Solving bilevel optimization via sequential minimax optimization

Zhaosong Lu * Sanyou Mei *

July 4, 2023

Abstract

In this paper we propose a novel method called sequential minimax optimization (SMO) for solving a class of constrained bilevel optimization problems in which the lower-level part is a possibly nonsmooth convex optimization problem, while the upper-level part is a possibly nonconvex optimization problem. Specifically, SMO applies a first-order method to solve a sequence of minimax subproblems, which are obtained by employing a hybrid of modified augmented Lagrangian and penalty schemes on the bilevel optimization problems. Under suitable assumptions, we establish an *operation complexity* of $\mathcal{O}(\varepsilon^{-7}\log\varepsilon^{-1})$, which is measured in terms of fundamental operations, for SMO in finding an ε-KKT (Karush-Kuhn-Tucker) solution of the bilevel optimization problems. Preliminary numerical results are presented to illustrate the performance of SMO.

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

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

1 Introduction

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

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

Bilevel optimization has widely been used in many areas, including adversarial training [43, 44, 55], continual learning [37], hyperparameter tuning [3, 17, 46], image reconstruction [8], meta-learning [4, 27, 49], neural architecture search [15, 35], reinforcement learning [22, 30], and Stackelberg games [58]. More applications about it can be found in [2, 7, 11, 12, 13, 52] and the references therein. Theoretical properties including optimality conditions of (1) have been extensively studied in the literature (e.g., see [13, 14, 41, 57, 61]).

Numerous methods have been developed for solving some special cases of (1). For example, constraint-based methods [21, 51], deterministic gradient-based methods [16, 17, 19, 23, 42, 48, 49], and stochastic gradient-based methods [6, 20, 22, 24, 25, 28, 29, 32, 33, 60] were proposed for solving (1) with $\tilde{g} \equiv 0$, f, $\tilde{f}$ being smooth, and $\tilde{f}$ being *strongly convex* with respect to y. For a similar case as this but with $\tilde{f}$ being *convex* with respect to y, a zeroth-order method was recently proposed in [5], and also numerical methods were developed in [34, 54, 63] by solving (1) as a single or sequential smooth constrained optimization problems. Besides, when all the functions in (1) are smooth and $\tilde{f}$, $\tilde{g}$ are *convex* with respect to y, gradient-type methods were proposed by solving a mathematical program with equilibrium constraints resulting from replacing the lower-level optimization problem of (1) by its first-order optimality conditions (e.g., see [1, 40, 47]). Recently, difference-of-convex (DC) algorithms were developed in [62] for solving (1) with f being a DC function, and $\tilde{f}$, $\tilde{g}$ being convex functions. In addition, penalty methods were proposed in [26, 39, 50] for solving (1). Notably, the paper [39] demonstrates *for the first time* that bilevel optimization can be approximately solved as minimax optimization. Specifically, it reformulates bilevel optimization as minimax optimization by a novel double penalty scheme and proposes a first-order method with *complexity guarantees* for bilevel optimization via solving a single minimax problem. Lately, a practically efficient multi-stage gradient descent and ascent algorithm (GDA) was developed in [59]

[*] 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.

for (1) with $\tilde{g} \equiv 0$, f being convex and Lipschitz continuous, and $\tilde{f}$ being strongly convex and Lipschitz smooth via solving the aforementioned minimax reformulation of (1). More discussion on algorithmic development for bilevel optimization can be found in [2, 7, 13, 36, 53, 57]) and the references therein.

In this paper, we consider problem (1) that has at least one optimal solution and satisfies the following assumptions.

Assumption 1. *(i) $f(x,y) = f_1(x,y) + f_2(x)$ and $\tilde{f}(x,y) = \tilde{f}_1(x,y) + \tilde{f}_2(y)$ are respectively L_f- and $L_{\tilde{f}}$-Lipschitz 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) $\tilde{g} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^l$ is $L_{\tilde{g}}$-Lipschitz continuous and $L_{\nabla \tilde{g}}$-smooth on $\mathcal{X} \times \mathcal{Y}$, and $\tilde{g}_i(x,\cdot)$ is convex for all $x \in \mathcal{X}$ and $i = 1, 2, \ldots, l$.

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

Due to the sophisticated structure described in Assumption 1, existing methods except the first-order penalty method [39] are not applicable to problem (1) in general. Besides, despite being applicable to (1), the first-order penalty method [39] may suffer practical inefficiency because (i) it solves a *single* minimax problem; (ii) the associated minimax problem results from a *double penalty* scheme. Indeed, it solves the single minimax problem

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

for some suitable penalty parameters $\rho, \mu > 0$. Notice that $\tilde{f}(x,z) + \mu \left\|[\tilde{g}(x,z)]_+\right\|^2$ is the classical quadratic penalty function associated with the lower level optimization problem of (1), while $\max_z f(x,y) + \rho\big(\tilde{f}(x,y) + \mu \left\|[\tilde{g}(x,y)]_+\right\|^2 - \tilde{f}(x,z) - \mu \left\|[\tilde{g}(x,z)]_+\right\|^2\big)$ can be viewed as an exact penalty function associated with the problem

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

It is well known that augmented Lagrangian methods typically outperform quadratic penalty methods in practice for nonlinear constrained optimization. Inspired by this and the above discussion, in this paper we propose a novel sequential minimax optimization (SMO) method to solve problem (1). Specifically, instead of using the aforementioned quadratic penalty function, we use a *modified* augmented Lagrangian function associated with the lower level optimization problem of (1) given by $\tilde{f}(x,z) + \frac{1}{2\rho\mu}\left(\|[\lambda + \mu\tilde{g}(x,z)]_+\|^2 - \|\lambda\|^2\right)$ for $\lambda \in \mathbb{R}^l_+$ and $\mu > 0$.[2] Performing such a replacement in (2) results in a new minimax problem

$$\min_{x,y} \max_{z} f(x,y) + \rho\Big(\tilde{f}(x,y) + \frac{1}{2\rho\mu}\|[\lambda + \mu\tilde{g}(x,y)]_+\|^2 - \tilde{f}(x,z) - \frac{1}{2\rho\mu}\|[\lambda + \mu\tilde{g}(x,z)]_+\|^2\Big). \tag{3}$$

Our SMO method solves a sequence of minimax subproblems in the form of (3). Specifically, let $\{\rho_k\}$, $\{\mu_k\}$, $(x^0, y^0, z^0, \lambda^0)$ be given. At each iteration $k \geq 0$, SMO finds an approximate solution $(x^{k+1}, y^{k+1}, z^{k+1})$ of problem (3) with $\lambda = \lambda^k$, $\rho = \rho_k$ and $\mu = \mu_k$, starting at (x^k, y^k, z^k), and then updates λ^{k+1} according to $\lambda^{k+1} = [\lambda^k + \mu_k \tilde{g}(x^{k+1}, z^{k+1})]_+$. The detailed motivation and presentation of our SMO are given in Section 2. It shall be mentioned that SMO enjoys the following notable features.

- It uses only the first-order information of the problem. Specifically, its fundamental operations consist only of gradient evaluation of $\tilde{g}$ and the smooth component of f and $\tilde{f}$ and also proximal operator evaluation of the nonsmooth component of f and $\tilde{f}$ (see Algorithm 1).

- It has theoretical guarantees on operation complexity, which is measured by the aforementioned fundamental operations, for finding an ε-KKT solution of (1). Specifically, it enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-7} \log \varepsilon^{-1})$ (see Theorem 1).

[2] The standard augmented Lagrangian function associated with the lower-level optimization of (1) is $\tilde{f}(x,z) + \frac{1}{2\mu}\left(\|[\lambda + \mu\tilde{g}(x,z)]_+\|^2 - \|\lambda\|^2\right)$. Thus, $\tilde{f}(x,z) + \frac{1}{2\rho\mu}\left(\|[\lambda + \mu\tilde{g}(x,z)]_+\|^2 - \|\lambda\|^2\right)$ can be viewed as a modified augmented Lagrangian function.

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

1.1 Notation and terminology

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

A function or mapping ϕ is said to be *L_ϕ-Lipschitz continuous* on a set S if $\|\phi(x) - \phi(x')\| \le L_\phi \|x - x'\|$ 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 $p : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, associated with p is denoted by prox_p, that is,

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

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

In addition, we introduce an (approximate) primal-dual stationary point (e.g., see [9, 10, 31]) for a general minimax problem

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

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

Definition 1. *A point (x, y) is said to be a primal-dual stationary point of the minimax problem (4) if*

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

In addition, for any $\epsilon > 0$, a point (x_ϵ, y_ϵ) is said to be an ϵ-primal-dual stationary point of the minimax problem (4) if

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

2 A sequential minimax optimization method for problem (1)

In this section, we propose a sequential minimax optimization (SMO) method for finding an approximate KKT solution of (1) and study its complexity.

To motivate our method, let $\rho, \mu > 0$ and $\lambda \in \mathbb{R}^l_+$ be given. Applying an augmented Lagrangian scheme, one can migrate the constraint $\tilde{g}(x, y) \le 0$ of the lower-level optimization of (1) to its objective function and obtain an approximation to (1) given by

$$\begin{array}{ll} \min & f(x, y) \\ \text{s.t.} & y \in \operatorname*{Argmin}_z \{\tilde{f}(x, z) + \frac{1}{2\rho\mu}\left(\|[\lambda + \mu\tilde{g}(x, z)]_+\|^2 - \|\lambda\|^2\right)\}. \end{array} \tag{5}$$

It shall be mentioned that the standard augmented Lagrangian function associated with the lower-level optimization of (1) is $\tilde{f}(x, z) + \frac{1}{2\mu}\left(\|[\lambda + \mu\tilde{g}(x, z)]_+\|^2 - \|\lambda\|^2\right)$. Therefore, strictly speaking, $\tilde{f}(x, z) + \frac{1}{2\rho\mu}\left(\|[\lambda + \mu\tilde{g}(x, z)]_+\|^2 - \|\lambda\|^2\right)$ is not an augmented Lagrangian function associated with it, albeit being motivated from an augmented Lagrangian perspective.

Further, by applying a penalty scheme, problem (5) can be approximated by

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

which is equivalent to the minimax problem

$$\min_{x,y} \max_z \mathcal{L}(x, y, z, \lambda; \rho, \mu), \tag{6}$$

where $\mathcal{L}$ is defined as

$$\mathcal{L}(x,y,z,\lambda;\rho,\mu) = f(x,y) + \rho\tilde{f}(x,y) + \frac{1}{2\mu}\|[\lambda+\mu\tilde{g}(x,y)]_+\|^2 - \rho\tilde{f}(x,z) - \frac{1}{2\mu}\|[\lambda+\mu\tilde{g}(x,z)]_+\|^2. \tag{7}$$

In view of Assumption 1, one can observe that $\mathcal{L}$ enjoys the following nice structure.

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

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

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

Thanks to the above nice structure of $\mathcal{L}$, an approximate primal-dual stationary point of problem (6) can be suitably found by Algorithm 4 (see Appendix B). Based on this observation, we propose to solve problem (1) by solving a sequence of minimax subproblems in the form of (6).

To present our method, we define

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

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

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

```
Input: $\varepsilon,\tau\in(0,1)$, $\epsilon_0\in(\tau\varepsilon,1]$, $x^0\in\mathcal{X}$, $z^0\in\mathcal{Y}$, $\epsilon_k=\epsilon_0\tau^k$, $\rho_k=\epsilon_k^{-1}$, $\mu_k=\epsilon_k^{-3}$, $\eta_k=\epsilon_k$ and $\lambda^0\in\mathbb{R}^l_+$.
1: for $k=0,1\ldots$ do
2:     Call Algorithm 2 (see Appendix A) with $\Psi(\cdot)\leftarrow\widetilde{\mathcal{L}}(x^k,\cdot,\lambda^k;\rho_k,\mu_k)$, $\tilde{\epsilon}\leftarrow\eta_k$, $L_{\nabla\phi}\leftarrow\widetilde{L}_k$, $\tilde{x}^0\leftarrow y^k$
       to find an approximate solution $y^k_{\rm init}$ of $\min_z\widetilde{L}(x^k,z,\lambda^k;\rho_k,\mu_k)$ such that
           $\widetilde{\mathcal{L}}(x^k,y^k_{\rm init},\lambda^k;\rho_k,\mu_k)-\min_z\widetilde{\mathcal{L}}(x^k,z,\lambda^k;\rho_k,\mu_k)\le\eta_k,$    (9)
       where $\widetilde{\mathcal{L}}$ is given in (8) and
           $\widetilde{L}_k=L_{\nabla\tilde{f}_1}+\rho_k^{-1}(\mu_kL_{\tilde{g}}^2+\mu_k\tilde{g}_{\rm hi}L_{\nabla\tilde{g}}+\|\lambda^k\|L_{\nabla\tilde{g}}).$    (10)
3:     Call Algorithm 4 (see Appendix B) with $\epsilon\leftarrow\epsilon_k$, $\hat{\epsilon}_0\leftarrow\epsilon_k/(2\sqrt{\mu_k})$, $\hat{x}^0\leftarrow(x^k,y^k_{\rm init})$, $\hat{y}^0\leftarrow z^k$ and
       $L_{\nabla h}\leftarrow L_k$ to find an $\epsilon_k$-primal-dual stationary point $(x^{k+1},y^{k+1},z^{k+1})$ of
           $\min_{x,y}\max_z\mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k),$    (11)
       where
           $L_k=L_{\nabla f_1}+2\rho_kL_{\nabla\tilde{f}_1}+2\mu_kL_{\tilde{g}}^2+2\mu_k\tilde{g}_{\rm hi}L_{\nabla\tilde{g}}+2\|\lambda^k\|L_{\nabla\tilde{g}}.$    (12)
4:     Set $\lambda^{k+1}=[\lambda^k+\mu_k\tilde{g}(x^{k+1},z^{k+1})]_+$.
5:     If $\epsilon_k\le\varepsilon$, terminate the algorithm and output $(x^{k+1},y^{k+1})$.
6: end for
```

Remark 1. *(i) Notice that* $\widetilde{L}(x^k,y,\lambda^k;\rho_k,\mu_k)=\phi(y)+\tilde{f}_2(y)$ *with* $\phi(y)=\tilde{f}_1(x^k,y)+\|[\lambda^k+\mu_k\tilde{g}(x^k,y)]_+\|^2/(2\rho_k\mu_k)$. *By Assumption 1 and* (27)*, one can see that* ϕ *is* $\widetilde{L}_k$*-smooth on* $\operatorname{dom}P$ *and the proximal operator of* $\tilde{f}_2$ *can be exactly evaluated. It then follows from this and Theorem 2 (see Appendix A) that* $y^k_{\rm init}$ *satisfying* (9) *can be successfully found in step 2 of Algorithm 1 by applying Algorithm 2 to the problem* $\min_z\widetilde{L}(x^k,z,\lambda^k;\rho_k,\mu_k)$.

(ii) In view of Theorem 3 (see Appendix B), one can see that an ϵ_k*-primal-dual stationary point of* (11) *can be successfully found in step 3 of Algorithm 1 by applying Algorithm 4 to problem* (11)*. Consequently, Algorithm 1 is well-defined.*

2.1 Complexity results for Algorithm 1

In this subsection we study *iteration and operation complexity* for Algorithm 1. In particular, in order to characterize the approximate solution found by Algorithm 1, we first introduce a terminology called an ε-KKT solution of problem (1). Then we establish iteration and operation complexity of Algorithm 1 for finding an $\mathcal{O}(\varepsilon)$-KKT solution of (1).

For notational convenience, we define

$$\tilde{f}^*(x) := \min\{\tilde{f}(x,z)|\tilde{g}(x,z) \le 0\}. \tag{13}$$

Observe that problem (1) can be equivalently reformulated as

$$\min_{x,y}\{f(x,y)|\tilde{f}(x,y) \le \tilde{f}^*(x),\ \tilde{g}(x,y) \le 0\}. \tag{14}$$

The Lagrangian function associated with (14) is given by

$$\widehat{\mathcal{L}}(x,y,\rho,\lambda_{\mathbf{y}}) = f(x,y) + \rho(\tilde{f}(x,y) - \tilde{f}^*(x)) + \langle \lambda_{\mathbf{y}}, \tilde{g}(x,y)\rangle. \tag{15}$$

In the same spirit of classical constrained optimization, one would naturally be interested in a KKT solution (x,y) of (14), namely, (x,y) satisfies

$$\tilde{f}(x,y) \le \tilde{f}^*(x), \quad \tilde{g}(x,y) \le 0, \quad \rho(\tilde{f}(x,y) - \tilde{f}^*(x)) = 0, \quad \langle \lambda_{\mathbf{y}}, \tilde{g}(x,y)\rangle = 0, \tag{16}$$

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

$$\min_{x',y'} \widehat{\mathcal{L}}(x',y',\rho,\lambda_{\mathbf{y}}) \tag{17}$$

for some $\rho \ge 0$ and $\lambda_{\mathbf{y}} \in \mathbb{R}^l_+$. Yet, due to the sophisticated problem structure, characterizing a stationary point of (17) is generally difficult. On another hand, notice from Lemma 1 and (15) that problem (17) is equivalent to the minimax problem

$$\min_{x',y',\lambda'_{\mathbf{z}}} \max_{z'} \left\{ f(x',y') + \rho\big(\tilde{f}(x',y') - \tilde{f}(x',z') - \langle \lambda'_{\mathbf{z}}, \tilde{g}(x',z')\rangle\big) + \langle \lambda_{\mathbf{y}}, \tilde{g}(x',y')\rangle + \mathscr{I}_{\mathbb{R}^l_+}(\lambda'_{\mathbf{z}}) \right\},$$ [3]

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

$$0 \in \partial f(x,y) + \rho\partial\tilde{f}(x,y) - \rho(\nabla_x\tilde{f}(x,z) + \nabla_x\tilde{g}(x,z)\lambda_{\mathbf{z}};0) + \nabla\tilde{g}(x,y)\lambda_{\mathbf{y}}, \tag{18}$$

$$0 \in \rho(\partial_z\tilde{f}(x,z) + \nabla_z\tilde{g}(x,z)\lambda_{\mathbf{z}}), \tag{19}$$

$$\lambda_{\mathbf{z}} \in \mathbb{R}^l_+, \quad \tilde{g}(x,z) \le 0, \quad \langle \lambda_{\mathbf{z}}, \tilde{g}(x,z)\rangle = 0. \tag{20}$$ [4]

Based on this observation, the equivalence of (1) and (14), and also the fact that (16) is equivalent to

$$\tilde{f}(x,y) = \tilde{f}^*(x), \quad \tilde{g}(x,y) \le 0, \quad \langle \lambda_{\mathbf{y}}, \tilde{g}(x,y)\rangle = 0, \tag{21}$$

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

Definition 2. *The pair (x,y) is said to be a KKT solution of problem* (1) *if there exists $(z,\rho,\lambda_{\mathbf{y}},\lambda_{\mathbf{z}}) \in \mathbb{R}^m \times \mathbb{R}_+ \times \mathbb{R}^l_+ \times \mathbb{R}^l_+$ such that* (18)*-*(21) *hold. In addition, for any $\varepsilon > 0$, (x,y) is said to be an ε-KKT solution of problem* (1) *if there exists $(z,\rho,\lambda_{\mathbf{y}},\lambda_{\mathbf{z}}) \in \mathbb{R}^m \times \mathbb{R}_+ \times \mathbb{R}^l_+ \times \mathbb{R}^l_+$ such that*

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

where $\tilde{f}^*$ is defined in (13).

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

[4] The relations in (20) are equivalent to $0 \in -\tilde{g}(x,z) + \partial\mathscr{I}_{\mathbb{R}^l_+}(\lambda_{\mathbf{z}})$.

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

$$\tilde{f}^*_{\mathrm{hi}} := \sup\{\tilde{f}^*(x) | x \in \mathcal{X}\}, \tag{22}$$

$$D_{\mathbf{x}} := \max\{\|u - v\| | u, v \in \mathcal{X}\}, \quad D_{\mathbf{y}} := \max\{\|u - v\| | u, v \in \mathcal{Y}\}, \tag{23}$$

$$f_{\mathrm{hi}} := \max\{f(x, y) | (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \quad f_{\mathrm{low}} := \min\{f(x, y) | (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \tag{24}$$

$$\tilde{f}_{\mathrm{low}} := \min\{\tilde{f}(x, z) | (x, z) \in \mathcal{X} \times \mathcal{Y}\}, \quad \tilde{g}_{\mathrm{hi}} := \max\{\|\tilde{g}(x, y)\| | (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \tag{25}$$

$$K := \lceil (\log \varepsilon - \log \epsilon_0)/\log \tau \rceil_+, \quad \mathbb{K} := \{0, 1, \ldots, K + 1\}, \quad \mathbb{K} - 1 = \{k - 1 | k \in \mathbb{K}\}. \tag{26}$$

It then follows from Assumption 1(iii) that

$$\|\nabla \tilde{g}(x, y)\| \leq L_{\tilde{g}} \qquad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{27}$$

In addition, by Assumption 1 and the compactness of $\mathcal{X}$ and $\mathcal{Y}$, one can observe that $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, f_{hi}, f_{low}, $\tilde{f}_{\mathrm{low}}$ and $\tilde{g}_{\mathrm{hi}}$ are finite. Besides, as will be subsequently shown in Lemma 1(ii), $\tilde{f}^*_{\mathrm{hi}}$ is finite.

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

Assumption 2 (Slater's condition). *There exists $\hat{z}_x \in \mathcal{Y}$ for each $x \in \mathcal{X}$ such that $\tilde{g}_i(x, \hat{z}_x) < 0$ for all $i = 1, 2, \ldots, l$ and $G := \inf\{-\tilde{g}_i(x, \hat{z}_x) | x \in \mathcal{X},\ i = 1, \ldots, l\} > 0$.* [5]

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

Theorem 1 (iteration and operation complexity of Algorithm 1). *Suppose that Assumptions 1 and 2 hold. Let $\{(x_k, y_k, z_k, \lambda^k)\}_{k \in \mathbb{K}}$ be generated by Algorithm 1, f^*, $\tilde{f}^*_{\mathrm{hi}}$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, f_{hi}, f_{low}, $\tilde{f}_{\mathrm{low}}$, $\tilde{g}_{\mathrm{hi}}$ and K be defined in (1), (22), (23), (24), (25) and (26), L_f, $L_{\tilde{f}}$, $L_{\nabla f_1}$, $L_{\nabla \tilde{f}_1}$, $L_{\tilde{g}}$, $L_{\nabla \tilde{g}}$ and G be given in Assumptions 1 and 2, and ε, ϵ_0, τ, μ_K, ρ_K and λ_0 be given in Algorithm 1. Let*

$$\vartheta = \frac{1}{2}\|\lambda^0\|^2 + \frac{\tilde{f}^*_{\mathrm{hi}} - \tilde{f}_{\mathrm{low}}}{1 - \tau^4} + \frac{D_{\mathbf{y}}\epsilon_0}{1 - \tau^3}, \tag{28}$$

$$L = L_{\nabla f_1} + 2L_{\nabla \tilde{f}_1} + 2L_{\tilde{g}}^2 + 2\tilde{g}_{\mathrm{hi}} L_{\nabla \tilde{g}} + 2\sqrt{2\vartheta} L_{\nabla \tilde{g}}, \quad \widetilde{L} = L_{\nabla \tilde{f}_1} + L_{\tilde{g}}^2 + \tilde{g}_{\mathrm{hi}} L_{\nabla \tilde{g}} + \sqrt{2\vartheta} L_{\nabla \tilde{g}}, \tag{29}$$

$$\alpha = \min\left\{1, \sqrt{4/(D_{\mathbf{y}} L)}\right\}, \quad \delta = (2 + \alpha^{-1}) L D_{\mathbf{x}}^2 + \max\{1/D_{\mathbf{y}}, L/4\} D_{\mathbf{y}}^2, \tag{30}$$

$$\begin{aligned} M = {} & 16 \max\left\{1/(4L_{\tilde{g}}^2), 2/(\alpha L_{\tilde{g}}^2)\right\} \left[(3L + 1/(2D_{\mathbf{y}}))^2 / \min\{2L_{\tilde{g}}^2, 1/(2D_{\mathbf{y}})\} + 3L + 1/(2D_{\mathbf{y}})\right]^2 \\ & \times \left(\delta + 2\alpha^{-1}\left(f^* - f_{\mathrm{low}} + \tilde{f}^*_{\mathrm{hi}} - \tilde{f}_{\mathrm{low}} + L_{\tilde{f}} D_{\mathbf{y}} + 3\vartheta + \tilde{g}_{\mathrm{hi}}^2 + D_{\mathbf{y}}/4 + L D_{\mathbf{x}}^2\right)\right), \end{aligned} \tag{31}$$

$$T = \left\lceil 16 \left(f_{\mathrm{hi}} - f_{\mathrm{low}} + 1 + D_{\mathbf{y}}/4\right) L + 8(1 + 4D_{\mathbf{y}}^2 L^2) \right\rceil_+, \tag{32}$$

$$\lambda_{\mathbf{y}}^{K+1} = [\lambda^K + \mu_K \tilde{g}(x^{K+1}, y^{K+1})]_+, \quad \lambda_{\mathbf{z}}^{K+1} = \rho_K^{-1} [\lambda^K + \mu_K \tilde{g}(x^{K+1}, z^{K+1})]_+. \tag{33}$$

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

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

$$\begin{aligned} & \mathrm{dist}\Big(0, \partial f(x^{K+1}, y^{K+1}) + \rho_K \partial \tilde{f}(x^{K+1}, y^{K+1}) - \rho_K (\nabla_x \tilde{f}(x^{K+1}, z^{K+1}) + \nabla_x \tilde{g}(x^{K+1}, z^{K+1}) \lambda_{\mathbf{z}}^{K+1}; 0) \\ & \qquad + \nabla \tilde{g}(x^{K+1}, y^{K+1}) \lambda_{\mathbf{y}}^{K+1}\Big) \leq \varepsilon, \end{aligned} \tag{34}$$

$$\mathrm{dist}\left(0, \rho_K (\partial_z \tilde{f}(x^{K+1}, z^{K+1}) + \nabla_z \tilde{g}(x^{K+1}, z^{K+1}) \lambda_{\mathbf{z}}^{K+1})\right) \leq \varepsilon, \tag{35}$$

$$\|[\tilde{g}(x^{K+1}, z^{K+1})]_+\| \leq 2\varepsilon^2 G^{-1} (\epsilon_0 + L_{\tilde{f}}) D_{\mathbf{y}}, \tag{36}$$

$$|\langle \lambda_{\mathbf{z}}^{K+1}, \tilde{g}(x^{K+1}, z^{K+1}) \rangle| \leq 2\varepsilon^2 G^{-1} (\epsilon_0 + L_{\tilde{f}}) D_{\mathbf{y}} \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + L_{\tilde{f}}) D_{\mathbf{y}}\}, \tag{37}$$

$$\|[\tilde{g}(x^{K+1}, y^{K+1})]_+\| \leq 2\varepsilon^2 G^{-1} (\epsilon_0 + L_f + L_{\tilde{f}}) D_{\mathbf{y}}, \tag{38}$$

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

$$
\begin{aligned}
&|\langle \lambda_{\mathbf{y}}^{K+1}, \tilde g(x^{K+1}, z^{K+1})\rangle| \le 2\varepsilon G^{-1}(\epsilon_0 + L_f + L_{\tilde f}) D_{\mathbf{y}} \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + L_f + L_{\tilde f}) D_{\mathbf{y}}\}, \qquad (39)\\
&|\tilde f(x^{K+1}, y^{K+1}) - \tilde f^*(x^{K+1})| \le \max\Big\{2\varepsilon^2 G^{-2} L_{\tilde f}(\epsilon_0 + L_f + L_{\tilde f}) D_{\mathbf{y}}^2,\ \varepsilon^3 \max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + L_{\tilde f}) D_{\mathbf{y}}\}/2\\
&\qquad\qquad + \varepsilon\left(f_{\mathrm{hi}} - f_{\mathrm{low}} + 1 + D_{\mathbf{y}}/4 + L_{\tilde g}^{-2}/4 + 2D_{\mathbf{y}}^2 L\right)\Big\}. \qquad (40)
\end{aligned}
$$

(ii) The total number of evaluations of ∇f_1, $\nabla \tilde f_1$, $\nabla \tilde g$ and proximal operator of f_2 and $\tilde f_2$ performed in Algorithm 1 is N, respectively, where

$$
\begin{aligned}
N = &\left(\left\lceil 96\sqrt{2}\left(1 + (12L + 2/D_y)/L_{\tilde g}^2\right)\right\rceil + 2\right) \max\left\{2, \sqrt{D_y L}\right\} T(1-\tau^7)^{-1}\\
&\times (\tau\varepsilon)^{-7}\left(56K\log(1/\tau) + 56\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right)\\
&+ (\tau\varepsilon)^{-3/2}(1-\tau^{3/2})^{-1} D_{\mathbf{y}}\sqrt{2\widetilde L} + K. \qquad (41)
\end{aligned}
$$

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

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

where $\tilde f^*$ is defined in (13)*, $\rho_K = (\epsilon_0\tau^K)^{-1}$, and $\lambda_{\mathbf{y}}^{K+1}, \lambda_{\mathbf{z}}^{K+1} \in \mathbb{R}_+^l$ are given in* (33)*.*

3 Numerical results

In this section we conduct some preliminary experiments to test the performance of our SMO method (namely, Algorithms 1), and compare it with a first-order penalty method (FPM) [39]. Both algorithms are coded in Matlab and all the computations are performed on a laptop with a 2.30 GHz Intel i9-9880H 8-core processor and 16 GB of RAM.

3.1 Constrained bilevel linear optimization

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

$$
\begin{aligned}
\min \quad & c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x)\\
\text{s.t.} \quad & y \in \operatorname*{Argmin}_z \left\{\tilde d^T z + \mathscr{I}_{[-1,1]^m}(z) \middle| \widetilde A x + \widetilde B z - \tilde b \le 0\right\},
\end{aligned} \qquad (42)
$$

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

For each triple (n, m, l), we randomly generate 10 instances of problem (42). Specifically, we first randomly generate c and d with all the entries independently chosen from the standard normal distribution. We then randomly generate $\widetilde A$ and $\widetilde B$ with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.01. In addition, we randomly generate $\hat y \in [-1,1]^m$ with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.1 and then projected to $[-1,1]^m$ and choose $\tilde d$ and $\tilde b$ such that $\hat y$ is an optimal solution of the lower level optimization of (42) with $x = 0$.

Notice that (42) is a special case of (1) with $f(x,y) = c^T x + d^T y + \mathscr{I}_{[-1,1]^n}(x)$, $\tilde f(x,z) = \tilde d^T z + \mathscr{I}_{[-1,1]^m}(z)$ and $\tilde g(x,z) = \widetilde A x + \widetilde B z - \tilde b$. We now apply SMO and FPM to solve (42). Specifically, we choose 0 as the initial point and set the parameters of SMO as $(\varepsilon, \epsilon_0, \tau) = (10^{-2}, 1, 0.8)$ and use a variant of FPM with dynamic update on penalty and tolerance parameters for the sake of efficiency. Specifically, we set $\rho_k = 5^{k-1}$, $\varepsilon_k = \rho_k^{-1}$ and $x_{-1} = 0$ for [39, Algorithm 2]. For each $k \geq 0$, let (x^k, y^k) be the output of [39, Algorithm 2], and we run [39, Algorithm 2] with $(\varepsilon, \rho) = (\varepsilon_k, \rho_k)$ and $(x^{k-1}, \tilde{y}^{k-1})$ as the initial point to generate (x^k, y^k), where $\tilde{y}^{k-1} \in \mathrm{Argmin}_z \tilde{f}(x^{k-1}, z)$ is found by CVX [18]. We terminate both algorithms once $\epsilon_k \leq 10^{-2}$ (or $\varepsilon_k \leq 10^{-2}$ for FPM) and (x^k, y^k) satisfies

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

for some k and output (x^k, y^k) as an approximate solution of (42), where $\tilde{f}^*$ is defined in (13) and the value $\tilde{f}^*(x^k)$ is computed by CVX [18].

The computational results of SMO and FPM for problem (42) with the instances randomly generated above are presented in Table 1. In detail, the values of n, m and l are listed in the first three columns. For each triple (n, m, l), the average initial objective value $f(x^0, \hat{y})$ with $\hat{y}$ being generated above,[6] and the average final objective value $f(x^k, y^k)$ and the average CPU time (in seconds) over 10 random instances are given in the rest of the columns. One can observe that both SMO and FPM found an approximate solution with much lower objective value than the initial objective value. Moreover, SMO outputs an approximate solution with lower objective value than FPM, while SMO significantly outperforms FPM in terms of average CPU time.

			Initial objective value	Final objective value		CPU time (seconds)	
n	m	l		SMO	FPM	SMO	FPM
100	100	5	0.22	-77.51	-77.32	6.47	68.80
200	200	10	-0.38	-153.43	-153.09	13.63	140.22
300	300	15	-0.11	-249.92	-246.71	18.76	213.45
400	400	20	0.34	-307.83	-307.54	28.74	321.81
500	500	25	0.96	-396.68	-396.29	52.30	710.53

Table 1: Numerical results for problem (42)

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

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

$$\begin{aligned} \min \quad & x^TAx + x^TBy + y^TCy + c^Tx + d^Ty + \mathscr{I}_{[-1,1]^n}(x) \\ \text{s.t.} \quad & y \in \underset{z}{\mathrm{Argmin}} \left\{ \tilde{d}^Tz + \mathscr{I}_{[-1,1]^m}(z) \middle| \widetilde{A}x + \widetilde{B}z - \tilde{b} \leq 0 \right\}, \end{aligned} \tag{43}$$

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

For each triple (n, m, l), we randomly generate 10 instances of problem (43). Specifically, we first randomly generate A, B, C, c and d with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.1. We then randomly generate $\widetilde{A}$ and $\widetilde{B}$ with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.01. In addition, we randomly generate $\hat{y} \in [-1,1]^m$ with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.1 and then projected to $[-1,1]^m$ and choose $\tilde{d}$ and $\tilde{b}$ such that $\hat{y}$ is an optimal solution of the lower level optimization of (43) with $x = 0$.

Notice that (43) is a special case of (1) with $f(x,y) = x^TAx + x^TBy + y^TCy + c^Tx + d^Ty + \mathscr{I}_{[-1,1]^n}(x)$, $\tilde{f}(x,z) = \tilde{d}^Tz + \mathscr{I}_{[-1,1]^m}(z)$ and $\tilde{g}(x,z) = \widetilde{A}x + \widetilde{B}z - \tilde{b}$. We now apply our SMO method (namely, Algorithms 1) to solve (43).[7] Specifically, we choose 0 as the initial point and set the parameters of SMO as $(\varepsilon, \epsilon_0, \tau) = (10^{-2}, 1, 0.8)$. We terminate SMO once $\epsilon_k \leq 10^{-2}$ and (x^k, y^k) satisfies

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

[6] Note that $(x^0, y^0_{\mathrm{init}})$ may not be a feasible point of (42). Nevertheless, $(x^0, \hat{y})$ is a feasible point of (42) due to $x^0 = 0$ and the particular way for generating instances of (42). Besides, (42) can be viewed as an implicit optimization problem in terms of the variable x. It is thus reasonable to use $f(x^0, \hat{y})$ as the initial objective value for the purpose of comparison.

[7] Clearly, problem (43) is more sophisticated than (42). As seen from Table 1, problem (42) is challenging to FPM [39] when the dimension n is relatively large. Consequently, we will not apply FPM to solve (43).

for some k and output (x^k, y^k) as an approximate solution of (43), where $\tilde{f}^*$ is defined in (13) and the value $\tilde{f}^*(x^k)$ is computed by CVX [18].

The computational results of SMO for problem (43) with the instances randomly generated above are presented in Table 2. In detail, the values of n, m and l are listed in the first three columns. For each triple (n, m, l), the average initial objective value $f(x^0, \hat{y})$ with $\hat{y}$ being generated above[8] and the average final objective value $f(x^k, y^k)$ over 10 random instances are given in the rest of the columns. One can see that the approximate solution (x^k, y^k) found by SMO significantly reduces the initial objective value.

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

Table 2: Numerical results for problem (43)

4 Proof of main results

In this section we provide a proof of our main result presented in Subsection 2.1, which is particularly Theorem 1. Before proceeding, one can observe from (8) and (13) that

$$\min_z \widetilde{\mathcal{L}}(x, z, \lambda; \rho, \mu) \le \tilde{f}^*(x) + \frac{\|\lambda\|^2}{2\rho\mu} \qquad \forall x \in \mathcal{X}, \lambda \in \mathbb{R}^l_+, \rho, \mu > 0, \tag{44}$$

which will be frequently used later.

We next establish several technical lemmas that will be used to prove Theorem 1 subsequently.

Lemma 1. *Suppose that Assumptions 1 and 2 hold. Let $\tilde{f}^*$, $\tilde{f}^*_{\rm hi}$, $D_{\mathbf{y}}$, $L_{\tilde{f}}$ and G be given in* (13)*,* (22)*,* (23)*, and Assumptions 1 and 2, respectively. Then the following statements hold.*

(i) $\lambda^* \ge 0$ and $\|\lambda^*\| \le G^{-1} L_{\tilde{f}} D_{\mathbf{y}}$ for all $\lambda^* \in \Lambda^*(x)$ and $x \in \mathcal{X}$, where $\Lambda^*(x)$ denotes the set of optimal Lagrangian multipliers of problem (13) *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

$$\tilde{f}^*(x) = \max_\lambda \min_z \tilde{f}(x, z) + \langle \lambda, \tilde{g}(x, z) \rangle - \mathscr{I}_{\mathbb{R}^l_+}(\lambda) \qquad \forall x \in \mathcal{X}, \tag{45}$$

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 (13). 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 1(iv). 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 $\lambda^* \ge 0$, (23) and Assumption 1 implies that

$$G \sum_{i=1}^l \lambda_i^* \le \langle \lambda^*, -\tilde{g}(x, \hat{z}_x) \rangle \le \tilde{f}(x, \hat{z}_x) - \tilde{f}(x, z^*) \le L_{\tilde{f}} \|\hat{z}_x - z^*\| \le L_{\tilde{f}} D_{\mathbf{y}}, \tag{46}$$

[8] Note that $(x^0, y^0_{\rm init})$ may not be a feasible point of (43). Nevertheless, $(x^0, \hat{y})$ is a feasible point of (43) due to $x^0 = 0$ and the particular way for generating instances of (43). Besides, (43) can be viewed as an implicit optimization problem in terms of the variable x. It is thus reasonable to use $f(x^0, \hat{y})$ as the initial objective value for the purpose of comparison.

where the first inequality is due to Assumption 1(iv), and the third inequality follows from (23) and $L_{\tilde{f}}$-Lipschitz continuity of $\tilde{f}$ (see Assumption 1(i)). It then follows from (71) that $\|\lambda^*\| \le \sum_{i=1}^{l} \lambda_i^* \le G^{-1} L_{\tilde{f}} D_{\mathbf{y}}$.

(ii) Recall from Assumptions 1(i) and 1(iv) that $\tilde{f}(x,\cdot)$ and $\tilde{g}_i(x,\cdot)$, $i=1,\ldots,l$, are convex for any given $x \in \mathcal{X}$. Using this, (13) and the first statement of this lemma, we observe that

$$\tilde{f}^*(x) = \min_{z} \max_{\lambda\in\mathbb{B}} \tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle \qquad \forall x \in \mathcal{X}, \tag{47}$$

where

$$\mathbb{B} := \{\lambda \in \mathbb{R}^l_+ : \|\lambda\| \le G^{-1} L_{\tilde{f}} D_{\mathbf{y}}\}.$$

Notice from Assumption 1(i) that $\tilde{f}$ and $\tilde{g}$ are Lipschitz continuous on their domain. Then it is not hard to observe that $\max\{\tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle | \lambda \in \mathbb{B}\}$ is a Lipschitz continuous function of (x,z) on its domain. By this and (47), one can easily verify that $\tilde{f}^*$ is Lipschitz continuous on $\mathcal{X}$. In addition, the finiteness of $\tilde{f}^*_{\mathrm{hi}}$ follows from (22), the continuity of $\tilde{f}^*$, and the compactness of $\mathcal{X}$.

(iii) It follows from Assumption 1 that the domain of $\tilde{f}(x,\cdot)$ is compact for all $x \in \mathcal{X}$. By this, (47) and the strong duality, one has

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

and hence

$$\tilde{f}^*(x) \le \max_{\lambda} \min_{z} \tilde{f}(x,z) + \langle \lambda, \tilde{g}(x,z)\rangle - \mathscr{I}_{\mathbb{R}^l_+}(\lambda) \qquad \forall x \in \mathcal{X}. \tag{48}$$

In addition, one can observe from (13) 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. This inequality together with (48) implies that (45) holds. □

Lemma 2. *Suppose that Assumption 1 holds. Let $\mathbb{K}$ and ϑ be defined in* (26) *and* (28)*, μ_k and ρ_k be given in Algorithm 1, and $\{\lambda^k\}_{k\in\mathbb{K}}$ be generated by Algorithm 1. Then we have*

$$\|\lambda^k\|^2 \le 2\rho_k \mu_k \vartheta \qquad \forall 0 \le k \in \mathbb{K} - 1. \tag{49}$$

Proof. One can observe from (22), (25) and Algorithm 1 that $\tilde{f}^*_{\mathrm{hi}} \ge \tilde{f}_{\mathrm{low}}$ and $\mu_0 \ge \rho_0 \ge 1 > \tau > 0$, which together with (28) imply that (49) holds for $k=0$. It remains to show that (49) holds for all $1 \le k \in \mathbb{K} - 1$.

Since $(x^{t+1}, y^{t+1}, z^{t+1})$ is an ϵ_t-primal-dual stationary point of (11) for all $0 \le t \in \mathbb{K} - 1$, it follows from Definition 1 that there exists some $u \in \partial_z \mathcal{L}(x^{t+1}, y^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t)$ with $\|u\| \le \epsilon_t$. Notice from (7) and (8) that $\partial_z \mathcal{L}(x^{t+1}, y^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t) = -\rho_t \partial_z \widetilde{\mathcal{L}}(x^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t)$. Hence, $-\rho_t^{-1} u \in \partial_z \widetilde{\mathcal{L}}(x^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t)$. Also, observe from (8) and Assumption 1 that $\widetilde{\mathcal{L}}(x^{t+1}, \cdot, \lambda^t; \rho_t, \mu_t)$ is convex. Using this, (23), $-\rho_t^{-1} u \in \partial_z \widetilde{\mathcal{L}}(x^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t)$ and $\|u\| \le \epsilon_t$, we obtain

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

which implies that

$$\min_{z} \widetilde{\mathcal{L}}(x^{t+1}, z, \lambda^t; \rho_t, \mu_t) \ge \widetilde{\mathcal{L}}(x^{t+1}, z^{t+1}, \lambda^t; \rho_t, \mu_t) - \rho_t^{-1} D_{\mathbf{y}} \epsilon_t. \tag{50}$$

By this, (8) and (44), one has

$$\begin{aligned}\tilde{f}^*(x^{t+1}) &\overset{(44)}{\ge} \min_{z} \widetilde{\mathcal{L}}(x^{t+1}, z, \lambda^t; \rho_t, \mu_t) - \frac{\|\lambda^t\|^2}{2\rho_t \mu_t} \\ &\overset{(8)(50)}{\ge} \tilde{f}(x^{t+1}, z^{t+1}) + \frac{1}{2\rho_t \mu_t}\left(\|[\lambda^t + \mu_t \tilde{g}(x^{t+1}, z^{t+1})]_+\|^2 - \|\lambda^t\|^2\right) - \rho_t^{-1} D_{\mathbf{y}} \epsilon_t \\ &= \tilde{f}(x^{t+1}, z^{t+1}) + \frac{1}{2\rho_t \mu_t}\left(\|\lambda^{t+1}\|^2 - \|\lambda^t\|^2\right) - \rho_t^{-1} D_{\mathbf{y}} \epsilon_t,\end{aligned}$$

where the equality follows from the relation $\lambda^{t+1} = [\lambda^t + \mu_t \tilde{g}(x^{t+1}, z^{t+1})]_+$ (see Algorithm 1). Using this inequality, (22), (25) and $\epsilon_t \le \epsilon_0$ (see Algorithm 1), we have

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

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

$$\|\lambda^k\|^2 \le \|\lambda^0\|^2 + 2(\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}}) \sum_{t=0}^{k-1} \rho_t\mu_t + 2D_{\mathbf{y}}\epsilon_0 \sum_{t=0}^{k-1} \mu_t. \tag{51}$$

Recall from Algorithm 1 that $\epsilon_t = \epsilon_0\tau^t$, $\mu_t = \epsilon_t^{-3}$ and $\rho_t = \epsilon_t^{-1}$. It is not hard to verify that $\sum_{t=0}^{k-1} \rho_t\mu_t \le \rho_{k-1}\mu_{k-1}/(1-\tau^4)$ and $\sum_{t=0}^{k-1} \mu_t \le \mu_{k-1}/(1-\tau^3)$. Using these, (51), $\rho_k > \rho_{k-1} \ge 1$ and $\mu_k > \mu_{k-1} \ge 1$ (see Algorithm 1), we obtain that for all $1 \le k \in \mathbb{K} - 1$,

$$\begin{aligned}
\rho_k^{-1}\mu_k^{-1}\|\lambda^k\|^2 &\le \rho_k^{-1}\mu_k^{-1}\left(\|\lambda^0\|^2 + \frac{2\rho_{k-1}\mu_{k-1}(\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}})}{1-\tau^4} + \frac{2\mu_{k-1}D_{\mathbf{y}}\epsilon_0}{1-\tau^3}\right) \\
&\le \|\lambda^0\|^2 + \frac{2(\tilde{f}^*_{\text{hi}} - \tilde{f}_{\text{low}})}{1-\tau^4} + \frac{2D_{\mathbf{y}}\epsilon_0}{1-\tau^3} \overset{(28)}{=} 2\vartheta.
\end{aligned}$$

It implies that the conclusion of this lemma holds. □

Lemma 3. *Suppose that Assumptions 1 and 2 hold. Let $D_{\mathbf{y}}$, $\mathbb{K}$ and ϑ be defined in* (23), (26) *and* (28), *L_f, $L_{\tilde{f}}$ and G be given in Assumptions 1 and 2, and ϵ_0, ρ_k and μ_k be given in Algorithm 1. Suppose that $(x^{k+1}, y^{k+1}, z^{k+1}, \lambda^{k+1})$ is generated by Algorithm 1 for some $0 \le k \in \mathbb{K} - 1$ with*

$$\rho_k^{-1}\mu_k \ge 8G^{-2}\vartheta. \tag{52}$$

Then we have

$$\|[\tilde{g}(x^{k+1}, z^{k+1})]_+\| \le \mu_k^{-1}\|\lambda^{k+1}\| \le 2\mu_k^{-1}G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}, \tag{53}$$

$$\|[\tilde{g}(x^{k+1}, y^{k+1})]_+\| \le \mu_k^{-1}\|[\lambda^k + \mu_k\tilde{g}(x^{k+1}, y^{k+1})]_+\| \le 2\mu_k^{-1}G^{-1}(\epsilon_0 + L_f + \rho_k L_{\tilde{f}})D_{\mathbf{y}}. \tag{54}$$

Proof. Suppose that $(x^{k+1}, y^{k+1}, z^{k+1}, \lambda^{k+1})$ is generated by Algorithm 1 for some $0 \le k \in \mathbb{K} - 1$ satisfying (52). Notice that $(x^{k+1}, y^{k+1}, z^{k+1})$ is an ϵ_k-primal-dual stationary point of (11). It then follows from (7), Definition 1 and Assumption 1 that

$$\text{dist}\left(0, \nabla_y f(x^{k+1}, y^{k+1}) + \rho_k\partial_y\tilde{f}(x^{k+1}, y^{k+1}) + \nabla_y\tilde{g}(x^{k+1}, y^{k+1})[\lambda^k + \mu_k\tilde{g}(x^{k+1}, y^{k+1})]_+\right) \le \epsilon_k, \tag{55}$$

$$\text{dist}\left(0, -\rho_k\partial_z\tilde{f}(x^{k+1}, z^{k+1}) - \nabla_z\tilde{g}(x^{k+1}, z^{k+1})[\lambda^k + \mu_k\tilde{g}(x^{k+1}, z^{k+1})]_+\right) \le \epsilon_k. \tag{56}$$

We first show that (53) holds. Notice from Algorithm 1 that $\lambda^{k+1} = [\lambda^k + \mu_k\tilde{g}(x^{k+1}, z^{k+1})]_+$. Hence, it follows from (56) that there exists some $u \in \partial_z\tilde{f}(x^{k+1}, z^{k+1})$ such that

$$\|\rho_k u + \nabla_z\tilde{g}(x^{k+1}, z^{k+1})\lambda^{k+1}\| \le \epsilon_k. \tag{57}$$

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

$$-\langle\lambda^{k+1}, \mu_k^{-1}\lambda^k\rangle \le \langle\lambda^{k+1}, \tilde{g}(x^{k+1}, z^{k+1})\rangle. \tag{58}$$

Using these, (57), $\lambda^{k+1} \ge 0$ and $u \in \partial_z\tilde{f}(x^{k+1}, z^{k+1})$, we have

$$\begin{aligned}
&\rho_k\tilde{f}(x^{k+1}, z^{k+1}) - \rho_k\tilde{f}(x^{k+1}, \hat{z}^{k+1}) + G\|\lambda^{k+1}\|_1 - \langle\lambda^{k+1}, \mu_k^{-1}\lambda^k\rangle \\
&\le \rho_k\tilde{f}(x^{k+1}, z^{k+1}) - \rho_k\tilde{f}(x^{k+1}, \hat{z}^{k+1}) + \langle\lambda^{k+1}, -\tilde{g}(x^{k+1}, \hat{z}^{k+1}) - \mu_k^{-1}\lambda^k\rangle \\
&\overset{(58)}{\le} \rho_k\tilde{f}(x^{k+1}, z^{k+1}) - \rho_k\tilde{f}(x^{k+1}, \hat{z}^{k+1}) + \langle\lambda^{k+1}, \tilde{g}(x^{k+1}, z^{k+1}) - \tilde{g}(x^{k+1}, \hat{z}^{k+1}))\rangle \\
&\le \langle\rho_k u, z^{k+1} - \hat{z}^{k+1}\rangle + \langle\nabla_z\tilde{g}(x^{k+1}, z^{k+1})\lambda^{k+1}, z^{k+1} - \hat{z}^{k+1}\rangle \\
&= \langle\rho_k u + \nabla_y\tilde{g}(x^{k+1}, z^{k+1})\lambda^{k+1}, z^{k+1} - \hat{z}^{k+1}\rangle \le D_{\mathbf{y}}\epsilon_k,
\end{aligned} \tag{59}$$

where the first inequality is due to $\lambda^{k+1}\geq 0$ and $-\tilde g_i(x^{k+1},\hat z^{k+1})\geq G$ for all i, the third inequality follows from $u\in\partial_z\tilde f(x^{k+1},z^{k+1})$, $\lambda^{k+1}\geq 0$ and the convexity of $\tilde f(x^{k+1},\cdot)$ and $\tilde g_i(x^{k+1},\cdot)$ for all i, and the last inequality is due to (23), (57) and $z^{k+1},\hat z^{k+1}\in\mathcal{Y}$.

In view of (23), (59), $z^{k+1},\hat z^{k+1}\in\mathcal{Y}$, and the Lipschitz continuity of $\tilde f$, one has

$$\begin{aligned}D_{\mathbf{y}}\epsilon_k+\rho_k L_{\tilde f}D_{\mathbf{y}} &\overset{(23)}{\geq} D_{\mathbf{y}}\epsilon_k+\rho_k L_{\tilde f}\|z^{k+1}-\hat z^{k+1}\| \geq D_{\mathbf{y}}\epsilon_k+\rho_k(\tilde f(x^{k+1},\hat z^{k+1})-\tilde f(x^{k+1},z^{k+1}))\\ &\overset{(59)}{\geq} G\|\lambda^{k+1}\|_1-\langle\lambda^{k+1},\mu_k^{-1}\lambda^k\rangle \geq (G-\mu_k^{-1}\|\lambda^k\|)\|\lambda^{k+1}\|,\end{aligned}\tag{60}$$

where the first inequality is due to (23) and $z^{k+1},\hat z^{k+1}\in\mathcal{Y}$, the second inequality follows from $L_{\tilde f}$-Lipschitz continuity of $\tilde f$, and the last inequality is due to $\|\lambda^{k+1}\|_1\geq\|\lambda^{k+1}\|$. In addition, it follows from (49) and (52) that

$$G-\mu_k^{-1}\|\lambda^k\| \overset{(49)}{\geq} G-\sqrt{2\rho_k\mu_k^{-1}\vartheta} \overset{(52)}{\geq} G/2,$$

which together with (60) yields

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

The statement (53) then follows from this, $\epsilon_k\leq\epsilon_0$, and

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

We next show that (54) holds. Indeed, let $\tilde\lambda^{k+1}=[\lambda^k+\mu_k\tilde g(x^{k+1},y^{k+1})]_+$. It then follows from (55) that

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

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

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

The rest of the proof of (54) is similar to the one of (53) with u, z^{k+1} and λ^{k+1} being replaced with v, y^{k+1} and $\tilde\lambda^{k+1}$ respectively and thus omitted. □

Lemma 4. *Suppose that Assumptions 1 and 2 hold. Let $D_{\mathbf{y}}$, $\mathbb{K}$ and ϑ be defined in (23), (26) and (28), L_f, $L_{\tilde f}$ and G be given in Assumptions 1 and 2, and ϵ_0, τ, ρ_k and μ_k be given in Algorithm 1. Suppose that $(x^{k+1},y^{k+1},z^{k+1},\lambda^{k+1})$ is generated by Algorithm 1 for some $0\leq k\in\mathbb{K}-1$ with*

$$\rho_k^{-1}\mu_k\geq 8\tau^{-2}G^{-2}\vartheta.\tag{61}$$

Let

$$\lambda_{\mathbf{y}}^{k+1}=[\lambda^k+\mu_k\tilde g(x^{k+1},y^{k+1})]_+,\qquad \lambda_{\mathbf{z}}^{k+1}=\rho_k^{-1}[\lambda^k+\mu_k\tilde g(x^{k+1},z^{k+1})]_+.\tag{62}$$

Then we have

$$\begin{aligned}&\operatorname{dist}\Big(0,\partial f(x^{k+1},y^{k+1})+\rho_k\partial\tilde f(x^{k+1},y^{k+1})-\rho_k\big(\nabla_x\tilde f(x^{k+1},z^{k+1})+\nabla_x\tilde g(x^{k+1},z^{k+1})\lambda_{\mathbf{z}}^{k+1};0\big)\\&\qquad+\nabla\tilde g(x^{k+1},y^{k+1})\lambda_{\mathbf{y}}^{k+1}\Big)\leq\epsilon_k,\end{aligned}\tag{63}$$

$$\operatorname{dist}\Big(0,\rho_k(\partial_z\tilde f(x^{k+1},z^{k+1})+\nabla_z\tilde g(x^{k+1},z^{k+1})\lambda_{\mathbf{z}}^{k+1})\Big)\leq\epsilon_k,\tag{64}$$

$$\|[\tilde g(x^{k+1},z^{k+1})]_+\|\leq 2\mu_k^{-1}G^{-1}(\epsilon_0+\rho_k L_{\tilde f})D_{\mathbf{y}},\tag{65}$$

$$|\langle\lambda_{\mathbf{z}}^{k+1},\tilde g(x^{k+1},z^{k+1})\rangle|\leq 2\rho_k^{-1}\mu_k^{-1}G^{-1}(\epsilon_0+\rho_k L_{\tilde f})D_{\mathbf{y}}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0+\rho_k L_{\tilde f})D_{\mathbf{y}}\},\tag{66}$$

$$\|[\tilde g(x^{k+1},y^{k+1})]_+\|\leq 2\mu_k^{-1}G^{-1}(\epsilon_0+L_f+\rho_k L_{\tilde f})D_{\mathbf{y}},\tag{67}$$

$$|\langle\lambda_{\mathbf{y}}^{k+1},\tilde g(x^{k+1},z^{k+1})\rangle|\leq 2\mu_k^{-1}G^{-1}(\epsilon_0+L_f+\rho_k L_{\tilde f})D_{\mathbf{y}}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0+L_f+\rho_k L_{\tilde f})D_{\mathbf{y}}\}.\tag{68}$$

Proof. Suppose that $(x^{k+1},y^{k+1},z^{k+1},\lambda^{k+1})$ is generated by Algorithm 1 for some $0\leq k\in\mathbb{K}-1$ satisfying (61). Notice that $(x^{k+1},y^{k+1},z^{k+1})$ is an ϵ_k-primal-dual stationary point of (11). It then follows from Definition 1 that

$$\operatorname{dist}\big(0,\partial_{(x,y)}\mathcal{L}(x^{k+1},y^{k+1},z^{k+1},\lambda^k;\rho_k,\mu_k)\big)\leq\epsilon_k,\tag{69}$$

$$\operatorname{dist}\big(0,\partial_z\mathcal{L}(x^{k+1},y^{k+1},z^{k+1},\lambda^k;\rho_k,\mu_k)\big)\leq\epsilon_k.\tag{70}$$

In view of these, (7) and (62), one has

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

These relations together with (69) and (70) imply that (63) and (64) hold.

Notice from Algorithm 1 that $0<\tau<1$, which together with (61) implies that (52) holds for μ_k and ρ_k. It then follows that (53) and (54) hold, which immediately yields (65), (67), and

$$\|\lambda^{k+1}\|\le 2G^{-1}(\epsilon_0+\rho_k L_{\tilde{f}})D_{\mathbf{y}},\quad \|[\lambda^k+\mu_k\tilde{g}(x^{k+1},y^{k+1})]_+\|\le 2G^{-1}(\epsilon_0+L_f+\rho_k L_{\tilde{f}})D_{\mathbf{y}}. \tag{71}$$

Also, notice from (62) and $\lambda^{k+1}=[\lambda^k+\mu_k\tilde{g}(x^{k+1},z^{k+1})]_+$ that $\lambda_{\mathbf{z}}^{k+1}=\rho_k^{-1}\lambda^{k+1}$. By this, (62) and (71), one has

$$\|\lambda_{\mathbf{z}}^{k+1}\|\le 2\rho_k^{-1}G^{-1}(\epsilon_0+\rho_k L_{\tilde{f}})D_{\mathbf{y}},\qquad \|\lambda_{\mathbf{y}}^{k+1}\|\le 2G^{-1}(\epsilon_0+L_f+\rho_k L_{\tilde{f}})D_{\mathbf{y}}. \tag{72}$$

Observe from (62) that $\langle\lambda_{\mathbf{y}}^{k+1},\lambda^k+\mu_k\tilde{g}(x^{k+1},y^{k+1})\rangle=\|[\lambda^k+\mu_k\tilde{g}(x^{k+1},y^{k+1})]_+\|^2\ge 0$, which implies that

$$-\langle\lambda_{\mathbf{y}}^{k+1},\mu_k^{-1}\lambda^k\rangle\le\langle\lambda_{\mathbf{y}}^{k+1},\tilde{g}(x^{k+1},y^{k+1})\rangle. \tag{73}$$

In addition, we claim that

$$\|\lambda^k\|\le\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0+\rho_k L_{\tilde{f}})D_{\mathbf{y}}\}. \tag{74}$$

Indeed, (74) clearly holds if $k=0$. We now assume that $k>0$. Notice from Algorithm 1 that $\mu_{k-1}=\tau^3\mu_k$ and $\rho_{k-1}=\tau\rho_k$, which along with (61) imply that $\rho_{k-1}^{-1}\mu_{k-1}\ge 8G^{-2}\vartheta$. By this and Lemma 2 with k replaced by $k-1$, one can conclude that $\|\lambda^k\|\le 2G^{-1}(\epsilon_0+\rho_{k-1}L_{\tilde{f}})D_{\mathbf{y}}$. This together with $\rho_{k-1}<\rho_k$ implies that (74) holds as desired.

We next show that (66) and (68) hold. By $\lambda_{\mathbf{y}}^{k+1},\lambda_{\mathbf{z}}^{k+1}\ge 0$, (65), (67), (72), (73) and (74), one has

$$\begin{aligned}
\langle\lambda_{\mathbf{z}}^{k+1},\tilde{g}(x^{k+1},z^{k+1})\rangle\ &\le\ \langle\lambda_{\mathbf{z}}^{k+1},[\tilde{g}(x^{k+1},z^{k+1})]_+\rangle\le\|\lambda_{\mathbf{z}}^{k+1}\|\|[\tilde{g}(x^{k+1},z^{k+1})]_+\|\\
&\overset{(65)(72)}{\le}\ 4\rho_k^{-1}\mu_k^{-1}G^{-2}(\epsilon_0+\rho_k L_{\tilde{f}})^2D_{\mathbf{y}}^2,\\
\langle\lambda_{\mathbf{z}}^{k+1},\tilde{g}(x^{k+1},z^{k+1})\rangle&=\rho_k^{-1}\langle\lambda^{k+1},\tilde{g}(x^{k+1},z^{k+1})\rangle\overset{(58)}{\ge}-\rho_k^{-1}\langle\lambda^{k+1},\mu_k^{-1}\lambda^k\rangle\ge-\rho_k^{-1}\mu_k^{-1}\|\lambda^{k+1}\|\|\lambda^k\|\\
&\overset{(71)(74)}{\ge}\ -2\rho_k^{-1}\mu_k^{-1}G^{-1}(\epsilon_0+\rho_k L_{\tilde{f}})D_{\mathbf{y}}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0+\rho_k L_{\tilde{f}})D_{\mathbf{y}}\},\\
\langle\lambda_{\mathbf{y}}^{k+1},\tilde{g}(x^{k+1},y^{k+1})\rangle\ &\le\ \langle\lambda_{\mathbf{y}}^{k+1},[\tilde{g}(x^{k+1},y^{k+1})]_+\rangle\le\|\lambda_{\mathbf{y}}^{k+1}\|\|[\tilde{g}(x^{k+1},y^{k+1})]_+\|\\
&\overset{(67)(72)}{\le}\ 4\mu_k^{-1}G^{-2}(\epsilon_0+L_f+\rho_k L_{\tilde{f}})^2D_{\mathbf{y}}^2,\\
\langle\lambda_{\mathbf{y}}^{k+1},\tilde{g}(x^{k+1},y^{k+1})\rangle\ &\overset{(73)}{\ge}\ \langle\lambda_{\mathbf{y}}^{k+1},-\mu_k^{-1}\lambda^k\rangle\ge-\mu_k^{-1}\|\lambda_{\mathbf{y}}^{k+1}\|\|\lambda^k\|\\
&\overset{(72)(74)}{\ge}\ -2\mu_k^{-1}G^{-1}(\epsilon_0+L_f+\rho_k L_{\tilde{f}})D_{\mathbf{y}}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0+\rho_k L_{\tilde{f}})D_{\mathbf{y}}\}.
\end{aligned}$$

These relations imply that (66) and (68) hold. □

Lemma 5. *Suppose that Assumption 1 holds. Let* f^*, L_k, $\tilde{f}^*_{\rm hi}$, $D_{\mathbf{x}}$, $D_{\mathbf{y}}$, $f_{\rm hi}$, $f_{\rm low}$ $\tilde{f}_{\rm low}$, $\tilde{g}_{\rm hi}$, $\mathbb{K}$ *and* ϑ *be defined in* (1), (12), (22), (23), (24), (25), (26) *and* (28), $L_{\tilde{f}}$ *be given in Assumption 1,* ϵ_k, ρ_k, μ_k *and*

η_k be given in Algorithm 1, and

$$
\begin{aligned}
\alpha_k =&\ \min\left\{1, \sqrt{4\epsilon_k/(D_{\mathbf{y}}L_k)}\right\}, && (75)\\
\delta_k =&\ (2+\alpha_k^{-1})L_kD_{\mathbf{x}}^2 + \max\left\{\epsilon_k/D_{\mathbf{y}}, \alpha_kL_k/4\right\}D_{\mathbf{y}}^2, && (76)\\
M_k =&\ \frac{16\max\left\{1/(2L_k), \min\left\{D_{\mathbf{y}}/\epsilon_k, 4/(\alpha_kL_k)\right\}\right\}\mu_k}{\left[(3L_k+\epsilon_k/(2D_{\mathbf{y}}))^2/\min\{L_k, \epsilon_k/(2D_{\mathbf{y}})\}+3L_k+\epsilon_k/(2D_{\mathbf{y}})\right]^{-2}\epsilon_k^2}\\
&\times\left(\delta_k + 2\alpha_k^{-1}\left(f^* - f_{\mathrm{low}} + \rho_k(\tilde{f}^*_{\mathrm{hi}} - \tilde{f}_{\mathrm{low}}) + \rho_kL_{\tilde{f}}D_{\mathbf{y}} + 3\rho_k\vartheta + \mu_k\tilde{g}_{\mathrm{hi}}^2 + \epsilon_kD_{\mathbf{y}}/4 + L_kD_{\mathbf{x}}^2\right)\right), && (77)\\
T_k =&\ \left\lceil 16\left(f_{\mathrm{hi}} - f_{\mathrm{low}} + \rho_k\eta_k + \epsilon_kD_{\mathbf{y}}/4\right)L_k\epsilon_k^{-2} + 8(1+4D_{\mathbf{y}}^2L_k^2\epsilon_k^{-2})\mu_k^{-1} - 1\right\rceil_+, && (78)\\
N_k =&\ \left(\left\lceil 96\sqrt{2}\left(1+\left(24L_k+4\epsilon_k/D_{\mathbf{y}}\right)L_k^{-1}\right)\right\rceil + 2\right)\max\left\{2, \sqrt{D_{\mathbf{y}}L_k\epsilon_k^{-1}}\right\}\\
&\times\left((T_k+1)(\log M_k)_+ + T_k + 1 + 2T_k\log(T_k+1)\right). && (79)
\end{aligned}
$$

Then for all $0 \le k \in \mathbb{K}-1$, an ϵ_k-primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of problem (11) is successfully found at step 3 of Algorithm 1 that satisfies

$$
\max_z \mathcal{L}(x^{k+1}, y^{k+1}, z, \lambda^k; \rho_k, \mu_k) \le f_{\mathrm{hi}} + \rho_k\eta_k + \frac{\epsilon_kD_{\mathbf{y}}}{4} + \frac{1}{2\mu_k}\left(L_k^{-1}\epsilon_k^2 + 4D_{\mathbf{y}}^2L_k\right). \qquad (80)
$$

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

Proof. Observe from (7) and Assumption 1 that problem (11) can be viewed as

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

with

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

By (27) and Assumption 1, it can be verified that $\|[\lambda^k+\mu_k\tilde{g}(x,y)]_+\|^2/(2\mu_k)$ and $\|[\lambda^k+\mu_k\tilde{g}(x,z)]_+\|^2/(2\mu_k)$ are both $(\mu_kL_{\tilde{g}}^2 + \mu_k\tilde{g}_{\mathrm{hi}}L_{\nabla\tilde{g}} + \|\lambda^k\|L_{\nabla\tilde{g}})$-smooth on $\mathcal{X}\times\mathcal{Y}$. Using this and the fact that f_1 and $\tilde{f}_1$ are respectively $L_{\nabla f_1}$- and $L_{\nabla\tilde{f}_1}$-smooth on $\mathcal{X}\times\mathcal{Y}$, we can see that $h(x,y,z)$ is L_k-smooth on $\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}$ for all $0 \le k \in \mathbb{K}-1$, where L_k is given in (12). Consequently, it follows from Theorem 3 (see Appendix B) that an ϵ_k-primal-dual stationary point $(x^{k+1}, y^{k+1}, z^{k+1})$ of problem (11) is successfully found by Algorithm 4 at step 3 of Algorithm 1.

In addition, by (7), (8) and (24), one has

$$
\begin{aligned}
\min_{x,y}\max_z \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k) &\overset{(7)(8)}{=} \min_{x,y}\left\{f(x,y) + \rho_k\widetilde{L}(x,y,\lambda^k;\rho_k,\mu_k) - \min_z \rho_k\widetilde{L}(x,z,\lambda^k;\rho_k,\mu_k)\right\}\\
&\ge \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x,y) \overset{(24)}{=} f_{\mathrm{low}}. \qquad (81)
\end{aligned}
$$

Let (x^*, y^*) be an optimal solution of (1). It then follows that $f(x^*,y^*) = f^*$, $\tilde{f}(x^*,y^*) = \tilde{f}^*(x^*)$ and $\tilde{g}(x^*,y^*) \le 0$, where f^* and $\tilde{f}^*$ are defined in (1) and (13), respectively. Using these, (7), (8), (22), (25) and (49), we obtain that

$$
\begin{aligned}
&\min_{x,y}\max_z \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k) \le \max_z \mathcal{L}(x^*,y^*,z,\lambda^k;\rho_k,\mu_k)\\
&\overset{(7)(8)}{=} f(x^*,y^*) + \rho_k\tilde{f}(x^*,y^*) + \frac{1}{2\mu_k}\|[\lambda^k + \mu_k\tilde{g}(x^*,y^*)]_+\|^2 - \min_z \rho_k\widetilde{L}(x^*,z,\lambda^k;\rho_k,\mu_k)\\
&\le f^* + \rho_k\tilde{f}^*(x^*) + \frac{1}{2\mu_k}\|\lambda^k\|^2 - \min_z\left\{\rho_k\tilde{f}(x^*,z) + \frac{1}{2\mu_k}\|[\lambda^k + \mu_k\tilde{g}(x^*,z)]_+\|^2\right\}\\
&\overset{(22)(25)}{\le} f^* + \rho_k(\tilde{f}^*_{\mathrm{hi}} - \tilde{f}_{\mathrm{low}}) + \frac{1}{2\mu_k}\|\lambda^k\|^2 \overset{(49)}{\le} f^* + \rho_k(\tilde{f}^*_{\mathrm{hi}} - \tilde{f}_{\mathrm{low}}) + \rho_k\vartheta, \qquad (82)
\end{aligned}
$$

where the second inequality is due to $\tilde f(x^*,y^*) = \tilde f^*(x^*)$, $\tilde g(x^*,y^*) \le 0$ and (8). Also, by (7), (23), (24), (25) and (49), one has

$$
\begin{aligned}
&\min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k)\\
&\overset{(7)}{\ge} \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \left\{ f(x,y) + \rho_k(\tilde f(x,y) - \tilde f(x,z)) - \frac{1}{2\mu_k}\|[\lambda^k + \mu_k \tilde g(x,z)]_+\|^2 \right\}\\
&\ge \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \left\{ f(x,y) - \rho_k L_{\tilde f}\|y-z\| - \frac{1}{2\mu_k}\left(\|\lambda^k\| + \mu_k\|[\tilde g(x,z)]_+\|\right)^2 \right\}\\
&\ge \min_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \left\{ f(x,y) - \rho_k L_{\tilde f}\|y-z\| - \frac{1}{\mu_k}\|\lambda^k\|^2 - \mu_k\|[\tilde g(x,z)]_+\|^2 \right\}\\
&\ge f_{\mathrm{low}} - \rho_k L_{\tilde f} D_{\mathbf{y}} - 2\rho_k\vartheta - \mu_k \tilde g_{\mathrm{hi}}^2, \qquad (83)
\end{aligned}
$$

where the second inequality is due to $\lambda^k \in \mathbb{R}^l_+$ and $L_{\tilde f}$-Lipschitz continuity of $\tilde f$ (see Assumption 1(i)), and the last inequality is due to (23), (24), (25) and (49). Notice from step 2 of Algorithm 1 that y^k_{init} is an approximate solution of $\min_z \widetilde{L}(x^k,z,\lambda^k;\rho_k,\mu_k)$ satisfying (9). It then follows from (7), (8), (9) and (24) that

$$
\begin{aligned}
\max_z \mathcal{L}(x^k,y^k_{\mathrm{init}},z,\lambda^k;\rho_k,\mu_k) &\overset{(7)(8)}{=} f(x^k,y^k_{\mathrm{init}}) + \rho_k\left(\widetilde{\mathcal{L}}(x^k,y^k_{\mathrm{init}},\lambda^k;\rho_k,\mu_k) - \min_z \widetilde{\mathcal{L}}(x^k,z,\lambda^k;\rho_k,\mu_k)\right)\\
&\overset{(9)}{\le} f(x^k,y^k_{\mathrm{init}}) + \rho_k\eta_k \overset{(24)}{\le} f_{\mathrm{hi}} + \rho_k\eta_k. \qquad (84)
\end{aligned}
$$

To complete the rest of the proof, let

$$
H(x,y,z) = \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k), \quad H^* = \min_{x,y}\max_z \mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k), \qquad (85)
$$

$$
H_{\mathrm{low}} = \min\left\{\mathcal{L}(x,y,z,\lambda^k;\rho_k,\mu_k) | (x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\right\}. \qquad (86)
$$

In view of these, (81), (82), (83) and (84), we obtain that

$$
\begin{aligned}
&\max_z H(x^k,y^k_{\mathrm{init}},z) \overset{(84)}{\le} f_{\mathrm{hi}} + \rho_k\eta_k,\\
&f_{\mathrm{low}} \overset{(81)}{\le} H^* \overset{(82)}{\le} f^* + \rho_k(\tilde f^*_{\mathrm{hi}} - \tilde f_{\mathrm{low}}) + \rho_k\vartheta,\\
&H_{\mathrm{low}} \overset{(83)}{\ge} f_{\mathrm{low}} - \rho_k L_{\tilde f} D_{\mathbf{y}} - 2\rho_k\vartheta - \mu_k\tilde g^2_{\mathrm{hi}}.
\end{aligned}
$$

Using these and Theorem 3 (see Appendix B) with $\hat x^0 = (x^k, y^k_{\mathrm{init}})$, $\epsilon = \epsilon_k$, $\hat\epsilon_0 = \epsilon_k/(2\sqrt{\mu_k})$, $L_{\nabla h} = L_k$, $\hat\alpha = \alpha_k$, $\hat\delta = \delta_k$, $D_p = D_{\mathbf{x}}$, $D_q = D_{\mathbf{y}}$, and H, H^*, H_{low} given in (85) and (86), we can conclude that the ϵ_k-primal-dual stationary point $(x^{k+1},y^{k+1},z^{k+1})$ of problem (11) found at step 3 of Algorithm 1 satisfies (80). Moreover, the total number of evaluations of ∇f_1, $\nabla\tilde f_1$, $\nabla\tilde g$ and proximal operator of f_2 and $\tilde f_2$ performed by Algorithm 4 at step 3 of Algorithm 1 is no more than N_k, respectively. □

Lemma 6. *Suppose that Assumptions 1 and 2 hold. Let $\tilde f^*$, L_k, $D_{\mathbf{y}}$, f_{hi}, f_{low} and $\mathbb{K}$ be defined in* (13), (12), (23), (24) *and* (26), *L_f, $L_{\tilde f}$ and G be given in Assumptions 1 and 2, and ϵ_k, ρ_k, μ_k, η_k and λ^0 be given in Algorithm 1. Suppose that $(x^{k+1},y^{k+1},\lambda^{k+1})$ is generated by Algorithm 1 for some $0\le k\in\mathbb{K}-1$ satisfying* (61). *Then we have*

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

Proof. Notice from (61) and the proof of Lemma 4 that (74) holds. Using this, (7), (8), (24) and (44), we have

$$
\begin{aligned}
&\max_z \mathcal{L}(x^{k+1}, y^{k+1}, z, \lambda^k; \rho_k, \mu_k) \\
&\overset{(7)(8)}{=} f(x^{k+1}, y^{k+1}) + \rho_k \tilde{f}(x^{k+1}, y^{k+1}) + \frac{1}{2\mu_k}\|[\lambda^k + \mu_k \tilde{g}(x^{k+1}, y^{k+1})]_+\|^2 - \min_z \rho_k \widetilde{\mathcal{L}}(x^{k+1}, z, \lambda^k; \rho_k, \mu_k) \\
&\geq f(x^{k+1}, y^{k+1}) + \rho_k \tilde{f}(x^{k+1}, y^{k+1}) - \min_z \rho_k \widetilde{\mathcal{L}}(x^{k+1}, z, \lambda^k; \rho_k, \mu_k) \\
&\overset{(24)(44)}{\geq} f_{\rm low} + \rho_k\big(\tilde{f}(x^{k+1}, y^{k+1}) - \tilde{f}^*(x^{k+1})\big) - \frac{1}{2\mu_k}\|\lambda^k\|^2 \\
&\overset{(74)}{\geq} f_{\rm low} + \rho_k\big(\tilde{f}(x^{k+1}, y^{k+1}) - \tilde{f}^*(x^{k+1})\big) - \mu_k^{-1}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\}/2.
\end{aligned}
$$

This together with (80) implies that

$$
\begin{aligned}
\tilde{f}(x^{k+1}, y^{k+1}) - \tilde{f}^*(x^{k+1}) \ &\leq \rho_k^{-1}\left(f_{\rm hi} - f_{\rm low} + \rho_k\eta_k + \frac{\epsilon_k D_{\mathbf{y}}}{4} + \frac{1}{2\mu_k}\left(L_k^{-1}\epsilon_k^2 + 4D_{\mathbf{y}}^2 L_k\right)\right) \\
&\quad + \rho_k^{-1}\mu_k^{-1}\max\{\|\lambda^0\|,\ 2G^{-1}(\epsilon_0 + \rho_k L_{\tilde{f}})D_{\mathbf{y}}\}/2. \qquad (87)
\end{aligned}
$$

On the other hand, let $\lambda^* \in \mathbb{R}^l_+$ be an optimal Lagrangian multiplier of problem (13) with $x = x^{k+1}$. It then follows from Lemma 1(i) that $\|\lambda^*\| \leq G^{-1}L_{\tilde{f}}D_{\mathbf{y}}$. Using these, Lemma 2, (13) and (67), we have

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

The conclusion of this lemma then follows from this and (87). □

Lemma 7. *Suppose that Assumption 1 holds. Let* $\widetilde{L}_k$, $D_{\mathbf{y}}$ *and* $\mathbb{K}$ *be defined in* (10), (23) *and* (26), η_k *be given in Algorithm 1, and*

$$\widetilde{N}_k = \left\lceil D_{\mathbf{y}}\sqrt{2\eta_k^{-1}\widetilde{L}_k}\right\rceil. \qquad (88)$$

Then for all $0 \leq k \in \mathbb{K} - 1$, $y^k_{\rm init}$ *satisfying* (9) *is found at step 3 of Algorithm 1 by Algorithm 2 in no more than* $\widetilde{N}_k$ *evaluations of* $\nabla\tilde{f}_1$, $\nabla\tilde{g}$ *and the proximal operator of* $\tilde{f}_2$*, respectively.*

Proof. Notice from (8) and Algorithm 1 that $y^k_{\rm init}$ satisfying (9) is found by Algorithm 2 applied to the problem

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

where $\phi(y) = \tilde{f}_1(x^k, y) + \|[\lambda^k + \mu_k\tilde{g}(x^k, y)]_+\|^2/(2\rho_k\mu_k)$ and $P(y) = \tilde{f}_2(y)$. By Assumption 1 and (27), one can see that ϕ is $\widetilde{L}_k$-smooth on $\operatorname{dom} P$, where $\widetilde{L}_k$ is given in (10). It then follows from this and Theorem 2 (see Appendix A) with $\tilde{\epsilon} = \eta_k$, $D_P = D_{\mathbf{y}}$ and $L_{\nabla\phi} = \widetilde{L}_k$ that Algorithm 2 finds $y^k_{\rm init}$ satisfying (9) in no more than $\widetilde{N}_k$ iterations. Notice that each iteration of Algorithm 2 requires one evaluation of $\nabla\phi$ and the proximal operator of P, respectively. Hence, the conclusion of this lemma holds. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. (i) Observe from the definition of K in (26) and $\epsilon_k = \epsilon_0\tau^k$ that K is the smallest nonnegative integer such that $\epsilon_K \leq \varepsilon$. Hence, Algorithm 1 terminates and outputs (x^{K+1}, y^{K+1}) after $K + 1$ outer iterations. Also, one can see from Algorithm 1 that

$$\rho_K = \epsilon_K^{-1}, \quad \mu_K = \epsilon_K^{-3}, \quad \eta_K = \epsilon_K. \qquad (89)$$

Moreover, notice from the assumption of Theorem 1 that $\varepsilon^{-2} - 8\tau^{-2}G^{-2}\vartheta \geq 0$. It then follows from this and (89) that

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

which implies that (61) holds for $k = K$. In addition, by (12), (29), (49) and $\mu_k \geq \rho_k \geq 1$, one has that for all $0 \leq k \in \mathbb{K} - 1$,

$$\begin{aligned}
2\mu_k L_{\tilde{g}}^2 \leq L_k &\overset{(12)}{=} L_{\nabla f_1} + 2\rho_k L_{\nabla \tilde{f}_1} + 2\mu_k L_{\tilde{g}}^2 + 2\mu_k \tilde{g}_{\rm hi} L_{\nabla \tilde{g}} + 2\|\lambda^k\| L_{\nabla \tilde{g}} \\
&\overset{(49)}{\leq} L_{\nabla f_1} + 2\rho_k L_{\nabla \tilde{f}_1} + 2\mu_k L_{\tilde{g}}^2 + 2\mu_k \tilde{g}_{\rm hi} L_{\nabla \tilde{g}} + 2\sqrt{2\rho_k \mu_k \vartheta} L_{\nabla \tilde{g}} \leq \mu_k L. \qquad (90)
\end{aligned}$$

It then follows from $\epsilon_K \leq \varepsilon$, (89) and Lemmas 4 and 6 that (34)-(40) hold.

(ii) Let K and N be given in (26) and (41). Recall from Lemmas 5 and 7 that the number of evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla \tilde{g}$, proximal operator of f_2 and $\tilde{f}_2$ performed by Algorithms 2 and 4 at iteration k of Algorithm 1 is at most $N_k + \widetilde{N}_k$, where N_k and $\widetilde{N}_k$ are given in (79) and (88), respectively. By this and statement (i) of this theorem, one can observe that the total number of evaluations of ∇f_1, $\nabla \tilde{f}_1$, $\nabla \tilde{g}$ and proximal operator of f_2 and $\tilde{f}_2$ performed in Algorithm 1 is no more than $\sum_{k=0}^K (N_k + \widetilde{N}_k)$, respectively. As a result, to prove statement (ii) of this theorem, it suffices to show that $\sum_{k=0}^K (N_k + \widetilde{N}_k) \leq N$.

To this end, using $\mu_k \geq 1 \geq \epsilon_k$, (30), (31), (32), (75), (76), (77), (78) and (90), we obtain that

$$\begin{aligned}
&1 \geq \alpha_k \geq \min\left\{1, \sqrt{4\epsilon_k/(\mu_k D_{\mathbf{y}} L)}\right\} \geq \epsilon_k^{1/2} \mu_k^{-1/2} \alpha, && (91)\\
&\delta_k \leq (2 + \epsilon_k^{-1/2} \mu_k^{1/2} \alpha^{-1}) \mu_k L D_{\mathbf{x}}^2 + \max\{1/D_{\mathbf{y}}, \mu_k L/4\} D_{\mathbf{y}}^2 \leq \epsilon_k^{-1/2} \mu_k^{3/2} \delta, && (92)\\
&M_k \leq \frac{16 \max\left\{1/(4\mu_k L_{\tilde{g}}^2), 2/(\epsilon_k^{1/2} \mu_k^{-1/2} \alpha \mu_k L_{\tilde{g}}^2)\right\} \mu_k}{\left[(3\mu_k L + 1/(2D_{\mathbf{y}}))^2 / \min\{2\mu_k L_{\tilde{g}}^2, \epsilon_k/(2D_{\mathbf{y}})\} + 3\mu_k L + 1/(2D_{\mathbf{y}})\right]^{-2} \epsilon_k^2} \times \Bigg(\epsilon_k^{-1/2} \mu_k^{3/2} \delta \\
&\qquad + 2\epsilon_k^{-1/2} \mu_k^{1/2} \alpha^{-1} \Big(f^* - f_{\rm low} + \rho_k(\tilde{f}_{\rm hi}^* - \tilde{f}_{\rm low}) + \rho_k L_{\tilde{f}} D_{\mathbf{y}} + 3\rho_k \vartheta + \mu_k \tilde{g}_{\rm hi}^2 + \frac{D_{\mathbf{y}}}{4} + \mu_k L D_{\mathbf{x}}^2\Big)\Bigg) && (93)\\
&\quad \leq \frac{16 \epsilon_k^{-1/2} \mu_k^{-1/2} \max\left\{1/(4L_{\tilde{g}}^2), 2/(\alpha L_{\tilde{g}}^2)\right\} \mu_k}{\epsilon_k^2 \mu_k^{-4} \left[(3L + 1/(2D_{\mathbf{y}}))^2 / \min\{2L_{\tilde{g}}^2, 1/(2D_{\mathbf{y}})\} + 3L + 1/(2D_{\mathbf{y}})\right]^{-2} \epsilon_k^2} \times (\epsilon_k^{-1/2} \mu_k^{3/2}) \\
&\qquad \times \left(\delta + 2\alpha^{-1} \Big(f^* - f_{\rm low} + \tilde{f}_{\rm hi}^* - \tilde{f}_{\rm low} + L_{\tilde{f}} D_{\mathbf{y}} + 3\vartheta + \tilde{g}_{\rm hi}^2 + \frac{D_{\mathbf{y}}}{4} + L D_{\mathbf{x}}^2\Big)\right) \; = \; \epsilon_k^{-5} \mu_k^6 M, && (94)\\
&T_k \leq \left\lceil 16 \left(f_{\rm hi} - f_{\rm low} + \rho_k \eta_k + \frac{D_{\mathbf{y}}}{4}\right) \epsilon_k^{-2} \mu_k L + 8(1 + 4D_{\mathbf{y}}^2 \mu_k^2 L^2 \epsilon_k^{-2}) \mu_k^{-1} - 1 \right\rceil_+ \leq \epsilon_k^{-2} \mu_k T, && (95)
\end{aligned}$$

where (91) follows from (30), (75) and (90); (92) is due to (30), (76), (91) and $\mu_k \geq 1 \geq \epsilon_k$; (93) is due to (77), (90), (91), (92) and $\epsilon_k \in (0, 1]$; (94) follows from $\mu_k \geq \rho_k \geq 1 \geq \epsilon_k$ and (31); and (95) is due to (90), (32) and the fact that $\epsilon_k \in (0, 1]$ and $\rho_k \eta_k = 1$. By the above inequalities, (79), (90), $T > 1$ and $\mu_k \geq 1 \geq \epsilon_k$, one has

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

where the first inequality follows from $\epsilon_k \in (0,1]$, (79), (90), (94) and (95), and the second and third inequalities are due to the fact that $\mu_k \geq 1 \geq \epsilon_k$ and $T > 1$. By the definition of K in (26), one has $\tau^K \geq \tau\varepsilon/\epsilon_0$. Also, notice from Algorithm 1 that $\mu_k = \epsilon_k^{-3} = (\epsilon_0\tau^k)^{-3}$. It then follows from these and (96) that

$$
\begin{aligned}
\sum_{k=0}^{K} N_k &\leq \sum_{k=0}^{K} \left(\left\lceil 96\sqrt{2}\left(1 + (12L + 2/D_y)/L_{\tilde g}^2\right) \right\rceil + 2 \right) \max\left\{2, \sqrt{D_y L}\right\} T \\
&\qquad \times \epsilon_k^{-7} \left(56\log(1/\epsilon_k) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&= \left(\left\lceil 96\sqrt{2}\left(1 + (12L + 2/D_y)/L_{\tilde g}^2\right) \right\rceil + 2 \right) \max\left\{2, \sqrt{D_y L}\right\} T \\
&\quad \times \sum_{k=0}^{K} \epsilon_0^{-7}\tau^{-7k} \left(56k\log(1/\tau) + 56\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&\leq \left(\left\lceil 96\sqrt{2}\left(1 + (12L + 2/D_y)/L_{\tilde g}^2\right) \right\rceil + 2 \right) \max\left\{2, \sqrt{D_y L}\right\} T \\
&\quad \times \sum_{k=0}^{K} \epsilon_0^{-7}\tau^{-7k} \left(56K\log(1/\tau) + 56\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&\leq \left(\left\lceil 96\sqrt{2}\left(1 + (12L + 2/D_y)/L_{\tilde g}^2\right) \right\rceil + 2 \right) \max\left\{2, \sqrt{D_y L}\right\} T\epsilon_0^{-7} \\
&\quad \times \tau^{-7K}(1-\tau^7)^{-1} \left(56K\log(1/\tau) + 56\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right) \\
&\leq \left(\left\lceil 96\sqrt{2}\left(1 + (12L + 2/D_y)/L_{\tilde g}^2\right) \right\rceil + 2 \right) \max\left\{2, \sqrt{D_y L}\right\} T\epsilon_0^{-7}(1-\tau^7)^{-1} \\
&\quad \times (\tau\varepsilon/\epsilon_0)^{-7} \left(56K\log(1/\tau) + 56\log(1/\epsilon_0) + 2(\log M)_+ + 2 + 2\log(2T)\right), \qquad (97)
\end{aligned}
$$

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

In addition, observe from (10), (29), (49) and $\rho_k^{-1}\mu_k \geq 1$, one has that for all $0 \leq k \in \mathbb{K} - 1$,

$$
\widetilde{L}_k = L_{\nabla \tilde f_1} + \rho_k^{-1}(\mu_k L_{\tilde g}^2 + \mu_k \tilde g_{\rm hi} L_{\nabla \tilde g} + \|\lambda^k\| L_{\nabla \tilde g}) \leq L_{\nabla \tilde f_1} + \rho_k^{-1}(\mu_k L_{\tilde g}^2 + \mu_k \tilde g_{\rm hi} L_{\nabla \tilde g} + \sqrt{2\rho_k\mu_k\vartheta} L_{\nabla \tilde g}) \leq \rho_k^{-1}\mu_k \widetilde{L}.
$$

Using this, (88), $\epsilon_k = \epsilon_0\tau^k$, $\rho_k = \epsilon_k^{-1}$, $\mu_k = \epsilon_k^{-3}$ and $\eta_k = \epsilon_k$, we have

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

where the second last inequality is due to $\sum_{k=0}^{K} \tau^{-3k/2} \leq \tau^{-3K/2}/(1-\tau^{3/2})$, and the last inequality follows from $\tau^K \geq \tau\varepsilon/\epsilon_0$. This together with (41) and (97) implies that $\sum_{k=1}^{K}(N_k + \widetilde{N}_k) \leq N$. Hence, statement (ii) of Theorem 1 holds. □

References

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

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

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

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

[5] L. Chen, J. Xu, and J. Zhang. On bilevel optimization without lower-level strong convexity. *arXiv preprint arXiv:2301.00712*, 2023.

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

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

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

[9] Y.-H. Dai, J. Wang, and L. Zhang. Optimality conditions and numerical algorithms for a class of linearly constrained minimax optimization problems. *arXiv preprint arXiv:2204.09185*, 2022.

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

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

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

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

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

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

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

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

[21] 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.

[22] 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.

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

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

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

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

[27] 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.

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

[29] P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283, 2021.

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

[31] 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.

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

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

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

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

[36] 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.

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

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

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

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

[41] 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.

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

[43] 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.

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

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

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

[47] 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.

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

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

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

[51] 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.

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

[53] 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.

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

[55] 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.

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

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

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

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

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

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

[62] 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.

[63] M. Ye, B. Liu, S. Wright, P. Stone, and Q. Liu. Bome! bilevel optimization made easy: A simple first-order approach. *arXiv preprint arXiv:2209.08709*, 2022.

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

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

$$\Psi^* = \min_x \{\Psi(x) := \phi(x) + P(x)\}, \tag{98}$$

where $\phi, P : \mathbb{R}^m \to (-\infty, \infty]$ are closed convex functions, ϕ is continuously differentiable on an open set containing dom P, and $\nabla\phi$ is $L_{\nabla\phi}$-Lipschitz continuous on dom P. In addition, we assume that dom P is compact and let $D_P := \max_{x,y\in \text{dom } P} \|x - y\|$.

We next present an optimal first-order method for solving (98). It is a variant of Nesterov's optimal first-order methods [45] and has been studied in, for example, [56, Section 3].

Algorithm 2 An optimal first-order method for (98) with general convex ϕ

```
Input: $\tilde{\epsilon} > 0$, $\tilde{x}^0 \in \mathrm{dom}\, P$ and $x^0 = z^0 = \tilde{x}^0$.
for k=0,1,... do
    Set $y^k = (kx^k + 2z^k)/(k+2)$.
    Compute $z^{k+1}$ as
        $z^{k+1} = \arg\min_z \left\{ \ell(z; y^k) + \frac{L_{\nabla\phi}}{k+2}\|z - z^k\|^2 \right\},$
    where
        $\ell(x; y) := \phi(y) + \langle \nabla\phi(y), x - y \rangle + P(x).$    (99)
    Set $x^{k+1} = (kx^k + 2z^{k+1})/(k+2)$.
    Terminate the algorithm and output $x^{k+1}$ if
        $\Psi(x^{k+1}) - \underline{\Psi}_{k+1} \leq \tilde{\epsilon}, \qquad \text{where} \quad \underline{\Psi}_{k+1} = \frac{4}{(k+1)(k+3)} \min \left\{ \sum_{i=0}^{k} \frac{i+2}{2} \ell(x; y^i) \right\}.$
end for
```

The following result provides an *iteration-complexity* of Algorithm 2 for finding an $\tilde{\epsilon}$-optimal solution[9] of (98). It is an immediate consequence of [56, Corollary 1] and its proof is thus omitted.

Theorem 2. *Let $\{(x^k, y^k)\}$ be generated by Algorithm 2 and $\ell(\cdot;\cdot)$ be defined in (99). Then, $\Psi(x^k) - \Psi^* \leq \Psi(x^k) - \underline{\Psi}_k$ for all $k \geq 1$. Moreover, for any given $\tilde{\epsilon} > 0$, Algorithm 2 finds an approximate solution x^{k+1} of problem (98) such that $\Psi(x^{k+1}) - \Psi^* \leq \Psi(x^{k+1}) - \underline{\Psi}_{k+1} \leq \tilde{\epsilon}$ in no more than $\widetilde{T}$ iterations, where*

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

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

In this part we present a first-order method proposed in [38, Algorithm 2] for finding an ϵ-stationary point of the nonconvex-concave minimax problem

$$H^* = \min_x \max_y \left\{ H(x, y) := h(x, y) + p(x) - q(y) \right\}, \tag{100}$$

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

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

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

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

For ease of presentation, we define

$$D_p = \max\{\|u - v\| \,|\, u, v \in \mathrm{dom}\, p\}, \quad D_q = \max\{\|u - v\| \,|\, u, v \in \mathrm{dom}\, q\}, \tag{101}$$

$$H_{\mathrm{low}} = \min\{H(x, y) | (x, y) \in \mathrm{dom}\, p \times \mathrm{dom}\, q\}. \tag{102}$$

Given an iterate (x^k, y^k), the first-order method [38, Algorithm 2] finds the next iterate (x^{k+1}, y^{k+1}) by applying a modified optimal first-order method [38, Algorithm 1] to the strongly-convex-strongly-concave minimax problem

$$\min_x \max_y \left\{ h_k(x, y) = h(x, y) - \epsilon\|y - y^0\|^2/(4D_q) + L_{\nabla h}\|x - x^k\|^2 \right\}. \tag{103}$$

[9] An $\tilde{\epsilon}$-optimal solution of problem (98) is a point x satisfying $\Psi(x) - \Psi^* \leq \tilde{\epsilon}$.

For ease reference, we next present the modified optimal first-order method [38, Algorithm 1] in Algorithm 3 below for solving the strongly-convex-strongly-concave minimax problem

$$\min_x \max_y \left\{\bar{h}(x,y) + p(x) - q(y)\right\}, \tag{104}$$

where $\bar{h}(x,y)$ is σ_x-strongly-convex-σ_y-strongly-concave and $L_{\nabla\bar{h}}$-smooth on $\mathrm{dom}\, p \times \mathrm{dom}\, q$ for some $\sigma_x, \sigma_y > 0$. In Algorithm 3, the functions $\hat{h}$, a_x^k and a_y^k are defined as follows:

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

where y_g^k and z_g^k are generated at iteration k of Algorithm 3 below.

Algorithm 3 A modified optimal first-order method for problem (104)

```
Input: $\tau > 0$, $\bar{z}^0 = z_f^0 \in -\sigma_x \mathrm{dom}\, p$,[10] $\bar{y}^0 = y_f^0 \in \mathrm{dom}\, q$, $(z^0, y^0) = (\bar{z}^0, \bar{y}^0)$, $\bar{\alpha} = \min\left\{1, \sqrt{8\sigma_y/\sigma_x}\right\}$,
    $\eta_z = \sigma_x/2$, $\eta_y = \min\{1/(2\sigma_y), 4/(\bar{\alpha}\sigma_x)\}$, $\beta_t = 2/(t+3)$, $\zeta = \left(2\sqrt{5}(1 + 8L_{\nabla\bar{h}}/\sigma_x)\right)^{-1}$, $\gamma_x = \gamma_y =$
    $8\sigma_x^{-1}$, 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

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

end for
```

We are now ready to present the first-order method [38, Algorithm 2] for finding an ϵ-stationary point of (100) in Algorithm 4 below.

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

Algorithm 4 A first-order method for problem (100)

```
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 $\hat{\epsilon}_k = \hat{\epsilon}_0/(k+1)$.
1: for $k = 0, 1, 2, \ldots$ do
2:     Call Algorithm 3 with $\bar{h} \leftarrow h_k$, $\tau \leftarrow \hat{\epsilon}_k$, $\sigma_x \leftarrow L_{\nabla h}$, $\sigma_y \leftarrow \epsilon/(2D_q)$, $L_{\nabla \bar{h}} \leftarrow 3L_{\nabla h} + \epsilon/(2D_q)$, $\bar{z}^0 = z_f^0 \leftarrow -\sigma_x x^k$, $\bar{y}^0 = y_f^0 \leftarrow y^k$, and denote its output by $(x^{k+1}, y^{k+1})$, where $h_k$ is given in (103).
3:     Terminate the algorithm and output $(x_\epsilon, y_\epsilon) = (x^{k+1}, y^{k+1})$ if
           $\|x^{k+1} - x^k\| \le \epsilon/(4L_{\nabla h})$.
4: end for
```

The following theorem presents the iteration complexity of Algorithm 4, whose proof is given in [38, Theorem 2].

Theorem 3 (**Complexity of Algorithm 4**). *Suppose that Assumption 3 holds. Let H^*, H D_p, D_q, and H_{low} be defined in* (100), (101) *and* (102), *$L_{\nabla h}$ be given in Assumption 3, ϵ, ϵ_0 and $\hat{x}^0$ be given in Algorithm 4, and*

$$\begin{aligned}
\hat{\alpha} =&\ \min\left\{1, \sqrt{4\epsilon/(D_q L_{\nabla h})}\right\}, \\
\hat{\delta} =&\ (2 + \hat{\alpha}^{-1}) L_{\nabla h} D_p^2 + \max\left\{\epsilon/D_q, \hat{\alpha} L_{\nabla h}/4\right\} D_q^2, \\
\widehat{T} =&\ \left\lceil 16(\max_y H(\hat{x}^0, y) - H^* + \epsilon D_q/4) L_{\nabla h} \epsilon^{-2} + 32\hat{\epsilon}_0^2 (1 + 4D_q^2 L_{\nabla h}^2 \epsilon^{-2}) \epsilon^{-2} - 1 \right\rceil_+, \\
\widehat{N} =&\ \left(\left\lceil 96\sqrt{2}\left(1 + (24 L_{\nabla h} + 4\epsilon/D_q) L_{\nabla h}^{-1}\right)\right\rceil + 2\right) \max\left\{2, \sqrt{D_q L_{\nabla h} \epsilon^{-1}}\right\} \\
&\times \left((T+1) \left(\log \frac{4 \max\left\{\frac{1}{2L_{\nabla h}}, \min\left\{\frac{D_q}{\epsilon}, \frac{4}{\hat{\alpha} L_{\nabla h}}\right\}\right\} \left(\hat{\delta} + 2\hat{\alpha}^{-1}(H^* - H_{\text{low}} + \epsilon D_q/4 + L_{\nabla h} D_p^2)\right)}{\left[(3L_{\nabla h} + \epsilon/(2D_q))^2 / \min\{L_{\nabla h}, \epsilon/(2D_q)\} + 3L_{\nabla h} + \epsilon/(2D_q)\right]^{-2} \hat{\epsilon}_0^2} \right)_+ \right. \\
&\left. + \widehat{T} + 1 + 2\widehat{T} \log(\widehat{T} + 1) \right).
\end{aligned}$$

Then Algorithm 4 terminates and outputs an ϵ-stationary point (x_ϵ, y_ϵ) of (100) *in at most $\widehat{T} + 1$ outer iterations that satisfies*

$$\max_y H(x_\epsilon, y) \le \max_y H(\hat{x}^0, y) + \epsilon D_q/4 + 2\hat{\epsilon}_0^2 \left(L_{\nabla h}^{-1} + 4D_q^2 L_{\nabla h} \epsilon^{-2}\right). \tag{105}$$

Moreover, the total number of evaluations of ∇h and proximal operator of p and q performed in Algorithm 4 is no more than $\widehat{N}$, respectively.