{-1}))$, measured by the amount of evaluations of ∇h and proximal operator of p and q, for finding an ϵ-stationary point of the nonconvex-concave minimax problem (6).*

3 Unconstrained bilevel optimization

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

$$
\begin{aligned}
f^* = \min \quad & f(x,y)\\
\text{s.t.} \quad & y \in \operatorname*{Argmin}_z \tilde{f}(x,z).
\end{aligned} \tag{29}
$$

Assume that problem (29) has at least one optimal solution. In addition, f and $\tilde{f}$ satisfy the following assumptions.

Assumption 3. *(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

$$
\begin{aligned}
& D_{\mathbf{x}} := \max\{\|u-v\| \,|\, u,v \in \mathcal{X}\}, \quad D_{\mathbf{y}} := \max\{\|u-v\| \,|\, u,v \in \mathcal{Y}\}, && (30)\\
& \tilde{f}_{\text{hi}} := \max\{\tilde{f}(x,y) | (x,y) \in \mathcal{X}\times\mathcal{Y}\}, \quad \tilde{f}_{\text{low}} := \min\{\tilde{f}(x,y) | (x,y) \in \mathcal{X}\times\mathcal{Y}\}, && (31)\\
& f_{\text{low}} := \min\{f(x,y) | (x,y) \in \mathcal{X}\times\mathcal{Y}\}. && (32)
\end{aligned}
$$

[4] For convenience, problem (29) 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, (29) is essentially a constrained bilevel optimization problem.

By Assumption 3, 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 subsection is to propose penalty methods for solving problem for solving (29). To this end, we observe that problem (29) can be viewed as

$$\min_{x,y}\{f(x,y)|\tilde{f}(x,y) \le \min_z \tilde{f}(x,z)\}. \tag{33}$$

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 (33) is

$$\min_{x,y} f(x,y) + \rho(\tilde{f}(x,y) - \min_z \tilde{f}(x,z)), \tag{34}$$

where $\rho > 0$ is a penalty parameter. We further observe that (34) is equivalent to the *minimax problem*

$$\min_{x,y}\max_z P_\rho(x,y,z), \quad \text{where} \quad P_\rho(x,y,z) := f(x,y) + \rho(\tilde{f}(x,y) - \tilde{f}(x,z)). \tag{35}$$

In view of Assumption 3(i), P_ρ can be rewritten as

$$P_\rho(x,y,z) = \big(f_1(x,y) + \rho\tilde{f}_1(x,y) - \rho\tilde{f}_1(x,z)\big) + \big(f_2(x) + \rho\tilde{f}_2(y) - \rho\tilde{f}_2(z)\big). \tag{36}$$

By this and Assumption 3, 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 the nice structure of P_ρ, an approximate stationary point of the minimax problem (35) can be found by Algorithm 2 proposed in Subsection 2.2.

Based on the above observations, we are now ready to propose penalty methods for the unconstrained bilevel optimization problem (29) by solving either a sequence of or a single minimax problem in the form of (35). In particular, we first propose an *ideal* penalty method for (29) by solving a sequence of minimax problems (see Algorithm 3). Then we propose a *practical* penalty method for (29) by finding an approximate stationary point of a single minimax problem (see Algorithm 4).

Algorithm 3 An ideal penalty method for problem (29)

```
Input: positive sequences $\{\rho_k\}$ and $\{\epsilon_k\}$ with $\lim_{k\to\infty}(\rho_k, \epsilon_k) = (\infty, 0)$.
for $k = 0, 1, 2, \ldots$ do
    Find an $\epsilon_k$-optimal solution $(x^k, y^k, z^k)$ of problem (35) with $\rho = \rho_k$.
end for
```

The following theorem states a convergence result of Algorithm 3, whose proof is deferred to Section 5.3.

Theorem 3 (Convergence of Algorithm 3). *Suppose that Assumption 3 holds 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* (29).

Notice that (35) is a *nonconvex*-concave minimax problem. It is typically hard to find an ϵ-optimal solution of (35) for an arbitrary $\epsilon > 0$. Consequently, Algorithm 3 is *not implementable* in general. We next propose a *practical* penalty method for problem (29) by finding an approximate stationary point of a single minimax problem (35) with a suitable choice of ρ.

Algorithm 4 A practical penalty method for problem (29)

```
Input: $\varepsilon \in (0, 1/4]$, $\rho = \varepsilon^{-1}$, $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$ with $\tilde{f}(x^0, y^0) \le \min_y \tilde{f}(x^0, y) + \varepsilon$.
Call Algorithm 2 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 $\epsilon$-stationary point $(x_\epsilon, y_\epsilon, z_\epsilon)$ of problem (35) with $\rho = \varepsilon^{-1}$.
Output: $(x_\epsilon, y_\epsilon)$.
```

Remark 4. *(i) The initial point (x^0, y^0) of Algorithm 4 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 [38] to the problem $\min_y \tilde{f}(x^0, y)$ for finding $y^0 \in \mathcal{Y}$ such that $\tilde{f}(x^0, y^0) \leq \min_y \tilde{f}(x^0, y) + \varepsilon$; (ii) As seen from Theorem 2, an ϵ-stationary point of* (35) *can be successfully found in step 1 of Algorithm 4 by applying Algorithm 2 to* (35)*; (iii) For the sake of simplicity, a single subproblem of the form* (35) *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* (35) *with dynamic penalty and tolerance parameters instead.*

In order to characterize the approximate solution found by Algorithm 4, we next introduce a terminology called an ε-KKT solution of problem (29).

Recall that problem (29) can be viewed as problem (33). In the spirit of classical constrained optimization, one would naturally be interested in a KKT solution (x, y) of (33) or equivalently (29), namely, (x, y) satisfies $\tilde{f}(x, y) \leq \min_z \tilde{f}(x, z)$ and moreover (x, y) is a stationary point of the problem

$$\min_{x',y'} f(x', y') + \rho\big(\tilde{f}(x', y') - \min_{z'} \tilde{f}(x', z')\big) \tag{37}$$

for some $\rho \geq 0$. Yet, due to the sophisticated problem structure, characterizing a stationary point of (37) is generally difficult. On another hand, notice that problem (37) 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 satisfies

$$0 \in \partial f(x, y) + \rho \partial \tilde{f}(x, y) - (\rho \nabla_x \tilde{f}(x, z); 0), \quad 0 \in \rho \partial_z \tilde{f}(x, z). \tag{38}$$

Based on this observation, we are instead interested in a (weak) KKT solution of problem (29) and its inexact counterpart that are defined below.

Definition 3. *The pair (x, y) is said to be a KKT solution of problem* (29) *if there exists $(z, \rho) \in \mathbb{R}^m \times \mathbb{R}_+$ such that* (38) *and $\tilde{f}(x, y) \leq \min_{z'} \tilde{f}(x, z')$ hold. In addition, for any $\varepsilon > 0$, (x, y) is said to be an ε-KKT solution of problem* (29) *if there exists $(z, \rho) \in \mathbb{R}^m \times \mathbb{R}_+$ such that*

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

We are now ready to present a theorem regarding *operation complexity* of Algorithm 4, 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 (29), whose proof is deferred to Subsection 5.3.

Theorem 4 (Complexity of Algorithm 4). *Suppose that Assumption 3 holds. Let f^*, f, $\tilde{f}$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $\tilde{f}_{\text{hi}}$, $\tilde{f}_{\text{low}}$ and f_{low} be defined in* (29), (30), (31) *and* (32), *$L_{\nabla f_1}$ and $L_{\nabla \tilde{f}_1}$ be given in Assumption 3, ε, ρ, x^0, y^0 and z_ϵ be given in Algorithm 4, and*

$$\begin{aligned}
&\widehat{L} = L_{\nabla f_1} + 2\varepsilon^{-1} L_{\nabla \tilde{f}_1}, \ \hat{\alpha} = \min\left\{1, \sqrt{4\varepsilon/(D_{\mathbf{y}} \widehat{L})}\right\}, \qquad (39)\\
&\hat{\delta} = (2 + \hat{\alpha}^{-1})(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2)\widehat{L} + \max\left\{\varepsilon/D_{\mathbf{y}}, \hat{\alpha}\widehat{L}/4\right\} D_{\mathbf{y}}^2, \\
&\widehat{C} = \frac{4 \max\left\{\frac{1}{2\widehat{L}}, \min\left\{\frac{D_{\mathbf{y}}}{\varepsilon}, \frac{4}{\hat{\alpha}\widehat{L}}\right\}\right\} \left[\hat{\delta} + 2\hat{\alpha}^{-1}(f^* - f_{\text{low}} + \varepsilon^{-1}(\tilde{f}_{\text{hi}} - \tilde{f}_{\text{low}}) + \varepsilon D_{\mathbf{y}}/4 + \widehat{L}(D_{\mathbf{x}}^2 + D_{\mathbf{y}}^2))\right]}{\left[(3\widehat{L} + \varepsilon/(2D_{\mathbf{y}}))^2 / \min\{\widehat{L}, \varepsilon/(2D_{\mathbf{y}})\} + 3\widehat{L} + \varepsilon/(2D_{\mathbf{y}})\right]^{-2} \varepsilon^3}, \\
&\widehat{K} = \left\lceil 16(1 + f(x^0, y^0) - f_{\text{low}} + \varepsilon D_{\mathbf{y}}/4)\widehat{L}\varepsilon^{-2} + 32(1 + 4D_{\mathbf{y}}^2 \widehat{L}^2 \varepsilon^{-2})\varepsilon - 1 \right\rceil_+, \\
&\widehat{N} = \left(\left\lceil 96\sqrt{2}(1 + (24\widehat{L} + 4\varepsilon/D_{\mathbf{y}})\widehat{L}^{-1}) \right\rceil + 2\right) \max\left\{2, \sqrt{D_{\mathbf{y}} \widehat{L} \varepsilon^{-1}}\right\} \\
&\qquad \times ((\widehat{K} + 1)(\log \widehat{C})_+ + \widehat{K} + 1 + 2\widehat{K} \log(\widehat{K} + 1)).
\end{aligned}$$

Then Algorithm 4 outputs an approximate solution (x_ϵ, y_ϵ) of (29) satisfying

$$\mathrm{dist}\left(0, \partial f(x_\epsilon, y_\epsilon) + \rho\partial\tilde{f}(x_\epsilon, y_\epsilon) - (\rho\nabla_x\tilde{f}(x_\epsilon, z_\epsilon); 0)\right) \le \varepsilon, \quad \mathrm{dist}\left(0, \rho\partial\tilde{f}(x_\epsilon, z_\epsilon)\right) \le \varepsilon, \tag{40}$$

$$\tilde{f}(x_\epsilon, y_\epsilon) \le \min_z \tilde{f}(x_\epsilon, z) + \varepsilon\left(1 + f(x^0, y^0) - f_{\mathrm{low}} + 2\varepsilon^3(\widehat{L}^{-1} + 4D_{\mathbf{y}}^2\widehat{L}\varepsilon^{-2}) + D_{\mathbf{y}}\varepsilon/4\right), \tag{41}$$

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 5. *One can observe from Theorem 4 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 $\widehat{N} = \mathcal{O}(\varepsilon^{-4}\log\varepsilon^{-1})$. Consequently, Algorithm 4 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_ϵ, y_ϵ) of (29) satisfying*

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

where z_ϵ is given in Algorithm 4 and $\rho = \varepsilon^{-1}$.

4 Constrained bilevel optimization

In this section, we consider a constrained bilevel optimization problem[5]

$$\begin{aligned} f^* = \min \quad & f(x,y) \\ \text{s.t.} \quad & y \in \operatorname*{Argmin}_z\{\tilde{f}(x,z) | \tilde{g}(x,z) \le 0\}, \end{aligned} \tag{42}$$

where f and $\tilde{f}$ satisfy Assumption 3. Recall from Assumption 3 that $\mathcal{X} = \mathrm{dom}\, f_2$ and $\mathcal{Y} = \mathrm{dom}\,\tilde{f}_2$. We now make some additional assumptions for problem (42).

Assumption 4. *(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 $G := \min\{-\tilde{g}_i(x, \hat{z}_x) | x \in \mathcal{X},\ i = 1, \ldots, l\} > 0$.[6]

For notational convenience, we define

$$\tilde{f}^*(x) := \min_z\{\tilde{f}(x,z) | \tilde{g}(x,z) \le 0\}, \tag{43}$$

$$\tilde{f}^*_{\mathrm{hi}} := \sup\{\tilde{f}^*(x) | x \in \mathcal{X}\}, \tag{44}$$

$$\tilde{g}_{\mathrm{hi}} := \max\{\|\tilde{g}(x,y)\| \,|\, (x,y) \in \mathcal{X} \times \mathcal{Y}\}, \tag{45}$$

It then follows from Assumption 4(ii) that

$$\|\nabla\tilde{g}(x,y)\| \le L_{\tilde{g}} \qquad \forall (x,y) \in \mathcal{X} \times \mathcal{Y}. \tag{46}$$

In addition, by Assumptions 3 and 4 and the compactness of $\mathcal{X}$ and $\mathcal{Y}$, one can observe that $\tilde{g}_{\mathrm{hi}}$ and G are finite. Besides, as will be shown in Lemma 6(ii), $\tilde{f}^*_{\mathrm{hi}}$ is finite.

The goal of this subsection is to propose penalty methods for solving problem (42). To this end, let us 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 (42), which is given by

$$\widetilde{P}_\mu(x,z) = \tilde{f}(x,z) + \mu\,\|[\tilde{g}(x,z)]_+\|^2 \tag{47}$$

[5] For convenience, problem (42) is referred to as a constrained bilevel optimization problem since its lower level part has at least one explicit constraint.

[6] 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 (42), 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 4(iii) clearly holds. Otherwise, one can solve the perturbed counterpart of (42) with $\tilde{g}(x,z)$ being replaced by $\tilde{g}(x,z) - \epsilon$ for some suitable $\epsilon > 0$ instead, which satisfies Assumption 4(iii).

for a penalty parameter $\mu > 0$. Observe that problem (42) can be approximately solved as the *unconstrained bilevel optimization* problem

$$f_\mu^* = \min_{x,y} \big\{ f(x,y) | y \in \operatorname*{Argmin}_z \widetilde{P}_\mu(x,z) \big\}. \tag{48}$$

Further, by the study in Section 3, problem (48) can be approximately solved as the *penalty problem*

$$\min_{x,y} f(x,y) + \rho\Big(\widetilde{P}_\mu(x,y) - \min_z \widetilde{P}_\mu(x,z)\Big) \tag{49}$$

for some suitable $\rho > 0$. One can also observe that problem (49) is equivalent to the *minimax problem*

$$\min_{x,y} \max_z P_{\rho,\mu}(x,y,z), \quad \text{where} \quad P_{\rho,\mu}(x,y,z) := f(x,y) + \rho(\widetilde{P}_\mu(x,y) - \widetilde{P}_\mu(x,z)). \tag{50}$$

In view of (47), (50) and Assumption 3(i), $P_{\rho,\mu}$ can be rewritten as

$$\begin{aligned} P_{\rho,\mu}(x,y,z) = {} & \big(f_1(x,y) + \rho\tilde{f}_1(x,y) + \rho\mu \left\|[\tilde{g}(x,y)]_+\right\|^2 - \rho\tilde{f}_1(x,z) - \rho\mu \left\|[\tilde{g}(x,z)]_+\right\|^2 \big) \\ & + \big(f_2(x) + \rho\tilde{f}_2(y) - \rho\tilde{f}_2(z)\big). \end{aligned} \tag{51}$$

By this and Assumptions 3 and 4, one can observe that $P_{\rho,\mu}$ enjoys the following nice properties.

- $P_{\rho,\mu}$ is the sum of smooth function $f_1(x,y)+\rho\tilde{f}_1(x,y)+\rho\mu \left\|[\tilde{g}(x,y)]_+\right\|^2-\rho\tilde{f}_1(x,z)-\rho\mu \left\|[\tilde{g}(x,z)]_+\right\|^2$ 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 the nice structure of $P_{\rho,\mu}$, an approximate stationary point of the minimax problem (50) can be found by Algorithm 2 proposed in Subsection 2.2.

Based on the above observations, we are now ready to propose penalty methods for the constrained bilevel optimization problem (42) by solving a sequence of or a single minimax problem of the form (50). In particular, we first propose an *ideal* penalty method for (42) by solving a sequence of minimax problems (see Algorithm 5). Then we propose a *practical* penalty method for (42) by finding an approximate stationary point of a single minimax problem (see Algorithm 6).

Algorithm 5 An ideal penalty method for problem (42)

```
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 $\epsilon_k$-optimal solution $(x^k, y^k, z^k)$ of problem (50) with $(\rho, \mu) = (\rho_k, \mu_k)$.
3: end for
```

To study convergence of Algorithm 5, we make the following error bound assumption on the solution set of the lower level optimization problem of (42). This type of error bounds has been considered in the context of set-value mappings in the literature (e.g., see [14]).

Assumption 5. *There exists a non-decreasing function $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{\theta\downarrow 0} \omega(\theta) = 0$ and $\bar{\theta} > 0$ such that $\operatorname{dist}(z, \mathcal{S}_\theta(x)) \le \omega(\theta)$ for all $x \in \mathcal{X}$, $z \in \mathcal{S}_0(x)$ and $\theta \in [0, \bar{\theta}]$, where*

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

We are now ready to state a convergence result of Algorithm 5, whose proof is deferred to Section 5.4.

Theorem 5 (Convergence of Algorithm 5). *Suppose that Assumptions 3-5 hold and that $\{(x^k, y^k, z^k)\}$ is generated by Algorithm 5. Then any accumulation point of $\{(x^k, y^k)\}$ is an optimal solution of problem (42).*

Notice that (50) is a *nonconvex*-concave minimax problem. It is generally hard to find an ϵ-optimal solution of (50) for an arbitrary $\epsilon > 0$. As a result, Algorithm 5 is generally *not implementable.* We next propose a *practical* penalty method for problem (42) by finding an approximate stationary point of (50) with a suitable choice of ρ and μ.

Algorithm 6 A practical penalty method for problem (42)

```
Input: $\varepsilon \in (0,1/4]$, $\rho = \varepsilon^{-1}$, $\mu = \varepsilon^{-2}$, $(x^0,y^0) \in \mathcal{X} \times \mathcal{Y}$ with $\widetilde{P}_\mu(x^0,y^0) \leq \min_y \widetilde{P}_\mu(x^0,y) + \varepsilon$.
1: Call Algorithm 2 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 $\epsilon$-stationary point $(x_\epsilon, y_\epsilon, z_\epsilon)$ of problem (50) with $\rho = \varepsilon^{-1}$ and $\mu = \varepsilon^{-2}$.
2: Output: $(x_\epsilon, y_\epsilon)$.
```

Remark 6. *(i) The initial point (x^0, y^0) of Algorithm 6 can be found by the similar procedure as described in Remark 4 with $\tilde{f}$ being replaced by $\widetilde{P}_\mu$; (ii) As seen from Theorem 2, an ϵ-stationary point of* (50) *can be successfully found in step 1 of Algorithm 6 by applying Algorithm 2 to* (50)*; (iii) For the sake of simplicity, a single subproblem of the form* (50) *with static penalty and tolerance parameters is solved in Algorithm 6. Nevertheless, Algorithm 6 can be modified into a perhaps practically more efficient algorithm by solving a sequence of subproblems of the form* (50) *with dynamic penalty and tolerance parameters instead.*

In order to characterize the approximate solution found by Algorithm 6, we next introduce a terminology called an ε-KKT solution of problem (42).

By the definition of $\tilde{f}^*$ in (43), problem (42) can be viewed as

$$\min_{x,y} \{f(x,y) | \tilde{f}(x,y) \leq \tilde{f}^*(x),\ \tilde{g}(x,y) \leq 0\}. \tag{52}$$

Its associated Lagrangian function is given by

$$\mathcal{L}(x,y,\rho,\lambda) = f(x,y) + \rho(\tilde{f}(x,y) - \tilde{f}^*(x)) + \langle \lambda, \tilde{g}(x,y) \rangle. \tag{53}$$

In the spirit of classical constrained optimization, one would naturally be interested in a KKT solution (x,y) of (52) or equivalently (42), namely, (x,y) satisfies

$$\tilde{f}(x,y) \leq \tilde{f}^*(x), \quad \tilde{g}(x,y) \leq 0, \quad \rho(\tilde{f}(x,y) - \tilde{f}^*(x)) = 0, \quad \langle \lambda, \tilde{g}(x,y) \rangle = 0, \tag{54}$$

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

$$\min_{x',y'} \mathcal{L}(x',y',\rho,\lambda) \tag{55}$$

for some $\rho \geq 0$ and $\lambda \in \mathbb{R}^l_+$. Yet, due to the sophisticated problem structure, characterizing a stationary point of (55) is generally difficult. On another hand, notice from Lemma 6 and (53) that problem (55) 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 satisfies

$$0 \in \partial f(x,y) + \rho \partial \tilde{f}(x,y) - \rho(\nabla_x \tilde{f}(x,z) + \nabla_x \tilde{g}(x,z)\tilde{\lambda}; 0) + \nabla \tilde{g}(x,y)\lambda, \tag{56}$$

$$0 \in \rho(\partial_z \tilde{f}(x,z) + \nabla_z \tilde{g}(x,z)\tilde{\lambda}), \tag{57}$$

$$\tilde{\lambda} \in \mathbb{R}^l_+, \quad \tilde{g}(x,z) \leq 0, \quad \langle \tilde{\lambda}, \tilde{g}(x,z) \rangle = 0. \tag{58}$$

Based on this observation and also the fact that (54) is equivalent to

$$\tilde{f}(x,y) = \tilde{f}^*(x), \quad \tilde{g}(x,y) \leq 0, \quad \langle \lambda, \tilde{g}(x,y) \rangle = 0, \tag{59}$$

we are instead interested in a (weak) KKT solution of problem (42) and its inexact counterpart that are defined below.

Definition 4. *The pair (x,y) is said to be a KKT solution of problem* (42) *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* (56)*-*(59) *hold. In addition, for any $\varepsilon > 0$, (x,y) is said to be an ε-KKT solution of problem* (42) *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*

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

where $\tilde{f}^*$ is defined in (43)*.*

We are now ready to present an *operation complexity* of Algorithm 6, 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 (42), whose proof is deferred to Subsection 5.4.

Theorem 6 (Complexity of Algorithm 6). *Suppose that Assumptions 3 and 4 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* (29), (30), (31), (32), (43), (44) *and* (45), *$L_{\nabla f_1}$, $L_{\nabla \tilde{f}_1}$, $L_{\tilde{f}}$, $L_{\nabla \tilde{g}}$, $L_{\tilde{g}}$ and G be given in Assumptions 3 and 4, ε, ρ, μ, x^0, y^0 and z_ϵ be given in Algorithm 6, and*

$$\tilde{\lambda} = 2\varepsilon^{-1}[\tilde{g}(x_\epsilon, z_\epsilon)]_+, \quad \hat{\lambda} = 2\varepsilon^{-3}[\tilde{g}(x_\epsilon, y_\epsilon)]_+, \tag{60}$$

$$\widetilde{L} = L_{\nabla f_1} + 2\varepsilon^{-1}L_{\nabla \tilde{f}_1} + 4\varepsilon^{-3}(\tilde{g}_{\rm hi}L_{\nabla\tilde{g}} + L_{\tilde{g}}^2), \tag{61}$$

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

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

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

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

Then Algorithm 6 outputs an approximate solution (x_ϵ, y_ϵ) of (42) *satisfying*

$$\operatorname{dist}\left(\partial f(x_\epsilon, y_\epsilon) + \rho\partial\tilde{f}(x_\epsilon, y_\epsilon) - \rho(\nabla_x\tilde{f}(x_\epsilon, z_\epsilon) + \nabla_x\tilde{g}(x_\epsilon, z_\epsilon)\tilde{\lambda}; 0) + \nabla\tilde{g}(x_\epsilon, y_\epsilon)\hat{\lambda}\right) \le \varepsilon, \tag{62}$$

$$\operatorname{dist}\left(0, \rho(\partial_z\tilde{f}(x_\epsilon, z_\epsilon) + \nabla_z\tilde{g}(x_\epsilon, z_\epsilon)\tilde{\lambda})\right) \le \varepsilon, \tag{63}$$

$$\|[\tilde{g}(x_\epsilon, z_\epsilon)]_+\| \le \varepsilon^2 G^{-1}D_{\mathbf{y}}(\varepsilon^2 + L_{\tilde{f}})/2, \tag{64}$$

$$|\langle\tilde{\lambda}, \tilde{g}(x_\epsilon, z_\epsilon)\rangle| \le \varepsilon^2 G^{-2}D_{\mathbf{y}}^2(\rho^{-1}\epsilon + L_{\tilde{f}})^2/2, \tag{65}$$

$$\begin{aligned}|\tilde{f}(x_\epsilon, y_\epsilon) - \tilde{f}^*(x_\epsilon)| \le \max\Big\{ & \varepsilon\left(1 + f(x^0,y^0) - f_{\rm low} + 2\varepsilon^5(\widetilde{L}^{-1} + 4D_{\mathbf{y}}^2\widetilde{L}\varepsilon^{-2}) + D_{\mathbf{y}}\varepsilon/4\right), \\ & \varepsilon^2 G^{-2}D_{\mathbf{y}}^2 L_{\tilde{f}}(\varepsilon^2 + \varepsilon L_f + L_{\tilde{f}})/2\Big\},\end{aligned} \tag{66}$$

$$\|[\tilde{g}(x_\epsilon, y_\epsilon)]_+\| \le \varepsilon^2 G^{-1}D_{\mathbf{y}}(\varepsilon^2 + \varepsilon L_f + L_{\tilde{f}})/2, \tag{67}$$

$$|\langle\hat{\lambda}, \tilde{g}(x_\epsilon, y_\epsilon)\rangle| \le \varepsilon G^{-2}D_{\mathbf{y}}^2(\varepsilon^2 + \varepsilon L_f + L_{\tilde{f}})^2/2, \tag{68}$$

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 7. *One can observe from Theorem 6 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 $\widetilde{N} = \mathcal{O}(\varepsilon^{-7}\log\varepsilon^{-1})$. Consequently, Algorithm 6 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_ϵ, y_ϵ) of* (42) *satisfying*

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

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

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

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

where $\tilde{f}^*$ is defined in (43), *$\hat{\lambda}, \tilde{\lambda} \in \mathbb{R}^l_+$ are defined in* (60), *z_ϵ is given in Algorithm 6 and $\rho = \varepsilon^{-1}$.*

5 Proof of the main results

In this section we provide a proof of our main results presented in Sections 2, 3 and 4, which are particularly Theorems 1-6.

5.1 Proof of the main results in Subsection 2.1

In this subsection we prove Theorem 1. Before proceeding, we establish a lemma below.

Lemma 1. *Suppose that Assumptions 1 and 2 hold. Let $\bar{H}^*$, $\bar{H}_{\rm low}$, ϑ_0 and $\bar{\delta}$ be defined in* (8), (10), (14) *and* (16), *and $\bar{\alpha}$ be given in Algorithm 1. Then we have*

$$\vartheta_0 \le \bar{\delta} + 2\bar{\alpha}^{-1}\left(\bar{H}^* - \bar{H}_{\rm low}\right). \tag{69}$$

Proof. By (8), (10), (11) and (12), one has

$$\begin{aligned}
\mathcal{G}(\bar{z}^0,\bar{y}^0) &\overset{(12)}{=} \sup_x\left\{\langle x,\bar{z}^0\rangle - p(x) - \hat{h}(x,\bar{y}^0) + q(\bar{y}^0)\right\}\\
&\overset{(11)}{=} \max_{x\in\mathrm{dom}\,p}\left\{\langle x,\bar{z}^0\rangle - p(x) - \bar{h}(x,\bar{y}^0) + \frac{\sigma_x}{2}\|x\|^2 - \frac{\sigma_y}{2}\|\bar{y}^0\|^2 + q(\bar{y}^0)\right\}\\
&\overset{(8)(10)}{\le} \max_{x\in\mathrm{dom}\,p}\left\{\langle x,\bar{z}^0\rangle + \frac{\sigma_x}{2}\|x\|^2\right\} - \frac{\sigma_y}{2}\|\bar{y}^0\|^2 - \bar{H}_{\rm low}\\
&= \max_{x\in\mathrm{dom}\,p}\frac{\sigma_x}{2}\|x+\sigma_x^{-1}\bar{z}^0\|^2 - \frac{\sigma_x^{-1}}{2}\|\bar{z}^0\|^2 - \frac{\sigma_y}{2}\|\bar{y}^0\|^2 - \bar{H}_{\rm low}\\
&\le \frac{\sigma_x D_p^2}{2} - \frac{\sigma_x^{-1}}{2}\|\bar{z}^0\|^2 - \frac{\sigma_y}{2}\|\bar{y}^0\|^2 - \bar{H}_{\rm low},
\end{aligned} \tag{70}$$

where the last inequality follows from (9) and the fact that $z^0 \in -\sigma_x\mathrm{dom}\,p$.

Recall that (x^*,y^*) is the optimal solution of (8) and $z^* = -\sigma_x x^*$. It follows from (8), (11) and (12) that

$$\begin{aligned}
\mathcal{G}(z^*,y^*) &\overset{(12)}{=} \sup_x\left\{\langle x,z^*\rangle - p(x) - \hat{h}(x,y^*) + q(y^*)\right\} \ge \langle x^*,z^*\rangle - p(x^*) - \hat{h}(x^*,y^*) + q(y^*)\\
&\overset{(11)}{=} \langle x^*,z^*\rangle + \frac{\sigma_x}{2}\|x^*\|^2 - \frac{\sigma_y}{2}\|y^*\|^2 - p(x^*) - \bar{h}(x^*,y^*) + q(y^*)\\
&= -\frac{\sigma_x^{-1}}{2}\|z^*\|^2 - \frac{\sigma_y}{2}\|y^*\|^2 - \bar{H}^*,
\end{aligned}$$

where the last equality follows from (8), the definition of (x^*,y^*), and $z^* = -\sigma_x x^*$. This together with (13) and (70) implies that

$$\begin{aligned}
\mathcal{P}(\bar{z}^0,\bar{y}^0) - \mathcal{P}(z^*,y^*) &= \frac{\sigma_x^{-1}}{2}\|\bar{z}^0\|^2 + \frac{\sigma_y}{2}\|\bar{y}^0\|^2 + \mathcal{G}(\bar{z}^0,\bar{y}^0) - \frac{\sigma_x^{-1}}{2}\|z^*\|^2 - \frac{\sigma_y}{2}\|y^*\|^2 - \mathcal{G}(z^*,y^*)\\
&\le \sigma_x D_p^2/2 - \bar{H}_{\rm low} + \bar{H}^*.
\end{aligned}$$

Notice from Algorithm 1 that $z^0 = z_f^0 = \bar{z}^0 \in -\sigma_x\mathrm{dom}\,p$ and $y^0 = y_f^0 = \bar{y}^0 \in \mathrm{dom}\,q$. By these, $z^* = -\sigma_x x^*$, (9), (14), and the above inequality, one has

$$\begin{aligned}
\vartheta_0 &\overset{(14)}{=} \eta_z^{-1}\|\bar{z}^0 - z^*\|^2 + \eta_y^{-1}\|\bar{y}^0 - y^*\|^2 + 2\bar{\alpha}^{-1}(\mathcal{P}(\bar{z}^0,\bar{y}^0) - \mathcal{P}(z^*,y^*))\\
&\le \eta_z^{-1}\sigma_x^2 D_p^2 + \eta_y^{-1}D_q^2 + 2\bar{\alpha}^{-1}\left(\sigma_x D_p^2/2 - \bar{H}_{\rm low} + \bar{H}^*\right)\\
&= \eta_z^{-1}\sigma_x^2 D_p^2 + \bar{\alpha}^{-1}\sigma_x D_p^2 + \eta_y^{-1}D_q^2 + 2\bar{\alpha}^{-1}\left(\bar{H}^* - \bar{H}_{\rm low}\right).
\end{aligned}$$

Hence, the conclusion follows from this, (16), $\eta_z = \sigma_x/2$ and $\eta_y = \min\{1/(2\sigma_y), 4/(\bar{\alpha}\sigma_x)\}$. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. Suppose for contradiction that Algorithm 1 runs for more than $\bar{K}$ outer iterations, where $\bar{K}$ is given in (17). By this and Algorithm 1, one can assert that (15) does not hold for $k = \bar{K} - 1$. On the other hand, by (17) and [29, Theorem 3], one has

$$\|(x^{\bar{K}}, y^{\bar{K}}) - (x^*,y^*)\| \le (\hat{\zeta}^{-1} + L_{\nabla\bar{h}})^{-1}\tau/2, \tag{71}$$

where (x^*,y^*) is the optimal solution of problem (8) and $\hat{\zeta}$ is an input of Algorithm 1. Notice from Algorithm 1 that $(\hat{x}^{\bar{K}}, \hat{y}^{\bar{K}})$ results from the forward-backward splitting (FBS) step applied to the strongly monotone inclusion problem $0 \in (\nabla_x\bar{h}(x,y), -\nabla_y\bar{h}(x,y)) + (\partial p(x), \partial q(y))$ at the point $(x^{\bar{K}}, y^{\bar{K}})$. It then follows from this, $\hat{\zeta} = \min\{\sigma_x, \sigma_y\}/L^2_{\nabla\bar{h}}$ (see Algorithm 1), and the contraction property of FBS [5, Corollary 2.5] that $\|(\hat{x}^{\bar{K}}, \hat{y}^{\bar{K}}) - (x^*, y^*)\| \leq \|(x^{\bar{K}}, y^{\bar{K}}) - (x^*, y^*)\|$. Using this and (71), we have

$$\begin{aligned}
&\|\hat{\zeta}^{-1}(x^{\bar{K}} - \hat{x}^{\bar{K}}, \hat{y}^{\bar{K}} - y^{\bar{K}}) - (\nabla\bar{h}(x^{\bar{K}}, y^{\bar{K}}) - \nabla\bar{h}(\hat{x}^{\bar{K}}, \hat{y}^{\bar{K}}))\| \\
&\leq \ \hat{\zeta}^{-1}\|(x^{\bar{K}}, y^{\bar{K}}) - (\hat{x}^{\bar{K}}, \hat{y}^{\bar{K}})\| + \|\nabla\bar{h}(x^{\bar{K}}, y^{\bar{K}}) - \nabla\bar{h}(\hat{x}^{\bar{K}}, \hat{y}^{\bar{K}})\| \\
&\leq \ (\hat{\zeta}^{-1} + L_{\nabla\bar{h}})\|(x^{\bar{K}}, y^{\bar{K}}) - (\hat{x}^{\bar{K}}, \hat{y}^{\bar{K}})\| \\
&\leq \ (\hat{\zeta}^{-1} + L_{\nabla\bar{h}})(\|(x^{\bar{K}}, y^{\bar{K}}) - (x^*, y^*)\| + \|(\hat{x}^{\bar{K}}, \hat{y}^{\bar{K}}) - (x^*, y^*)\|) \\
&\leq \ 2(\hat{\zeta}^{-1} + L_{\nabla\bar{h}})\|(x^{\bar{K}}, y^{\bar{K}}) - (x^*, y^*)\| \overset{(71)}{\leq} \tau,
\end{aligned}$$

where the second inequality uses the fact that $\bar{h}$ is $L_{\nabla\bar{h}}$-smooth on $\operatorname{dom} p \times \operatorname{dom} q$. It follows that (15) holds for $k = \bar{K} - 1$, which contradicts the above assertion. Hence, Algorithm 1 must terminate in at most $\bar{K}$ outer iterations.

We next show that the output of Algorithm 1 is a τ-stationary point of (8). To this end, suppose that Algorithm 1 terminates at some iteration k at which (15) is satisfied. Then by (4) and the definition of $\hat{x}^{k+1}$ and $\hat{y}^{k+1}$ (see steps 23 and 24 of Algorithm 1), one has

$$\begin{aligned}
0 &\in \hat{\zeta}\partial p(\hat{x}^{k+1}) + \hat{x}^{k+1} - x^{k+1} + \hat{\zeta}\nabla_x\bar{h}(x^{k+1}, y^{k+1}), \\
0 &\in \hat{\zeta}\partial q(\hat{y}^{k+1}) + \hat{y}^{k+1} - y^{k+1} - \hat{\zeta}\nabla_y\bar{h}(x^{k+1}, y^{k+1}),
\end{aligned}$$

which yield

$$\hat{\zeta}^{-1}(x^{k+1} - \hat{x}^{k+1}) - \nabla_x\bar{h}(x^{k+1}, y^{k+1}) \in \partial p(\hat{x}^{k+1}), \ \hat{\zeta}^{-1}(y^{k+1} - \hat{y}^{k+1}) + \nabla_y\bar{h}(x^{k+1}, y^{k+1}) \in \partial q(\hat{y}^{k+1}).$$

These together with the definition of $\bar{H}$ in (8) imply that

$$\begin{aligned}
\nabla_x\bar{h}(\hat{x}^{k+1}, \hat{y}^{k+1}) + \hat{\zeta}^{-1}(x^{k+1} - \hat{x}^{k+1}) - \nabla_x\bar{h}(x^{k+1}, y^{k+1}) &\in \partial_x\bar{H}(\hat{x}^{k+1}, \hat{y}^{k+1}), \\
\nabla_y\bar{h}(\hat{x}^{k+1}, \hat{y}^{k+1}) - \hat{\zeta}^{-1}(y^{k+1} - \hat{y}^{k+1}) - \nabla_y\bar{h}(x^{k+1}, y^{k+1}) &\in \partial_y\bar{H}(\hat{x}^{k+1}, \hat{y}^{k+1}).
\end{aligned}$$

Using these and (15), we obtain

$$\begin{aligned}
&\operatorname{dist}(0, \partial_x\bar{H}(\hat{x}^{k+1}, \hat{y}^{k+1}))^2 + \operatorname{dist}(0, \partial_y\bar{H}(\hat{x}^{k+1}, \hat{y}^{k+1}))^2 \\
&\leq \|\hat{\zeta}^{-1}(x^{k+1} - \hat{x}^{k+1}) + \nabla_x\bar{h}(\hat{x}^{k+1}, \hat{y}^{k+1}) - \nabla_x\bar{h}(x^{k+1}, y^{k+1})\|^2 \\
&\quad + \|\hat{\zeta}^{-1}(\hat{y}^{k+1} - y^{k+1}) + \nabla_y\bar{h}(\hat{x}^{k+1}, \hat{y}^{k+1}) - \nabla_y\bar{h}(x^{k+1}, y^{k+1})\|^2 \\
&= \|\hat{\zeta}^{-1}(x^{k+1} - \hat{x}^{k+1}, \hat{y}^{k+1} - y^{k+1}) - (\nabla\bar{h}(x^{k+1}, y^{k+1}) - \nabla\bar{h}(\hat{x}^{k+1}, \hat{y}^{k+1}))\|^2 \overset{(15)}{\leq} \tau^2,
\end{aligned}$$

which implies that $\operatorname{dist}(0, \partial_x\bar{H}(\hat{x}^{k+1}, \hat{y}^{k+1})) \leq \tau$ and $\operatorname{dist}(0, \partial_y\bar{H}(\hat{x}^{k+1}, \hat{y}^{k+1})) \leq \tau$. It then follows from these and Definition 2 that the output $(\hat{x}^{k+1}, \hat{y}^{k+1})$ of Algorithm 1 is a τ-stationary point of (8).

Finally, we show that the total number of evaluations of $\nabla\bar{h}$ and proximal operator of p and q performed in Algorithm 1 is no more than $\bar{N}$, respectively. Indeed, notice from Algorithm 1 that $\bar{\alpha} = \min\left\{1, \sqrt{8\sigma_y/\sigma_x}\right\}$, which implies that $2/\bar{\alpha} = \max\{2, \sqrt{\sigma_x/(2\sigma_y)}\}$ and $\bar{\alpha} \leq \sqrt{8\sigma_y/\sigma_x}$. By these, one has

$$\max\left\{\frac{2}{\bar{\alpha}}, \frac{\bar{\alpha}\sigma_x}{4\sigma_y}\right\} \leq \max\left\{2, \sqrt{\frac{\sigma_x}{2\sigma_y}}, \sqrt{\frac{8\sigma_y}{\sigma_x}}\frac{\sigma_x}{4\sigma_y}\right\} = \max\left\{2, \sqrt{\frac{\sigma_x}{2\sigma_y}}\right\}. \tag{72}$$

In addition, by [29, Lemma 4], the number of inner iterations performed in each outer iteration of Algorithm 1 is at most

$$T = \left\lceil 48\sqrt{2}\left(1 + 8L_{\nabla\bar{h}}\sigma_x^{-1}\right)\right\rceil - 1.$$

Then one can observe that the number of evaluations of $\nabla\bar{h}$ and proximal operator of p and q performed in Algorithm 1 is at most

$$\begin{aligned}
(2T+3)\bar{K} &\leq \left(\left\lceil 96\sqrt{2}\left(1+8L_{\nabla\bar{h}}\sigma_x^{-1}\right)\right\rceil+2\right)\left\lceil \max\left\{\frac{2}{\bar{\alpha}},\frac{\bar{\alpha}\sigma_x}{4\sigma_y}\right\}\log\frac{4\max\{\eta_z\sigma_x^{-2},\eta_y\}\vartheta_0}{(\hat{\zeta}^{-1}+L_{\nabla\bar{h}})^{-2}\tau^2}\right\rceil_+ \\
&\overset{(72)}{\leq} \left(\left\lceil 96\sqrt{2}\left(1+8L_{\nabla\bar{h}}\sigma_x^{-1}\right)\right\rceil+2\right)\left\lceil \max\left\{2,\sqrt{\frac{\sigma_x}{2\sigma_y}}\right\}\log\frac{4\max\{\eta_z\sigma_x^{-2},\eta_y\}\vartheta_0}{(\hat{\zeta}^{-1}+L_{\nabla\bar{h}})^{-2}\tau^2}\right\rceil_+ \\
&\leq \left(\left\lceil 96\sqrt{2}\left(1+8L_{\nabla\bar{h}}\sigma_x^{-1}\right)\right\rceil+2\right) \\
&\quad\times\left\lceil \max\left\{2,\sqrt{\frac{\sigma_x}{2\sigma_y}}\right\}\log\frac{4\max\{1/(2\sigma_x),\min\{1/(2\sigma_y),4/(\bar{\alpha}\sigma_x)\}\}\,\vartheta_0}{(L_{\nabla\bar{h}}^2/\min\{\sigma_x,\sigma_y\}+L_{\nabla\bar{h}})^{-2}\tau^2}\right\rceil_+ \overset{(69)(18)}{\leq} \bar{N},
\end{aligned}$$

where the second last inequality follows from the definition of η_y, η_z and $\hat{\zeta}$ in Algorithm 1. Hence, the conclusion holds as desired. □

5.2 Proof of the main results in Subsection 2.2

In this subsection we prove Theorem 2. Before proceeding, let $\{(x^k,y^k)\}_{k\in\mathbb{K}}$ denote all the iterates generated by Algorithm 2, where $\mathbb{K}$ is a subset of consecutive nonnegative integers starting from 0. Also, we define $\mathbb{K}-1=\{k-1:k\in\mathbb{K}\}$. We first establish two lemmas and then use them to prove Theorem 2 subsequently.

Lemma 2. *Suppose that Assumption 1 holds. Let $\{(x^k,y^k)\}_{k\in\mathbb{K}}$ be generated by Algorithm 2, H^*, D_p, D_q, $H_{\rm low}$, α, δ be defined in* (6), (9), (23), (24) *and* (25), *$L_{\nabla h}$ be given in Assumption 1, ϵ, ϵ_k be given in Algorithm 2, and*

$$\begin{aligned}
N_k = &\left(\left\lceil 96\sqrt{2}\left(1+(24L_{\nabla h}+4\epsilon/D_q)\,L_{\nabla h}^{-1}\right)\right\rceil+2\right)\times\left\lceil \max\left\{2,\sqrt{\frac{D_qL_{\nabla h}}{\epsilon}}\right\}\right. \\
&\left.\times\log\frac{4\max\left\{\frac{1}{2L_{\nabla h}},\min\left\{\frac{D_q}{\epsilon},\frac{4}{\alpha L_{\nabla h}}\right\}\right\}\left(\delta+2\alpha^{-1}(H^*-H_{\rm low}+\epsilon D_q/4+L_{\nabla h}D_p^2)\right)}{[(3L_{\nabla h}+\epsilon/(2D_q))^2/\min\{L_{\nabla h},\epsilon/(2D_q)\}+3L_{\nabla h}+\epsilon/(2D_q)]^{-2}\,\epsilon_k^2}\right\rceil_+. \qquad (73)
\end{aligned}$$

Then for all $0\leq k\in\mathbb{K}-1$, (x^{k+1},y^{k+1}) is an ϵ_k-stationary point of (20). *Moreover, the total number of evaluations of ∇h and proximal operator of p and q performed at iteration k of Algorithm 2 for generating (x^{k+1},y^{k+1}) is no more than N_k, respectively.*

Proof. Let (x^*,y^*) be an optimal solution of (6). Recall that H, H_k and h_k are given in (6), (20) and (21), respectively. Then we have

$$\begin{aligned}
H_{k,*} := \min_x\max_y H_k(x,y) &= \min_x\max_y\left\{H(x,y)-\frac{\epsilon}{4D_q}\|y-\hat{y}^0\|^2+L_{\nabla h}\|x-x^k\|^2\right\} \\
&\leq \max_y\{H(x^*,y)+L_{\nabla h}\|x^*-x^k\|^2\} \overset{(6)(9)}{\leq} H^*+L_{\nabla h}D_p^2. \qquad (74)
\end{aligned}$$

Moreover, by (9) and (23), one has

$$\begin{aligned}
H_{k,\rm low} := \min_{(x,y)\in\operatorname{dom}p\times\operatorname{dom}q} H_k(x,y) &= \min_{(x,y)\in\operatorname{dom}p\times\operatorname{dom}q}\left\{H(x,y)-\frac{\epsilon}{4D_q}\|y-\hat{y}^0\|^2+L_{\nabla h}\|x-x^k\|^2\right\} \\
&\overset{(23)}{\geq} H_{\rm low}-\max_{y\in\operatorname{dom}q}\frac{\epsilon}{4D_q}\|y-\hat{y}^0\|^2 \overset{(9)}{\geq} H_{\rm low}-\epsilon D_q/4. \qquad (75)
\end{aligned}$$

In addition, by Assumption 1 and the definition of h_k in (21), it is not hard to verify that $h_k(x,y)$ is $L_{\nabla h}$-strongly-convex in x, $\epsilon/(2D_q)$-strongly-concave in y, and $(3L_{\nabla h}+\epsilon/(2D_q))$-smooth on its domain. Also, recall from Remark 2 that (x^{k+1},y^{k+1}) results from applying Algorithm 1 to problem (20). The conclusion of this lemma then follows by using (74) and (75) and applying Theorem 1 to (20) with $\tau=\epsilon_k$, $\sigma_x=L_{\nabla h}$, $\sigma_y=\epsilon/(2D_q)$, $L_{\nabla\bar{h}}=3L_{\nabla h}+\epsilon/(2D_q)$, $\bar{\alpha}=\alpha$, $\bar{\delta}=\delta$, $\bar{H}_{\rm low}=H_{k,\rm low}$, and $\bar{H}^*=H_{k,*}$. □

Lemma 3. *Suppose that Assumption 1 holds. Let $\{x^k\}_{k\in\mathbb{K}}$ be generated by Algorithm 2, H, H^* and D_q be defined in* (6) *and* (9)*, $L_{\nabla h}$ be given in Assumption 1, and ϵ, ϵ_0 and $\hat{x}^0$ be given in Algorithm 2. Then for all $0 \le K \in \mathbb{K}-1$, we have*

$$\min_{0\le k\le K} \|x^{k+1}-x^k\| \le \frac{\max_y H(\hat{x}^0,y) - H^* + \epsilon D_q/4}{L_{\nabla h}(K+1)} + \frac{2\epsilon_0^2(1+4D_q^2L_{\nabla h}^2\epsilon^{-2})}{L_{\nabla h}^2(K+1)}, \tag{76}$$

$$\max_y H(x^{K+1},y) \le \max_y H(\hat{x}^0,y) + \epsilon D_q/4 + 2\epsilon_0^2\left(L_{\nabla h}^{-1} + 4D_q^2 L_{\nabla h}\epsilon^{-2}\right). \tag{77}$$

Proof. For convenience of the proof, let

$$H_\epsilon^*(x) = \max_y \left\{H(x,y) - \epsilon\|y-\hat{y}^0\|^2/(4D_q)\right\}, \tag{78}$$

$$H_k^*(x) = \max_y H_k(x,y), \quad y_*^{k+1} = \arg\max_y H_k(x^{k+1},y). \tag{79}$$

One can observe from these, (20) and (21) that

$$H_k^*(x) = H_\epsilon^*(x) + L_{\nabla h}\|x-x^k\|^2. \tag{80}$$

By this and Assumption 1, one can also see that H_k^* is $L_{\nabla h}$-strongly convex on $\operatorname{dom} p$. In addition, recall from Lemma 2 that (x^{k+1},y^{k+1}) is an ϵ_k-stationary point of problem (20) for all $0\le k\in\mathbb{K}-1$. It then follows from Definition 2 that there exist some $u\in\partial_x H_k(x^{k+1},y^{k+1})$ and $v\in\partial_y H_k(x^{k+1},y^{k+1})$ with $\|u\|\le\epsilon_k$ and $\|v\|\le\epsilon_k$. Also, by (79), one has $0\in\partial_y H_k(x^{k+1},y_*^{k+1})$, which together with $v\in\partial_y H_k(x^{k+1},y^{k+1})$ and $\epsilon/(2D_q)$-strong concavity of $H_k(x^{k+1},\cdot)$, implies that $\langle -v, y^{k+1}-y_*^{k+1}\rangle \ge \epsilon\|y^{k+1}-y_*^{k+1}\|^2/(2D_q)$. This and $\|v\|\le\epsilon_k$ yield

$$\|y^{k+1}-y_*^{k+1}\| \le 2\epsilon_k D_q/\epsilon. \tag{81}$$

In addition, by $u\in\partial_x H_k(x^{k+1},y^{k+1})$, (20) and (21), one has

$$u \in \nabla_x h(x^{k+1},y^{k+1}) + \partial p(x^{k+1}) + 2L_{\nabla h}(x^{k+1}-x^k). \tag{82}$$

Also, observe from (20), (21) and (79) that

$$\partial H_k^*(x^{k+1}) = \nabla_x h(x^{k+1},y_*^{k+1}) + \partial p(x^{k+1}) + 2L_{\nabla h}(x^{k+1}-x^k),$$

which together with (82) yields

$$u + \nabla_x h(x^{k+1},y_*^{k+1}) - \nabla_x h(x^{k+1},y^{k+1}) \in \partial H_k^*(x^{k+1}).$$

By this and $L_{\nabla h}$-strong convexity of H_k^*, one has

$$H_k^*(x^k) \ge H_k^*(x^{k+1}) + \langle u + \nabla_x h(x^{k+1},y_*^{k+1}) - \nabla_x h(x^{k+1},y^{k+1}), x^k - x^{k+1}\rangle + L_{\nabla h}\|x^k-x^{k+1}\|^2/2. \tag{83}$$

Using this, (80), (81), (83), $\|u\|\le\epsilon_k$, and the Lipschitz continuity of ∇h, we obtain

$$\begin{aligned}
&H_\epsilon^*(x^k) - H_\epsilon^*(x^{k+1}) \overset{(80)}{=} H_k^*(x^k) - H_k^*(x^{k+1}) + L_{\nabla h}\|x^k-x^{k+1}\|^2\\
&\overset{(83)}{\ge} \langle u + \nabla_x h(x^{k+1},y_*^{k+1}) - \nabla_x h(x^{k+1},y^{k+1}), x^k-x^{k+1}\rangle + 3L_{\nabla h}\|x^k-x^{k+1}\|^2/2\\
&\ge \big(-\|u+\nabla_x h(x^{k+1},y_*^{k+1}) - \nabla_x h(x^{k+1},y^{k+1})\|\|x^k-x^{k+1}\| + L_{\nabla h}\|x^k-x^{k+1}\|^2/2\big) + L_{\nabla h}\|x^k-x^{k+1}\|^2\\
&\ge -(2L_{\nabla h})^{-1}\|u+\nabla_x h(x^{k+1},y_*^{k+1}) - \nabla_x h(x^{k+1},y^{k+1})\|^2 + L_{\nabla h}\|x^k-x^{k+1}\|^2\\
&\ge -L_{\nabla h}^{-1}\|u\|^2 - L_{\nabla h}^{-1}\|\nabla_x h(x^{k+1},y_*^{k+1}) - \nabla_x h(x^{k+1},y^{k+1})\|^2 + L_{\nabla h}\|x^k-x^{k+1}\|^2\\
&\ge -L_{\nabla h}^{-1}\epsilon_k^2 - L_{\nabla h}\|y^{k+1}-y_*^{k+1}\|^2 + L_{\nabla h}\|x^k-x^{k+1}\|^2\\
&\overset{(81)}{\ge} -(L_{\nabla h}^{-1} + 4D_q^2L_{\nabla h}\epsilon^{-2})\epsilon_k^2 + L_{\nabla h}\|x^k-x^{k+1}\|^2,
\end{aligned}$$

where the second and fourth inequalities follow from Cauchy-Schwartz inequality, and the third inequality is due to Young's inequality, and the fifth inequality follows from $L_{\nabla h}$-Lipschitz continuity of ∇h. Summing up the above inequality for $k=0,1,\ldots,K$ yields

$$L_{\nabla h}\sum_{k=0}^K \|x^k-x^{k+1}\|^2 \le H_\epsilon^*(x^0) - H_\epsilon^*(x^{K+1}) + (L_{\nabla h}^{-1} + 4D_q^2L_{\nabla h}\epsilon^{-2})\sum_{k=0}^K \epsilon_k^2. \tag{84}$$

In addition, it follows from (6), (9) and (78) that

$$H_\epsilon^*(x^{K+1}) = \max_y \left\{H(x^{K+1}, y) - \epsilon\|y - \hat{y}^0\|^2/(4D_q)\right\} \ge \min_x \max_y H(x, y) - \epsilon D_q/4 = H^* - \epsilon D_q/4,$$
$$H_\epsilon^*(x^0) = \max_y \left\{H(x^0, y) - \epsilon\|y - \hat{y}^0\|^2/(4D_q)\right\} \le \max_y H(x^0, y). \tag{85}$$

These together with (84) yield

$$\begin{aligned} L_{\nabla h}(K+1) \min_{0\le k\le K} \|x^{k+1} - x^k\|^2 &\le L_{\nabla h} \sum_{k=0}^{K} \|x^k - x^{k+1}\|^2 \\ &\le \max_y H(x^0, y) - H^* + \epsilon D_q/4 + (L_{\nabla h}^{-1} + 4D_q^2 L_{\nabla h}\epsilon^{-2}) \sum_{k=0}^{K} \epsilon_k^2, \end{aligned}$$

which together with $x^0 = \hat{x}^0$, $\epsilon_k = \epsilon_0(k+1)^{-1}$ and $\sum_{k=0}^{K}(k+1)^{-2} < 2$ implies that (76) holds.

Finally, we show that (77) holds. Indeed, it follows from (9), (78), (84), (85), $\epsilon_k = \epsilon_0(k+1)^{-1}$, and $\sum_{k=0}^{K}(k+1)^{-2} < 2$ that

$$\begin{aligned} \max_y H(x^{K+1}, y) &\overset{(9)}{\le} \max_y \left\{H(x^{K+1}, y) - \epsilon\|y - \hat{y}^0\|^2/(4D_q)\right\} + \epsilon D_q/4 \overset{(78)}{=} H_\epsilon^*(x^{K+1}) + \epsilon D_q/4 \\ &\overset{(84)}{\le} H_\epsilon^*(x^0) + \epsilon D_q/4 + (L_{\nabla h}^{-1} + 4D_q^2 L_{\nabla h}\epsilon^{-2}) \sum_{k=0}^{K} \epsilon_k^2 \\ &\overset{(85)}{\le} \max_y H(x^0, y) + \epsilon D_q/4 + 2\epsilon_0^2(L_{\nabla h}^{-1} + 4D_q^2 L_{\nabla h}\epsilon^{-2}). \end{aligned}$$

It then follows from this and $x^0 = \hat{x}^0$ that (77) holds. □

We are now ready to prove Theorem 2.

Proof of Theorem 2. Suppose for contradiction that Algorithm 2 runs for more than $K+1$ outer iterations, where K is given in (26). By this and Algorithm 2, one can then assert that (22) does not hold for all $0 \le k \le K$. On the other hand, by (26) and (76), one has

$$\min_{0\le k\le K} \|x^{k+1} - x^k\|^2 \overset{(76)}{\le} \frac{\max_y H(\hat{x}^0, y) - H^* + \epsilon D_q/4}{L_{\nabla h}(K+1)} + \frac{2\epsilon_0^2(1 + 4D_q^2 L_{\nabla h}^2 \epsilon^{-2})}{L_{\nabla h}^2(K+1)} \overset{(26)}{\le} \frac{\epsilon^2}{16L_{\nabla h}^2},$$

which implies that there exists some $0 \le k \le K$ such that $\|x^{k+1} - x^k\| \le \epsilon/(4L_{\nabla h})$, and thus (22) holds for such k, which contradicts the above assertion. Hence, Algorithm 2 must terminate in at most $K+1$ outer iterations.

Suppose that Algorithm 2 terminates at some iteration $0 \le k \le K$, namely, (22) holds for such k. We next show that its output $(x_\epsilon, y_\epsilon) = (x^{k+1}, y^{k+1})$ is an ϵ-stationary point of (6) and moreover it satisfies (28). Indeed, recall from Lemma 2 that (x^{k+1}, y^{k+1}) is an ϵ_k-stationary point of (20), namely, it satisfies $\mathrm{dist}(0, \partial_x H_k(x^{k+1}, y^{k+1})) \le \epsilon_k$ and $\mathrm{dist}(0, \partial_y H_k(x^{k+1}, y^{k+1})) \le \epsilon_k$. By these, (6), (20) and (21), there exists (u, v) such that

$$u \in \partial_x H(x^{k+1}, y^{k+1}) + 2L_{\nabla h}(x^{k+1} - x^k), \quad \|u\| \le \epsilon_k,$$
$$v \in \partial_y H(x^{k+1}, y^{k+1}) - \epsilon(y^{k+1} - \hat{y}^0)/(2D_q), \quad \|v\| \le \epsilon_k.$$

It then follows that $u - 2L_{\nabla h}(x^{k+1} - x^k) \in \partial_x H(x^{k+1}, y^{k+1})$ and $v + \epsilon(y^{k+1} - \hat{y}^0)/(2D_q) \in \partial_y H(x^{k+1}, y^{k+1})$. These together with (9), (22), and $\epsilon_k \le \epsilon_0 \le \epsilon/2$ (see Algorithm 2) imply that

$$\mathrm{dist}\left(0, \partial_x H(x^{k+1}, y^{k+1})\right) \le \|u - 2L_{\nabla h}(x^{k+1} - x^k)\| \le \|u\| + 2L_{\nabla h}\|x^{k+1} - x^k\| \overset{(22)}{\le} \epsilon_k + \epsilon/2 \le \epsilon,$$
$$\mathrm{dist}\left(0, \partial_y H(x^{k+1}, y^{k+1})\right) \le \|v + \epsilon(y^{k+1} - \hat{y}^0)/(2D_q)\| \le \|v\| + \epsilon\|y^{k+1} - \hat{y}^0\|/(2D_q) \overset{(9)}{\le} \epsilon_k + \epsilon/2 \le \epsilon.$$

Hence, the output (x^{k+1}, y^{k+1}) of Algorithm 2 is an ϵ-stationary point of (6). In addition, (28) holds due to Lemma 3.

Recall from Lemma 2 that the number of evaluations of ∇h and proximal operator of p and q performed at iteration k of Algorithm 2 is at most N_k, respectively, where N_k is defined in (73). Also, one can observe from the above proof and the definition of $\mathbb{K}$ that $|\mathbb{K}| \le K+2$. It then follows that the total number of evaluations of ∇h and proximal operator of p and q in Algorithm 2 is respectively no more than $\sum_{k=0}^{|\mathbb{K}|-2} N_k$. Consequently, to complete the rest of the proof of Theorem 2, it suffices to show that $\sum_{k=0}^{|\mathbb{K}|-2} N_k \le N$, where N is given in (27). Indeed, by (27), (73) and $|\mathbb{K}| \le K+2$, one has

$$\begin{aligned}
\sum_{k=0}^{|\mathbb{K}|-2} N_k &\overset{(73)}{\le} \sum_{k=0}^{K} \left(\left\lceil 96\sqrt{2}\left(1 + (24L_{\nabla h} + 4\epsilon/D_q)\, L_{\nabla h}^{-1}\right)\right\rceil + 2 \right) \times \left\lceil \max\left\{2, \sqrt{\frac{D_q L_{\nabla h}}{\epsilon}}\right\} \right. \\
&\qquad \left. \times \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_{\rm 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_k^2} \right\rceil_+ \\
&\le \left(\left\lceil 96\sqrt{2}\left(1 + (24L_{\nabla h} + 4\epsilon/D_q)\, L_{\nabla h}^{-1}\right)\right\rceil + 2 \right) \max\left\{2, \sqrt{\frac{D_q L_{\nabla h}}{\epsilon}}\right\} \\
&\qquad \times \sum_{k=0}^{K} \left(\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_{\rm 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_k^2} \right)_+ + 1 \right) \\
&\le \left(\left\lceil 96\sqrt{2}\left(1 + (24L_{\nabla h} + 4\epsilon/D_q)\, L_{\nabla h}^{-1}\right)\right\rceil + 2 \right) \max\left\{2, \sqrt{\frac{D_q L_{\nabla h}}{\epsilon}}\right\} \\
&\qquad \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_{\rm 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. \\
&\qquad \left. + K + 1 + 2\sum_{k=0}^{K} \log(k+1) \right) \overset{(27)}{\le} N,
\end{aligned}$$

where the last inequality is due to (27) and $\sum_{k=0}^{K} \log(k+1) \le K\log(K+1)$. This completes the proof of Theorem 2. □

5.3 Proof of the main results in Section 3

In this subsection we prove Theorems 3 and 4. We first establish a lemma below, which will be used to prove Theorem 3 subsequently.

Lemma 4. *Suppose that Assumption 3 holds and $(x_\epsilon, y_\epsilon, z_\epsilon)$ is an ϵ-optimal solution of problem (35) for some $\epsilon > 0$. Let f, $\tilde{f}$, f^*, $f_{\rm low}$ and ρ be given in (29), (32) and (35), respectively. Then we have*

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

Proof. Since $(x_\epsilon, y_\epsilon, z_\epsilon)$ is an ϵ-optimal solution of (35), it follows from Definition 1 that

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

Summing up these inequalities yields

$$\max_z P_\rho(x_\epsilon, y_\epsilon, z) \le \min_{x,y}\max_z P_\rho(x, y, z) + 2\epsilon. \tag{86}$$

Let (x^*, y^*) be an optimal solution of (29). 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 (35), 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

$$\min_{x,y}\max_z P_\rho(x, y, z) \le \max_z P_\rho(x^*, y^*, z) = f^*. \tag{87}$$

It then follows from (35), (86) and (87) that

$$f(x_\epsilon, y_\epsilon) + \rho(\tilde{f}(x_\epsilon, y_\epsilon) - \min_z \tilde{f}(x_\epsilon, z)) \overset{(35)}{=} \max_z P_\rho(x_\epsilon, y_\epsilon, z) \overset{(86)(87)}{\leq} f^* + 2\epsilon,$$

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

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

The conclusion of this lemma directly follows from these and (32). □

We are now ready to prove Theorem 3.

Proof of Theorem 3. Let $\{(x^k, y^k, z^k)\}$ be generated by Algorithm 3 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 (29). Indeed, since (x^k, y^k, z^k) is an ϵ_k-optimal solution of (35) with $\rho = \rho_k$, it follows from Lemma 4 with $(\rho, \epsilon) = (\rho_k, \epsilon_k)$ and $(x_\epsilon, y_\epsilon) = (x^k, y^k)$ that

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

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 $\tilde{f}(x^*, y^*) \leq \min_z \tilde{f}(x^*, z)$ and $f(x^*, y^*) \leq f^*$, 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 (29) as desired. □

We next prove Theorem 4. Before proceeding, we establish a lemma below, which will be used to prove Theorem 4 subsequently.

Lemma 5. *Suppose that Assumption 3 holds and $(x_\epsilon, y_\epsilon, z_\epsilon)$ is an ϵ-stationary point of (35). Let $D_{\mathbf{y}}$, f_{low}, $\tilde{f}$, ρ, and P_ρ be given in (30), (32) and (35), respectively. Then we have*

$$\operatorname{dist}\Big(0, \partial f(x_\epsilon, y_\epsilon) + \rho\partial\tilde{f}(x_\epsilon, y_\epsilon) - (\rho\nabla_x \tilde{f}(x_\epsilon, z_\epsilon); 0)\Big) \leq \epsilon, \quad \operatorname{dist}\big(0, \rho\partial\tilde{f}(x_\epsilon, z_\epsilon)\big) \leq \epsilon,$$
$$\tilde{f}(x_\epsilon, y_\epsilon) \leq \min_z \tilde{f}(x_\epsilon, z) + \rho^{-1}(\max_z P_\rho(x_\epsilon, y_\epsilon, z) - f_{\text{low}}).$$

Proof. Since $(x_\epsilon, y_\epsilon, z_\epsilon)$ is an ϵ-stationary point of (35), it follows from Definition 2 that

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

Using these and the definition of P_ρ in (35), we have

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

In addition, by (35), we have

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

which along with (32) implies that

$$\tilde{f}(x_\epsilon, y_\epsilon) - \min_z \tilde{f}(x_\epsilon, z) \leq \rho^{-1}(\max_z P_\rho(x_\epsilon, y_\epsilon, z) - f_{\text{low}}).$$

This completes the proof of this lemma. □

We are now ready to prove Theorem 4.

Proof of Theorem 4. Observe from (36) that problem (35) can be viewed as

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

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 (35) is in the form of (6) with $H = P_\rho$. By Assumption 3 and $\rho = \varepsilon^{-1}$, one can see that h is

$\widehat{L}$-smooth on its domain, where $\widehat{L}$ is given in (39). Also, notice from Algorithm 4 that $\epsilon_0 = \varepsilon^{3/2} \le \varepsilon/2$ due to $\varepsilon \in (0, 1/4]$. Consequently, Algorithm 2 can be suitably applied to problem (35) with $\rho = \varepsilon^{-1}$ for finding an ϵ-stationary point $(x_\epsilon, y_\epsilon, z_\epsilon)$ of it.

In addition, notice from Algorithm 4 that $\tilde{f}(x^0, y^0) \le \min_y \tilde{f}(x^0, y) + \varepsilon$. Using this, (35) and $\rho = \varepsilon^{-1}$, we obtain

$$\max_z P_\rho(x^0, y^0, z) = f(x^0, y^0) + \rho(\tilde{f}(x^0, y^0) - \min_z \tilde{f}(x^0, z)) \le f(x^0, y^0) + \rho\varepsilon = f(x^0, y^0) + 1. \tag{88}$$

By this and (28) 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

$$\begin{aligned} P_\rho(x_\epsilon, y_\epsilon, z_\epsilon) &\le \max_z P_\rho(x^0, y^0, z) + \varepsilon D_{\mathbf{y}}/4 + 2\varepsilon^3(\widehat{L}^{-1} + 4D_{\mathbf{y}}^2 \widehat{L}\varepsilon^{-2}) \\ &\overset{(88)}{\le} 1 + f(x^0, y^0) + \varepsilon D_{\mathbf{y}}/4 + 2\varepsilon^3(\widehat{L}^{-1} + 4D_{\mathbf{y}}^2 \widehat{L}\varepsilon^{-2}). \end{aligned}$$

It then follows from this and Lemma 5 with $\epsilon = \varepsilon$ and $\rho = \varepsilon^{-1}$ that $(x_\epsilon, y_\epsilon, z_\epsilon)$ satisfies (40) and (41).

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 4. Indeed, by (31), (32) and (35), one has

$$\min_{x,y} \max_z P_\rho(x, y, z) \overset{(35)}{=} \min_{x,y} \{f(x, y) + \rho(\tilde{f}(x, y) - \min_z \tilde{f}(x, z))\} \ge \min_{(x,y) \in \mathcal{X} \times \mathcal{Y}} f(x, y) \overset{(32)}{=} f_{\text{low}}, \tag{89}$$

$$\min_{(x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}} P_\rho(x, y, z) \overset{(35)}{=} \min_{(x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}} \{f(x, y) + \rho(\tilde{f}(x, y) - \tilde{f}(x, z))\} \overset{(31)(32)}{\ge} f_{\text{low}} + \rho(\tilde{f}_{\text{low}} - \tilde{f}_{\text{hi}}). \tag{90}$$

For convenience of the rest proof, let

$$H = P_\rho, \quad H^* = \min_{x,y} \max_z P_\rho(x, y, z), \quad H_{\text{low}} = \min\{P_\rho(x, y, z) | (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}\}. \tag{91}$$

In view of these, (87), (88), (89), (90), and $\rho = \varepsilon^{-1}$, we obtain that

$$\begin{aligned} &\max_z H(x^0, y^0, z) \overset{(88)}{\le} f(x^0, y^0) + 1, \qquad f_{\text{low}} \overset{(89)}{\le} H^* \overset{(87)}{\le} f^*, \\ &H_{\text{low}} \overset{(90)}{\ge} f_{\text{low}} + \rho(\tilde{f}_{\text{low}} - \tilde{f}_{\text{hi}}) = f_{\text{low}} + \varepsilon^{-1}(\tilde{f}_{\text{low}} - \tilde{f}_{\text{hi}}). \end{aligned}$$

Using these and Theorem 2 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_{low} given in (91), we can conclude that Algorithm 4 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_ϵ, y_ϵ) of problem (29) satisfying (40) and (41). □

5.4 Proof of the main results in Section 4

In this subsection we prove Theorems 5 and 6. Before proceeding, we define

$$r = G^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + L_{\tilde{f}}), \quad \mathbb{B}_r^+ = \{\lambda \in \mathbb{R}_+^l : \|\lambda\| \le r\}, \tag{92}$$

where $D_{\mathbf{y}}$ is defined in (30), G is given in Assumption 4(iii), and ϵ and ρ are given in Algorithm 6. In addition, one can observe from (43) and (47) that

$$\min_z \widetilde{P}_\mu(x, z) \le \tilde{f}^*(x) \qquad \forall x \in \mathcal{X}, \tag{93}$$

which will be frequently used later.

We next establish several technical lemmas that will be used to prove Theorem 5 subsequently.

Lemma 6. *Suppose that Assumptions 3 and 4 hold. Let $D_{\mathbf{y}}$, $L_{\tilde{f}}$, G, $\tilde{f}^*$, $\tilde{f}_{\text{hi}}^*$ and $\mathbb{B}_r^+$ be given in* (30), (43), (44), (92) *and Assumption 4, 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 (43) *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

$$\tilde{f}^*(x) = \max_{\lambda} \min_{z} \tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle - \mathscr{I}_{\mathbb{R}^l_+}(\lambda) \qquad \forall x \in \mathcal{X}, \tag{94}$$

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 (43). 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 4(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 $\lambda^* \ge 0$, (30) and Assumption 4 implies that

$$G\sum_{i=1}^{l} \lambda_i^* \le \langle \lambda^*, -\tilde{g}(x,\hat{z}_x)\rangle \le \tilde{f}(x,\hat{z}_x) - \tilde{f}(x,z^*) \le L_{\tilde{f}}\|\hat{z}_x - z^*\| \le L_{\tilde{f}} D_{\mathbf{y}}, \tag{95}$$

where the first inequality is due to Assumption 4(iii), and the third inequality follows from (30) and $L_{\tilde{f}}$-Lipschitz continuity of $\tilde{f}$ (see Assumption 4(i)). By (92), (95) 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 3(i) and 4(iii) that $\tilde{f}(x,\cdot)$ and $\tilde{g}_i(x,\cdot)$, $i = 1, \dots, l$, are convex for any given $x \in \mathcal{X}$. Using this, (43) and the first statement of this lemma, we observe that

$$\tilde{f}^*(x) = \min_{z} \max_{\lambda \in \mathbb{B}^+_r} \tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle \qquad \forall x \in \mathcal{X}. \tag{96}$$

Notice from Assumption 4 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 (96), one can easily verify that $\tilde{f}^*$ is Lipschitz continuous on $\mathcal{X}$. In addition, the finiteness of $\tilde{f}^*_{\rm hi}$ follows from (44), the continuity of $\tilde{f}^*$, and the compactness of $\mathcal{X}$.

(iii) One can observe from (43) 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 3 that the domain of $\tilde{f}(x,\cdot)$ is compact for all $x \in \mathcal{X}$. By this, (96) 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 (94) holds. □

Lemma 7. *Suppose that Assumptions 3 and 4 hold and that $(x_\epsilon, y_\epsilon, z_\epsilon)$ is an ϵ-optimal solution of problem* (50) *for some $\epsilon > 0$. Let $f_{\rm low}$, f, $\widetilde{P}_\mu$, f^*_μ, ρ and μ be given in* (32), (42), (47), (48) *and* (50), *respectively. Then we have*

$$\widetilde{P}_\mu(x_\epsilon, y_\epsilon) \le \min_{z} \widetilde{P}_\mu(x_\epsilon, z) + \rho^{-1}(f^*_\mu - f_{\rm low} + 2\epsilon), \qquad f(x_\epsilon, y_\epsilon) \le f^*_\mu + 2\epsilon. \tag{97}$$

Proof. The proof follows from the same argument as the one for Lemma 4 with f^* and $\tilde{f}$ being replaced by f^*_μ and $\widetilde{P}_\mu$, respectively. □

Lemma 8. *Suppose that Assumptions 3-5 hold. Let $\tilde{f}_{\rm low}$, f^*, $\tilde{f}^*_{\rm hi}$, f^*_μ be defined in* (31), (42), (44) *and* (48), *and L_f, ω and $\bar{\theta}$ be given in Assumptions 4 and 5. Suppose that $\mu \ge (\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})/\bar{\theta}^2$. Then we have*

$$f^*_\mu \le f^* + L_f\omega\Big(\sqrt{\mu^{-1}(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})}\Big). \tag{98}$$

Proof. Let $x \in \mathcal{X}$, $y \in \operatorname{Argmin}_z\{\tilde{f}(x,z)|\tilde{g}(x,z) \le 0\}$ and $z^* \in \operatorname{Argmin}_z \widetilde{P}_\mu(x,z)$ be arbitrarily chosen. One can easily see from (47) and (93) that $\tilde{f}(x,z^*) + \mu \|[\tilde{g}(x,z^*)]_+\|^2 \le \tilde{f}^*(x)$, which together with (31) and (44) implies that

$$\|[\tilde{g}(x,z^*)]_+\|^2 \le \mu^{-1}(\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low}). \tag{99}$$

Since $\mu \ge (\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})/\bar{\theta}^2$, it follows from (99) that $\|[\tilde{g}(x,z^*)]_+\| \le \bar{\theta}$. By this relation, $y \in \operatorname*{Argmin}_z\{\tilde{f}(x,z)|\tilde{g}(x,z) \le 0\}$ and Assumption 5, there exists some $\hat{z}^*$ such that

$$\|y - \hat{z}^*\| \le \omega(\|[\tilde{g}(x,z^*)]_+\|), \qquad \hat{z}^* \in \operatorname*{Argmin}_z \left\{\tilde{f}(x,z) \big| \, \|[\tilde{g}(x,z)]_+\| \le \|[\tilde{g}(x,z^*)]_+\|\right\}. \tag{100}$$

In view of (47), $z^* \in \operatorname{Argmin}_z \widetilde{P}_\mu(x,z)$ and the second relation in (100), one can observe that $\hat{z}^* \in \operatorname{Argmin}_z \widetilde{P}_\mu(x,z)$, which along with (48) yields $f(x,\hat{z}^*) \ge f^*_\mu$. Also, using (100) and L_f-Lipschitz continuity of f (see Assumption 4), we have

$$f(x,y) - f(x,\hat{z}^*) \ge -L_f\|y - \hat{z}^*\| \overset{(100)}{\ge} -L_f\omega(\|[\tilde{g}(x,z^*)]_+\|).$$

Taking minimum over $x \in \mathcal{X}$ and $y \in \operatorname{Argmin}_z\{\tilde{f}(x,z)|\tilde{g}(x,z) \le 0\}$ on both sides of this relation, and using (42), (99), $f(x,\hat{z}^*) \ge f^*_\mu$ and the monotonicity of ω, we can conclude that (98) holds. □

Lemma 9. *Suppose that Assumptions 3-5 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* (31), (32), (42), (43), (44) *and* (50), *and* L_f, ω *and* $\bar{\theta}$ *be given in Assumptions 4 and 5, respectively. Suppose that* $\mu \ge (\tilde{f}^*_{\rm hi} - \tilde{f}_{\rm low})/\bar{\theta}^2$ *and* $(x_\epsilon, y_\epsilon, z_\epsilon)$ *is an* ϵ*-optimal solution of problem* (50) *for some* $\epsilon > 0$. *Then we have*

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

Proof. By (47), (93), and the first relation in (97), one has

$$\tilde{f}(x_\epsilon,y_\epsilon) + \mu\|[\tilde{g}(x_\epsilon,y_\epsilon)]_+\|^2 \overset{(47)}{=} \widetilde{P}_\mu(x_\epsilon,y_\epsilon) \overset{(93)(97)}{\le} \tilde{f}^*(x_\epsilon) + \rho^{-1}(f^*_\mu - f_{\rm low} + 2\epsilon).$$

It then follows from this and (31) that

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

In addition, recall from (97) that $f(x_\epsilon,y_\epsilon) \le f^*_\mu + 2\epsilon$. The conclusion of this lemma then follows from these three relations and (98). □

We are now ready to prove Theorem 5.

Proof of Theorem 5. Let $\{(x^k,y^k,z^k)\}$ be generated by Algorithm 5 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 (42). Indeed, since (x^k,y^k,z^k) is an ϵ_k-optimal solution of (50) with $(\rho,\mu) = (\rho_k,\mu_k)$ and $\lim_{k\to\infty}\mu_k = \infty$, it follows from Lemma 9 with $(\rho,\mu,\epsilon) = (\rho_k,\mu_k,\epsilon_k)$ and $(x_\epsilon,y_\epsilon) = (x^k,y^k)$ that for all sufficiently large k, one has

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

By the continuity of f, $\tilde{f}$ and $\tilde{f}^*$ (see Assumption 3(i) and Lemma 6(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 (42) and (43) imply that $f(x^*,y^*) = f^*$ and $y^* \in \operatorname{Argmin}_z\{\tilde{f}(x^*,z)|\tilde{g}(x^*,z) \le 0\}$. Hence, (x^*,y^*) is an optimal solution of (42) as desired. □

We next prove Theorem 6. Before proceeding, we establish several technical lemmas below, which will be used to prove Theorem 6 subsequently.

Lemma 10. *Suppose that Assumptions 3 and 4 hold and that $(x_\epsilon, y_\epsilon, z_\epsilon)$ is an ϵ-stationary point of problem* (50) *for some $\epsilon > 0$. Let $D_{\mathbf{y}}$, $\tilde{g}$, ρ, μ, L_f, $L_{\tilde{f}}$ and G be given in* (30), (42), (50) *and Assumption 4, respectively. Then we have*

$$\|[\tilde{g}(x_\epsilon, z_\epsilon)]_+\| \leq (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + L_{\tilde{f}}), \tag{101}$$

$$\|[\tilde{g}(x_\epsilon, y_\epsilon)]_+\| \leq (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}}). \tag{102}$$

Proof. We first prove (101). Since $(x_\epsilon, y_\epsilon, z_\epsilon)$ is an ϵ-stationary point of (50), it follows from Definition 2 that $\mathrm{dist}(0, \partial_z P_{\rho,\mu}(x_\epsilon, y_\epsilon, z_\epsilon)) \leq \epsilon$. Also, by (47) and (50), one has

$$P_{\rho,\mu}(x, y, z) = f(x, y) + \rho(\tilde{f}(x, y) + \mu \|[\tilde{g}(x, y)]_+\|^2) - \rho(\tilde{f}(x, z) + \mu \|[\tilde{g}(x, z)]_+\|^2). \tag{103}$$

Using these relations, we have

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

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

$$\Big\|s + 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\epsilon, z_\epsilon)]_+ \nabla_z \tilde{g}_i(x_\epsilon, z_\epsilon)\Big\| \leq \rho^{-1}\epsilon. \tag{104}$$

Let $\hat{z}_{x_\epsilon}$ and G be given in Assumption 4(iii). It then follows that $\hat{z}_{x_\epsilon} \in \mathcal{Y}$ and $-\tilde{g}_i(x_\epsilon, \hat{z}_{x_\epsilon}) \geq G > 0$ for all i. Notice that $[\tilde{g}_i(x_\epsilon, z_\epsilon)]_+ \tilde{g}_i(x_\epsilon, z_\epsilon) \geq 0$ for all i and $\|z_\epsilon - \hat{z}_{x_\epsilon}\| \leq D_{\mathbf{y}}$ due to (30). Using these, (104), and the convexity of $\tilde{f}(x_\epsilon, \cdot)$ and $\tilde{g}_i(x_\epsilon, \cdot)$ for all i, we have

$$\begin{aligned}
&\tilde{f}(x_\epsilon, z_\epsilon) - \tilde{f}(x_\epsilon, \hat{z}_{x_\epsilon}) + 2\mu G \sum_{i=1}^{l} [\tilde{g}_i(x_\epsilon, z_\epsilon)]_+ \leq \tilde{f}(x_\epsilon, z_\epsilon) - \tilde{f}(x_\epsilon, \hat{z}_{x_\epsilon}) - 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\epsilon, z_\epsilon)]_+ \tilde{g}_i(x_\epsilon, \hat{z}_{x_\epsilon}) \\
&\leq \tilde{f}(x_\epsilon, z_\epsilon) - \tilde{f}(x_\epsilon, \hat{z}_{x_\epsilon}) + 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\epsilon, z_\epsilon)]_+ (\tilde{g}_i(x_\epsilon, z_\epsilon) - \tilde{g}_i(x_\epsilon, \hat{z}_{x_\epsilon})) \\
&\leq \langle s, z_\epsilon - \hat{z}_{x_\epsilon} \rangle + 2\mu \sum_{i=1}^{l} [\tilde{g}_i(x_\epsilon, z_\epsilon)]_+ \langle \nabla_z \tilde{g}_i(x_\epsilon, z_\epsilon), z_\epsilon - \hat{z}_{x_\epsilon} \rangle \\
&= \langle s + 2\mu \sum_{i=1}^{l} [\tilde{g}(x_\epsilon, z_\epsilon)]_+ \nabla_z \tilde{g}_i(x_\epsilon, z_\epsilon), z_\epsilon - \hat{z}_{x_\epsilon} \rangle \leq \rho^{-1} D_{\mathbf{y}} \epsilon,
\end{aligned} \tag{105}$$

where the first inequality is due to $-\tilde{g}_i(x_\epsilon, \hat{z}_{x_\epsilon}) \geq G$ for all i, the second inequality follows from $[\tilde{g}_i(x_\epsilon, z_\epsilon)]_+ \tilde{g}_i(x_\epsilon, z_\epsilon) \geq 0$ for all i, the third inequality is due to $s \in \partial_z \tilde{f}(x_\epsilon, z_\epsilon)$ and the convexity of $\tilde{f}(x_\epsilon, \cdot)$ and $\tilde{g}_i(x_\epsilon, \cdot)$ for all i, and the last inequality follows from (30) and (104). In view of (30), (105), and $L_{\tilde{f}}$-Lipschitz continuity of $\tilde{f}(x, y)$ (see Assumption 4), one has

$$\begin{aligned}
\|[\tilde{g}(x_\epsilon, z_\epsilon)]_+\| \leq \sum_{i=1}^{l} [\tilde{g}_i(x_\epsilon, z_\epsilon)]_+ &\overset{(105)}{\leq} (2\mu G)^{-1}(\rho^{-1} D_{\mathbf{y}} \epsilon + \tilde{f}(x_\epsilon, \hat{z}_{x_\epsilon}) - \tilde{f}(x_\epsilon, z_\epsilon)) \\
&\leq (2\mu G)^{-1}(\rho^{-1} D_{\mathbf{y}} \epsilon + L_{\tilde{f}} \|\hat{z}_{x_\epsilon} - z_\epsilon\|) \overset{(30)}{\leq} (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + L_{\tilde{f}}).
\end{aligned}$$

Hence, (101) holds.

We next prove (102). Since $(x_\epsilon, y_\epsilon, z_\epsilon)$ is an ϵ-stationary point of (50), it follows from Definition 2 that $\mathrm{dist}(0, \partial_y P_{\rho,\mu}(x_\epsilon, y_\epsilon, z_\epsilon)) \leq \epsilon$. This together with (103) implies that

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

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

$$\|s + \rho\tilde{s} + 2\rho\mu \nabla_y \tilde{g}(x_\epsilon, y_\epsilon)[\tilde{g}(x_\epsilon, y_\epsilon)]_+\| \leq \epsilon. \tag{106}$$

Let $\bar{\mathcal{A}}(x_\epsilon, y_\epsilon) = \{i | \tilde{g}_i(x_\epsilon, y_\epsilon) > 0, 1 \leq i \leq l\}$, $\hat{z}_{x_\epsilon}$ and G be given in Assumption 4(iii). It then follows that $\hat{z}_{x_\epsilon} \in \mathcal{Y}$ and $-\tilde{g}_i(x_\epsilon, \hat{z}_{x_\epsilon}) \geq G > 0$ for all i. Using these and the convexity of $\tilde{g}_i(x_\epsilon, \cdot)$ for all i, we have

$$\begin{aligned}
&\langle \nabla_y \tilde{g}(x_\epsilon, y_\epsilon)[\tilde{g}(x_\epsilon, y_\epsilon)]_+, y_\epsilon - \hat{z}_{x_\epsilon} \rangle = \sum_{i \in \bar{\mathcal{A}}(x_\epsilon, y_\epsilon)} \langle \nabla_y \tilde{g}_i(x_\epsilon, y_\epsilon), y_\epsilon - \hat{z}_{x_\epsilon} \rangle [g_i(x_\epsilon, y_\epsilon)]_+ \\
&\geq \sum_{i \in \bar{\mathcal{A}}(x_\epsilon, y_\epsilon)} (\tilde{g}_i(x_\epsilon, y_\epsilon) - \tilde{g}_i(x_\epsilon, \hat{z}_{x_\epsilon}))[\tilde{g}_i(x_\epsilon, y_\epsilon)]_+ \\
&\geq \sum_{i \in \bar{\mathcal{A}}(x_\epsilon, y_\epsilon)} G[\tilde{g}_i(x_\epsilon, y_\epsilon)]_+ = G \sum_{i=1}^{l} [\tilde{g}_i(x_\epsilon, y_\epsilon)]_+ \geq G \, \|[\tilde{g}(x_\epsilon, y_\epsilon)]_+\| , \qquad (107)
\end{aligned}$$

where the first inequality follows from the convexity of $\tilde{g}(x_\epsilon, \cdot)$ and the second inequality is due to $-\tilde{g}_i(x_\epsilon, \hat{z}_{x_\epsilon}) \geq G$. It then follows from this, (106) and (107) that

$$\begin{aligned}
D_{\mathbf{y}}\epsilon \ &\geq \|s + \rho\tilde{s} + 2\rho\mu\nabla_y \tilde{g}(x_\epsilon, y_\epsilon)[\tilde{g}(x_\epsilon, y_\epsilon)]_+\| \cdot \|y_\epsilon - \hat{z}_{x_\epsilon}\| \\
&\geq \langle s + \rho\tilde{s} + 2\rho\mu\nabla_y \tilde{g}(x_\epsilon, y_\epsilon)[\tilde{g}(x_\epsilon, y_\epsilon)]_+, y_\epsilon - \hat{z}_{x_\epsilon} \rangle \\
&= \langle s + \rho\tilde{s}, y_\epsilon - \hat{z}_{x_\epsilon} \rangle + 2\rho\mu \langle \nabla_y \tilde{g}(x_\epsilon, y_\epsilon)[\tilde{g}(x_\epsilon, y_\epsilon)]_+, y_\epsilon - \hat{z}_{x_\epsilon} \rangle \\
&\overset{(107)}{\geq} -(\|s\| + \rho\|\tilde{s}\|)\,\|y_\epsilon - \hat{z}_{x_\epsilon}\| + 2\rho\mu G \,\|[\tilde{g}(x_\epsilon, y_\epsilon)]_+\| \\
&\geq -(L_f + \rho L_{\tilde{f}})D_{\mathbf{y}} + 2\rho\mu G \,\|[\tilde{g}(x_\epsilon, y_\epsilon)]_+\| , \qquad (108)
\end{aligned}$$

where the last inequality follows from $\|y_\epsilon - \hat{z}_{x_\epsilon}\| \leq D_{\mathbf{y}}$ and the fact that $\|s\| \leq L_f$ and $\|\tilde{s}\| \leq L_{\tilde{f}}$, which are due to (30), $s \in \partial_y f(x_\epsilon, y_\epsilon)$, $\tilde{s} \in \partial_y \tilde{f}(x_\epsilon, y_\epsilon)$ and Assumption 4(i). By (108), one can immediately see that (102) holds. □

Lemma 11. *Suppose that Assumptions 3 and 4 hold. Let* f, $\tilde{f}$, $\tilde{g}$, $D_{\mathbf{y}}$, $f_{\rm low}$, $\tilde{f}^*$ *and* $P_{\rho,\mu}$ *be given in* (29)*,* (30)*,* (32)*,* (43) *and* (50)*,* L_f, $L_{\tilde{f}}$ *and* G *be given in Assumptions 3 and 4,* $(x_\epsilon, y_\epsilon, z_\epsilon)$ *be an* ϵ*-stationary point of* (50) *for some* $\epsilon > 0$*, and*

$$\tilde{\lambda} = 2\mu[\tilde{g}(x_\epsilon, z_\epsilon)]_+, \quad \hat{\lambda} = 2\rho\mu[\tilde{g}(x_\epsilon, y_\epsilon)]_+. \qquad (109)$$

Then we have

$$\mathrm{dist}\Big(\partial f(x_\epsilon, y_\epsilon) + \rho\partial\tilde{f}(x_\epsilon, y_\epsilon) - \rho(\nabla_x \tilde{f}(x_\epsilon, z_\epsilon) + \nabla_x \tilde{g}(x_\epsilon, z_\epsilon)\tilde{\lambda}; 0) + \nabla\tilde{g}(x_\epsilon, y_\epsilon)\hat{\lambda}\Big) \leq \epsilon, \qquad (110)$$

$$\mathrm{dist}\Big(0, \rho(\partial_z \tilde{f}(x_\epsilon, z_\epsilon) + \nabla_z \tilde{g}(x_\epsilon, z_\epsilon)\tilde{\lambda})\Big) \leq \epsilon, \qquad (111)$$

$$\|[\tilde{g}(x_\epsilon, z_\epsilon)]_+\| \leq (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + L_{\tilde{f}}), \qquad (112)$$

$$|\langle \tilde{\lambda}, \tilde{g}(x_\epsilon, z_\epsilon) \rangle| \leq (2\mu)^{-1} G^{-2} D_{\mathbf{y}}^2 (\rho^{-1}\epsilon + L_{\tilde{f}})^2, \qquad (113)$$

$$|\tilde{f}(x_\epsilon, y_\epsilon) - \tilde{f}^*(x_\epsilon)| \leq \max\big\{\rho^{-1}(\max_z P_{\rho,\mu}(x_\epsilon, y_\epsilon, z) - f_{\rm low}), (2\mu)^{-1} G^{-2} D_{\mathbf{y}}^2 L_{\tilde{f}}(\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}})\big\}, \qquad (114)$$

$$\|[\tilde{g}(x_\epsilon, y_\epsilon)]_+\| \leq (2\mu G)^{-1} D_{\mathbf{y}}(\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}}), \qquad (115)$$

$$|\langle \hat{\lambda}, \tilde{g}(x_\epsilon, y_\epsilon) \rangle| \leq (2\mu)^{-1} \rho G^{-2} D_{\mathbf{y}}^2 (\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}})^2. \qquad (116)$$

Proof. Since $(x_\epsilon, y_\epsilon, z_\epsilon)$ is an ϵ-stationary point of (50), it easily follows from (103), (109) and Definition 2 that (110) and (111) hold. Also, it follows from (101) and (102) that (112) and (115) hold. In addition, in view of (109), (112) and (115), one has

$$\begin{aligned}
|\langle \tilde{\lambda}, \tilde{g}(x_\epsilon, z_\epsilon) \rangle| &\overset{(109)}{=} 2\mu \,\|[\tilde{g}(x_\epsilon, z_\epsilon)]_+\|^2 \overset{(112)}{\leq} (2\mu)^{-1} G^{-2} D_{\mathbf{y}}^2 (\rho^{-1}\epsilon + L_{\tilde{f}})^2, \\
|\langle \hat{\lambda}, \tilde{g}(x_\epsilon, y_\epsilon) \rangle| &\overset{(109)}{=} 2\rho\mu \,\|[\tilde{g}(x_\epsilon, y_\epsilon)]\|_+\|^2 \overset{(115)}{\leq} (2\mu)^{-1} \rho G^{-2} D_{\mathbf{y}}^2 (\rho^{-1}\epsilon + \rho^{-1}L_f + L_{\tilde{f}})^2,
\end{aligned}$$

and hence (113) and (116) hold. Also, observe from the definition of $P_{\rho,\mu}$ in (50) that

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

Using this, (32), (47) and (93), we obtain that

$$\begin{aligned}\tilde f(x_\epsilon,y_\epsilon)+\mu\left\|[\tilde g(x_\epsilon,y_\epsilon)]_+\right\|^2 \overset{(47)}{=} \widetilde{P}_\mu(x_\epsilon,y_\epsilon) = \ &\min_z \widetilde{P}_\mu(x_\epsilon,z)+\rho^{-1}(\max_z P_{\rho,\mu}(x_\epsilon,y_\epsilon,z)-f(x_\epsilon,y_\epsilon))\\ \overset{(32)(93)}{\le}\ &\tilde f^*(x_\epsilon)+\rho^{-1}(\max_z P_{\rho,\mu}(x_\epsilon,y_\epsilon,z)-f_{\rm low}). \qquad (117)\end{aligned}$$

On the other hand, let $\lambda^*\in\mathbb{R}^l_+$ be an optimal Lagrangian multiplier of problem (43) with $x=x_\epsilon$. It then follows from Lemma 6(i) that $\|\lambda^*\|\le G^{-1}L_{\tilde f}D_{\mathbf{y}}$. Using these and (115), we have

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

By this and (117), one can see that (114) holds. □

We are now ready to prove Theorem 6.

Proof of Theorem 6. Observe from (51) that problem (50) can be viewed as

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

where $h(x,y,z)=f_1(x,y)+\rho\tilde f_1(x,y)+\rho\mu\left\|[\tilde g(x,y)]_+\right\|^2-\rho\tilde f_1(x,z)-\rho\mu\left\|[\tilde g(x,z)]_+\right\|^2$, $p(x,y)=f_2(x)+\rho\tilde f_2(y)$ and $q(z)=\rho\tilde f_2(z)$. Hence, problem (50) is in the form of (6) with $H=P_{\rho,\mu}$. By Assumption 3, (45), (46), $\rho=\varepsilon^{-1}$ and $\mu=\varepsilon^{-2}$, one can see that h is $\widetilde L$-smooth on its domain, where $\widetilde L$ is given in (61). Also, notice from Algorithm 6 that $\epsilon_0=\varepsilon^{5/2}\le\varepsilon/2=\epsilon/2$ due to $\varepsilon\in(0,1/4]$. Consequently, Algorithm 2 can be suitably applied to problem (50) with $\rho=\varepsilon^{-1}$ and $\mu=\varepsilon^{-2}$ for finding an ϵ-stationary point $(x_\epsilon,y_\epsilon,z_\epsilon)$ of it.

In addition, notice from Algorithm 6 that $\widetilde P_\mu(x^0,y^0)\le\min_y\widetilde P_\mu(x^0,y)+\varepsilon$. Using this, (50) and $\rho=\varepsilon^{-1}$, we obtain

$$\max_z P_{\rho,\mu}(x^0,y^0,z)\overset{(50)}{=} f(x^0,y^0)+\rho(\widetilde P_\mu(x^0,y^0)-\min_z\widetilde P_\mu(x^0,z))\le f(x^0,y^0)+\rho\varepsilon=f(x^0,y^0)+1. \quad (118)$$

By this and (28) 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

$$\begin{aligned}P_{\rho,\mu}(x_\epsilon,y_\epsilon,z_\epsilon)\le\ &\max_z P_{\rho,\mu}(x^0,y^0,z)+\varepsilon D_{\mathbf{y}}/4+2\varepsilon^5(\widetilde L^{-1}+4D_{\mathbf{y}}^2\widetilde L\varepsilon^{-2})\\ \overset{(118)}{\le}\ &1+f(x^0,y^0)+\varepsilon D_{\mathbf{y}}/4+2\varepsilon^5(\widetilde L^{-1}+4D_{\mathbf{y}}^2\widetilde L\varepsilon^{-2}).\end{aligned}$$

It then follows from this and Lemma 11 with $\epsilon=\varepsilon$, $\rho=\varepsilon^{-1}$ and $\mu=\varepsilon^{-2}$ that $(x_\epsilon,y_\epsilon,z_\epsilon)$ satisfies the relations (62)-(68).

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 6. Indeed, by (31), (32), (45), (47) and (50), one has

$$\min_{x,y}\max_z P_{\rho,\mu}(x,y,z)\overset{(50)}{=}\min_{x,y}\{f(x,y)+\rho(\widetilde P_\mu(x,y)-\min_z\widetilde P_\mu(x,z))\}\ge\min_{(x,y)\in\mathcal{X}\times\mathcal{Y}}f(x,y)\overset{(32)}{=}f_{\rm low}, \quad (119)$$

$$\begin{aligned}&\min\{P_{\rho,\mu}(x,y,z)|(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\}\overset{(50)}{=}\min\{f(x,y)+\rho(\widetilde P_\mu(x,y)-\widetilde P_\mu(x,z))|(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\}\\ &\overset{(47)}{=}\min\{f(x,y)+\rho(\tilde f(x,y)+\mu\|[\tilde g(x,y)]_+\|^2-\tilde f(x,z)-\mu\|[\tilde g(x,z)]_+\|^2)|(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\}\\ &\ge f_{\rm low}+\rho(\tilde f_{\rm low}-\tilde f_{\rm hi})-\rho\mu\tilde g_{\rm hi}^2, \qquad (120)\end{aligned}$$

where the last inequality follows from (31), (32) and (45). In addition, let (x^*,y^*) be an optimal solution of (42). It then follows that $f(x^*,y^*)=f^*$ and $[\tilde g(x^*,y^*)]_+=0$. By these, (31), (47) and (50), one has

$$\begin{aligned}\min_{x,y}\max_z P_{\rho,\mu}(x,y,z)\le\ &\max_z P_{\rho,\mu}(x^*,y^*,z)\overset{(50)}{=}f(x^*,y^*)+\rho\left(\widetilde P_\mu(x^*,y^*)-\min_z\widetilde P_\mu(x^*,z)\right)\\ &\overset{(47)}{=}f(x^*,y^*)+\rho(\tilde f(x^*,y^*)+\mu\|[\tilde g(x^*,y^*)]_+\|^2-\min_z\{\tilde f(x^*,z)+\mu\|[\tilde g(x^*,z)]_+\|^2\})\\ &\overset{(31)}{\le}f^*+\rho(\tilde f_{\rm hi}-\tilde f_{\rm low}). \qquad (121)\end{aligned}$$

For convenience of the rest proof, let

$$H = P_{\rho,\mu}, \quad H^* = \min_{x,y}\max_{z} P_{\rho,\mu}(x,y,z), \quad H_{\rm low} = \min\{P_{\rho,\mu}(x,y,z)|(x,y,z) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\}. \tag{122}$$

In view of these, (118), (119), (120), (121), $\rho = \varepsilon^{-1}$ and $\mu = \varepsilon^{-2}$, we obtain that

$$\max_z H(x^0,y^0,z) \overset{(118)}{\leq} f(x^0,y^0)+1, \quad f_{\rm low} \overset{(119)}{\leq} H^* \overset{(121)}{\leq} f^* + \rho(\tilde{f}_{\rm hi} - \tilde{f}_{\rm low}) = f^* + \varepsilon^{-1}(\tilde{f}_{\rm hi} - \tilde{f}_{\rm low}),$$
$$H_{\rm low} \overset{(120)}{\geq} f_{\rm low} + \rho(\tilde{f}_{\rm low} - \tilde{f}_{\rm hi}) - \rho\mu\tilde{g}_{\rm hi}^2 = f_{\rm low} + \varepsilon^{-1}(\tilde{f}_{\rm low} - \tilde{f}_{\rm hi}) - \varepsilon^{-3}\tilde{g}_{\rm hi}^2.$$

Using these and Theorem 2 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 (122), we can conclude that Algorithm 6 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_ϵ, y_ϵ) of problem (42) satisfying (62)-(68). □

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 [39]) for handling the constraint $g(x,y) \leq 0$. In addition, a single subproblem with static penalty and tolerance parameters is solved in our methods (Algorithms 4 and 6), which may be conservative in practice. To make the methods possibly practically more efficient, it would be natural to modify them by solving a sequence of subproblems with dynamic penalty and tolerance parameters instead. These along with numerical experiments 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] G. H. Chen and R. T. Rockafellar. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

[6] T. Chen, Y. Sun, and W. Yin. A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*, 2021.

[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] Z. Guo and T. Yang. Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *arXiv e-prints*, pages arXiv–2105, 2021.

[19] 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.

[20] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

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

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

[23] 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.

[24] K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Nonasymptotic analysis and faster algorithms. *arXiv preprint arXiv:2010.07962*, 2020.

[25] A. Kaplan and R. Tichatschke. Proximal point methods and nonconvex optimization. *Journal of global Optimization*, 13(4):389–406, 1998.

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

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

[28] 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.

[29] D. Kovalev and A. Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. *arXiv preprint arXiv:2205.05653*, 2022.

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

[31] 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.

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

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

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

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

[36] 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.

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

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

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

[40] 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.

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

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

[43] 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.

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

[45] 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.

[46] 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.

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

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

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

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

[51] 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.