BILEVEL OPTIMIZATION WITHOUT LOWER-LEVEL STRONG CONVEXITY FROM THE HYPER-OBJECTIVE PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Bilevel optimization reveals the inner structure of otherwise oblique optimization problems, such as hyperparameter tuning, neural architecture search, and meta-learning. A common goal in bilevel optimization is to find stationary points of the hyper-objective function. Although this hyper-objective approach is widely used, its theoretical properties have not been thoroughly investigated in cases where the lower-level functions lack strong convexity. This work takes a step forward when the typical lower-level strong convexity assumption is absent. Our hardness results show that bilevel optimization for general convex lower-level functions is intractable to solve. We then identify several regularity conditions of the lower-level problems that can provably confer tractability. Under these conditions, we propose the Inexact Gradient-Free Method (IGFM), which uses the Switching Gradient Method (SGM) as an efficient sub-routine, to find an approximate stationary point of the hyper-objective in polynomial time.

1 INTRODUCTION

The goal of bilevel optimization (BLO) is to minimize the upper-level (UL) function $f(x,y)$ under the constraint that y is minimized w.r.t. the lower-level (LL) function $g(x,y)$ on a closed convex set $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$. Mathematically, it can be formulated as:

$$\min_{x \in \mathbb{R}^{d_x}, y \in Y^*(x)} f(x,y), \quad Y^*(x) = \arg\min_{y \in \mathcal{Y}} g(x,y). \tag{1}$$

BLO in this form has received increasing attention due to its wide applications in many machine learning problems, including hyperparameter tuning (Franceschi et al., 2018; Pedregosa, 2016), neural architecture search (Liu et al., 2019; Wang et al., 2022b; Zoph & Le, 2016; Zhang et al., 2021), meta-learning (Franceschi et al., 2018; Hospedales et al., 2021; Ravi & Larochelle, 2017; Pham et al., 2021), out-of-distribution learning (Zhou et al., 2022), adversarial training (Goodfellow et al., 2020; Sinha et al., 2018; Lin et al., 2020a;b), reinforcement learning (Konda & Tsitsiklis, 1999; Hong et al., 2023), causal learning (Jiang & Veitch, 2022; Arjovsky et al., 2019).

The hyper-objective approaches (Dempe, 2002; Dempe & Zemkoho, 2020; Liu et al., 2020; 2021) reformulate Problem 1 by

$$\min_{x \in \mathbb{R}^d} \varphi(x), \text{ where } \varphi(x) = \min_{y \in Y^*(x)} f(x,y) \text{ is called the hyper-objective.} \tag{2}$$

It transforms the problem into the composition of a simple BLO (Sabach & Shtern, 2017) w.r.t. the LL variable y and an unconstrained single-level optimization w.r.t. the UL variable x. This reformulation naturally leads to two foundational problems: The first one involves

P1: Find an optimal LL variable $\hat{y} \in Y^*(\hat{x})$ such that $\varphi(\hat{x}) = f(\hat{x},\hat{y})$ for a given $\hat{x}$.

The second one involves

P2: Find a UL variable $\hat{x}$ that is a stationary point of $\varphi(x)$.

Both problems are easy to solve when the LL function is strongly convex. The lower-level strong convexity (LLSC) ensures $Y^*(x)$ to be a singleton, and therefore simplifies Equation 2 into $\varphi(x) = f(x, y^*(x))$, where the LL optimal solution $y^*(x) = \arg\min_{y \in \mathcal{Y}} g(x, y)$ can be found via gradient descent on g. If we further assume $\mathcal{Y} = \mathbb{R}^{d_y}$, then the implicit function theorem indicates:

$$\nabla \varphi(x) = \nabla_x f(x, y^*(x)) - \nabla_{xy}^2 g(x, y^*(x))[\nabla_{yy}^2 g(x, y^*(x))]^{-1} \nabla_y f(x, y^*(x)). \tag{3}$$

Then one can apply the gradient step with $\nabla \varphi(x)$ to find a UL stationary point. This forms the basis of the classical hyper-objective approaches for BLO with LLSC (Ji et al., 2021). However, these methods heavily rely on the LLSC condition that may not hold in many applications.

This paper investigates BLO with only LL convexity, but without LLSC. Adding a regularization term to the LL function is a natural idea to ensure LLSC (Rajeswaran et al., 2019), but we show in Proposition 4.1 that any small regularization may lead to a large deviation on the hyper-objective. Furthermore, we construct hard instances to illustrate the intractability of BLO without LLSC, for both finding an LL optimal solution and a UL stationary point: Firstly, we prove a lower bound in Proposition 4.2 to show that $\varphi(x)$ is not computable in finite iterations for general convex functions. Secondly, we give a pair of $f(x, y)$ and $g(x, y)$ in Example 4.1 such that the resulting hyper-objective $\varphi(x)$ is discontinuous and thus intractable to optimize.

The constructions of these hard instances rely on the fact that a general convex LL function can be arbitrarily "flat". To avoid the intractability caused by the undesirable "flatness", we introduce two sufficient conditions that can provably confer tractability to BLO with only LL convexity: the gradient dominance condition (Assumption 5.1) and the weak sharp minimum condition (Assumption 5.2). Under these conditions, we propose novel algorithms to find an LL optimal solution and a UL stationary point, with non-asymptotic convergence guarantees:

Finding an LL Optimal Solution. We show that both conditions fall into a general class of the Hölderian error bound condition (Proposition G.1), under which we propose the Switching Gradient Method (SGM, Algorithm 1) to find an LL optimal solution in polynomial time (Theorem 6.1).

Finding a UL Stationary Point. We prove in Proposition 5.1 that both conditions imply the Lipschitz continuity of the solution mapping $Y^*(x)$, which is proved to be both sufficient and necessary for the Lipschitz continuity of $\varphi(x)$ by Proposition 4.3. Under the Lipschitz continuity of $\varphi(x)$, we then propose the Inexact Gradient-Free Method (IGFM, Algorithm 2), which can provably converge to a Goldstein stationary point (Zhang et al., 2020) of the hyper-objective by incorporating SGM as an efficient sub-routine.

We compare the intractability and tractability results under different assumptions on the LL function in Table 1, and summarize our contributions as follows:

1. We formulate the LL optimality and UL stationary as valid criteria for BLO without LLSC (Section 3), which are necessary for an optimistic optimal solution (Dempe et al., 2006).
2. We provide hardness results to show that BLO without LLSC is generally intractable. Our analysis highlights the importance of sharpness in LL functions (Section 4).
3. We prove that when the LL function satisfies either the gradient dominance condition or the weak sharp minimum condition, the hyper-objective $\varphi(x)$ is Lipschitz and thus Clarke differentiable (Section 5).
4. We propose novel polynomial time algorithms for BLO with LL convexity under either the gradient dominance or the weak sharp minimum condition (Section 6).
5. We conduct numerical experiments on adversarial training and hyperparameter tuning that showcase the superiority of our methods (Section 7).

2 Related Works

BLO with LLSC. Approximate implicit differentiation (AID) (Domke, 2012; Ghadimi & Wang, 2018; Pedregosa, 2016; Franceschi et al., 2018; Grazzi et al., 2020; Ji et al., 2021) and iterative differentiation (ITD) (Gould et al., 2016; Franceschi et al., 2017; Shaban et al., 2019; Bolte et al.,

Assumption on LL function	LL Optimality	UL Stationary	Reference
Strongly convex	Tractable	Tractable	Known result
Convex with dominant gradients	Tractable	Tractable	Proved by this work
Convex with weak sharp minimum	Tractable	Tractable	Proved by this work
Only convex	Intractable	Intractable	Proved by this work

Table 1: An overview of the theoretical results in this paper. We show that BLO without LLSC is generally intractable, but becomes tractable when the LL function satisfies either the gradient dominance or the weak sharp minimum condition.

2021) are two representative methods that have non-asymptotically convergence to a UL stationary point for BLO with LLSC. Due to their popularity, many improvements to AID and ITD have also been proposed (Chen et al., 2022; Hong et al., 2023; Yang et al., 2021; Ji & Liang, 2021; Ji et al., 2022; Dagréou et al., 2022).

BLO without LLSC. In the absence of LLSC, Arbel & Mairal (2022) showed that one can extend AID by replacing the inverse in Equation 3 with the Moore-Penrose inverse under the Morse-Bott condition on the manifold $\{y \in \mathbb{R}^{d_y} : \nabla_y f(x,y) = 0\}$. Liu et al. (2021; 2020) extended ITD by proposing various methods to update the LL variable. However, all the methods mentioned above are limited to asymptotic convergence to an LL optimal solution and lack analysis for finding a UL stationary point. Due to the challenge of directly optimizing the hyper-objective, some concurrent works (Liu et al., 2022; Sow et al., 2022) reformulate Problem 1 via the value-function approach and show non-asymptotic convergence to the KKT points of this equivalent problem. However, since classical constraint qualifications provably fail for the reformulated problem (Ye & Zhu, 1995), the KKT condition is even not a necessary condition for a local minimum (Example A.1). In contrast, a UL stationary point is always a necessary condition. We leave a detailed comparison of our hyper-objective approach and value-function approach in Appendix A.

3 Preliminaries

3.1 Notations and Backgrounds

Basic Notation. Throughout this paper, we denote the LL solution mapping as $Y^*(x) = \arg\min_{y\in\mathcal{Y}} g(x,y)$, the LL value function as $g^*(x) = \min_{y\in\mathcal{Y}} g(x,y)$, and the hyper-objective as $\varphi(x) = \min_{y\in Y^*(x)} f(x,y)$. If $\varphi(x)$ has a finite minimum, we denote $\varphi^* = \inf_{x\in\mathbb{R}^{d_x}} \varphi(x)$. We use $\|\cdot\|$ to denote the ℓ_2-norm of a vector, and $z_{[j]}$ to denote the j-th coordinate of vector z. We use $\mathbb{B}_\delta(z) = \{z' : \|z' - z\| \le \delta\}$ to denote the ℓ_2-ball centered at z with radius δ. We let $\sigma_{\max}(A)$ to be the largest singular value of matrix A, and $\sigma^+_{\min}(A)$ to be its smallest non-zero singular value.

Constrained Optimization. To tackle the possible constraint in y, we introduce the definitions of projection and generalized gradient (Nesterov, 2018) as follows.

Definition 3.1 (Projection). *We define the projection onto a set $\mathcal{Y}$ by $\mathcal{P}_\mathcal{Y}(\,\cdot\,) := \arg\min_{y\in\mathcal{Y}} \|y - \cdot\,\|$.*

Definition 3.2 (Generalized Gradient). *For a L-gradient Lipschitz function $g(x,y)$ with $y \in \mathcal{Y}$, we define the generalized gradient with respect to y by $\mathcal{G}_\eta(y;x) := (y - \mathcal{P}_\mathcal{Y}(y - \eta\nabla_y g(x,y)))/\eta$ with some $0 < \eta \le 1/L$.*

Note that the generalized gradient reduced to $\nabla_y g(x,y)$ when $\mathcal{Y} = \mathbb{R}^{d_y}$.

Set-Valued Analysis. A classic notion of distance in set-valued analysis is the Hausdorff distance (Rockafellar & Wets, 2009), formally defined as follows.

Definition 3.3 (Hausdorff Distance). *The Hausdorff distance between two sets S_1, S_2 is defined as*

$$\operatorname{dist}(S_1,S_2) = \max\left\{\sup_{x_1\in S_1}\inf_{x_2\in S_2}\|x_1 - x_2\|, \sup_{x_2\in S_2}\inf_{x_1\in S_1}\|x_1 - x_2\|\right\}$$

This allows us to define the Lipschitz continuity of set-valued mappings as follows.

Definition 3.4. *We call a set-valued mapping $S(x) : \mathbb{R}^{d_1} \rightrightarrows \mathbb{R}^{d_2}$ locally Lipschitz if for any $x \in \mathbb{R}^{d_1}$, there exists $\delta > 0$ and $L > 0$ such that for any $x' \in \mathbb{R}^{d_1}$ satisfying $\|x' - x\| \le \delta$, we have $\operatorname{dist}(S(x), S(x')) \le L\|x - x'\|$. We call $S(x)$ Lipschitz if we can let $\delta \to \infty$.*

Note that the above definition generalizes the Lipschitz continuity for a single-valued mapping.

Nonsmooth Analysis. The following Clarke subdifferential (Clarke, 1990) generalizes both the gradients of differentiable functions and the subgradients of convex functions.

Definition 3.5 (Clarke Subdifferential)**.** *The Clarke subdifferential of a locally Lipschitz function $h(x) : \mathbb{R}^d \to \mathbb{R}$ at a point $x \in \mathbb{R}^d$ is defined by*

$$\partial h(x) := \operatorname{Conv}\left\{ s \in \mathbb{R}^d : \exists x_k \to x, \nabla h(x_k) \to s, \text{ s.t. } \nabla h(x_k) \textit{ exists for all } k \right\}.$$

It can be proved that finding a point with a small Clarke subdifferential is generally intractable for a nonsmooth nonconvex function (Zhang et al., 2020). So we need to consider the following relaxed definition of stationarity for non-asymptotic analysis in nonsmooth nonconvex optimization (Zhang et al., 2020; Tian et al., 2022; Davis et al., 2022; Jordan et al., 2023; Kornowski & Shamir, 2021; Lin et al., 2022; Cutkosky et al., 2023; Kornowski & Shamir, 2023).

Definition 3.6 (Approximate Goldstein Stationary Point)**.** *Given a locally Lipschitz function $h(x) : \mathbb{R}^d \to \mathbb{R}$, we call $x \in \mathbb{R}^d$ a (δ, ε)-Goldstein stationary point if $\min\{\|s\| : s \in \partial_\delta h(x)\} \le \varepsilon$, where $\partial_\delta h(x) := \operatorname{Conv}\left\{ \cup_{x' \in \mathbb{B}_\delta(x)} \partial h(x') \right\}$ is the Goldstein subdifferential (Goldstein, 1977).*

3.2 The Optimality Conditions

This section introduces the optimality conditions for BLO without LLSC used in this paper. Firstly, we recall the definition of the optimistic optimal solution (Dempe et al., 2006), which is a standard optimality condition for the hyper-objective reformulation.

Definition 3.7. *A pair of point (x^*, y^*) is called a locally optimistic optimal solution to Problem 1 if $y^* \in Y^*(x^*)$ and there exists $\delta > 0$ such that we have $\varphi(x^*) \le \varphi(x)$ and $f(x^*, y^*) \le f(x^*, y)$ for all $(x, y) \in \mathbb{B}_\delta(x^*, y^*)$. It is called a globally optimistic optimal solution if we can let $\delta \to \infty$.*

A globally optimistic optimal solution is an exact solution to Problem 1, but its computation is NP-hard since $\varphi(x)$ is generally nonconvex (Danilova et al., 2020). A common relaxation is to find a locally optimistic optimal solution, for which we can derive the following necessary conditions.

Proposition 3.1. *Suppose $f(x, \cdot)$ and $g(x, \cdot)$ are convex, and $\varphi(x)$ is locally Lipschitz. Then for any locally optimistic optimal solution (x^*, y^*), we have $\partial \varphi(x^*) = 0$, $f(x^*, y^*) = \varphi(x^*)$ and $g(x^*, y^*) = g^*(x^*)$.*

It motivates us to use the following criteria for non-asymptotic analysis:

Definition 3.8 (UL Stationary)**.** *Suppose $\varphi(x)$ is locally Lipschitz. We call $\hat{x}$ a (δ, ε)-UL stationary point if it is a (δ, ε)-Goldstein stationary point of $\varphi(x)$.*

Definition 3.9 (LL Optimality)**.** *Fix an x. Suppose $f(x, \cdot)$ and $g(x, \cdot)$ are convex. We call $\hat{y}$ a (ζ_f, ζ_g)-LL optimal solution if we have $|f(x, \hat{y}) - \varphi(x)| \le \zeta_f$ and $g(x, \hat{y}) - g^*(x) \le \zeta_g$.*

The main focus of this paper is to discuss when and how one can design a polynomial time algorithm to achieve the above goals for any given positive precision $\delta, \varepsilon, \zeta_f, \zeta_g$.

Remark 3.1. *In Definition 3.8, we assume that $\varphi(x)$ is locally Lipschitz, which is a regular condition to ensure Clarke differentiability. However, it may not hold for BLO without LLSC, and we will give the sufficient and necessary condition for it later in Proposition 4.3. Definition 3.8 adopts the Goldstein stationary points since $\varphi(x)$ can be nonconvex nonsmooth such that traditional stationary points may be intractable, as we will show later in Example 5.1.*

4 Hardness Results for Intractability

In this section, we provide various hardness results to show the challenges of BLO without LLSC. We first explain why one can not manually regularize the LL function to ensure the LLSC condition. Subsequently, we demonstrate that both the tasks of finding an LL optimal solution and finding a UL stationary point can be intractable for BLO without LLSC.

4.1 Can Regularization Help?

One natural way to tackle BLO without LLSC is to add some small quadratic terms and then apply an algorithm designed under LLSC (Rajeswaran et al., 2019). However, we show that the regularization transforms $Y^*(x)$ from a set to a singleton, thus breaking the original problem structure.

Proposition 4.1. *Given a pivot $\hat{y}$, there exists a BLO instance, where both $f(x,y)$ and $g(x,y)$ are convex in y, and the resulting hyper-objective $\varphi(x)$ is a quadratic function, but for any $\lambda > 0$ the regularized hyper-objective*

$$\varphi_\lambda(x) = \min_{y \in Y_\lambda^*(x)} f(x,y), \quad Y_\lambda^*(x) = \arg\min_{y \in \mathcal{Y}} g(x,y) + \lambda \|y - \hat{y}\|^2$$

is a linear function with $|\inf_{x \in \mathbb{R}^{d_x}} \varphi_\lambda(x) - \inf_{x \in \mathbb{R}^{d_x}} \varphi(x)| = \infty$.

This example indicates that even if the regularization is arbitrarily small, the hyper-objective before and after regularization can be completely different objectives. Consequently, BLO without LLSC should be treated as a distinct research topic from BLO with LLSC.

4.2 Can we Find an LL Optimal Solution?

The goal of finding an LL optimal solution for a given $x \in \mathbb{R}^{d_x}$ is to solve the following problem:

$$\min_{y \in Y^*(x)} f(x,y), \quad Y^*(x) = \arg\min_{y \in \mathcal{Y}} g(x,y). \tag{4}$$

This problem is usually called simple BLO (Beck & Sabach, 2014; Sabach & Shtern, 2017; Kaushik & Yousefian, 2021) since it involves only one variable y. However, it is not a "simple" problem as the following theorem shows its intractability for general convex objectives.

Proposition 4.2. *Fix an x. For any $K \in \mathbb{N}^+$, there exists $d_y \in \mathbb{N}^+$, such that for any $y_0 \in \mathbb{R}^{d_y}$, there exists a pair of functions $f(x,\cdot), g(x,\cdot)$ that are both convex and 1-gradient Lipschitz , for any first-order algorithm $\mathcal{A}$ which initializes from $y_0 \in \mathcal{Y}$ with* $\operatorname{dist}(y_0, y^*(x)) \le 1$ *and generates a sequence of test points $\{y_k\}_{k=0}^K$ with*

$$y_k \in y_0 + \operatorname{Span}\{\nabla_y f(x,y_0), \nabla_y g(x,y_0), \cdots, \nabla_y f(x,y_{k-1}), \nabla_y g(x,y_{k-1})\}, \quad k \ge 1,$$

it holds that $|f(x,y_k) - \varphi(x)| \ge 1/4$, where $y^*(x)$ is the unique solution to $\min_{y \in Y^*(x)} f(x,y)$.

The key idea in the proof is to construct the LL function using the worst-case convex zero chain (Nesterov, 2018), such that any first-order algorithm will require a large number of steps to approach the vicinity of the LL solution mapping $Y^*(x)$. The proof is provided in Appendix D

4.3 Can we Find a UL Stationary Point?

Besides the difficulty in finding an LL optimal solution, the goal of finding a UL stationary point is also challenging. Below, we show that the hyper-objective $\varphi(x)$ can be discontinuous without LLSC. Since continuity is one of the basic assumptions for almost all numerical optimization schemes (Nocedal & Wright, 1999), our hard instance indicates that $\varphi(x)$ may be intrinsically intractable to optimize for BLO without LLSC.

Example 4.1. *Consider a BLO instance given by*

$$\min_{x \in \mathbb{R}, y \in Y^*(x)} x^2 + y, \quad Y^*(x) = \arg\min_{y \in [-1,1]} -xy.$$

The resulting hyper-objective $\varphi(x) = x^2 + \mathbb{I}[x > 0]$ is discontinuous at $x = 0$.

In the above example, the discontinuity of $\varphi(x)$ comes from the discontinuity of $Y^*(x) = \operatorname{sign}(x)$. Below, we prove that this statement and its reverse generally holds.

Proposition 4.3. *Suppose the solution mapping $Y^*(x)$ is non-empty and compact for any $x \in \mathbb{R}^{d_x}$.*

a. *If $f(x,y)$ and $Y^*(x)$ are locally Lipschitz , then $\varphi(x)$ is locally Lipschitz.*

b. *Conversely, if $\varphi(x)$ is locally Lipschitz for any locally Lipschitz function $f(x,y)$, then $Y^*(x)$ is locally Lipschitz.*

c. *If $f(x,y)$ is C_f-Lipschitz and $Y^*(x)$ is κ-Lipschitz, then $\varphi(x)$ is C_φ-Lipschitz with coefficient $C_\varphi = (\kappa+1)C_f$.*

d. *Conversely, if $\varphi(x)$ is C_φ-Lipschitz for any C_f-Lipschitz function $f(x,y)$, then $Y^*(x)$ is κ-Lipschitz with coefficient $\kappa = C_\varphi / C_f$.*

Local Lipschitz continuity ensures UL stationary points (Definition 3.8) are well-defined, while global Lipschitz continuity enables uniform complexity bounds for non-asymptotic analysis (as we will use in Section 6.2). According to the above theorem, ensuring the continuity of $Y^*(x)$ is the key to obtaining the desired continuity of $\varphi(x)$. This motivates us to focus on well-behaved LL functions that confer continuity of $Y^*(x)$.

5 Sufficient Conditions for Tractability

5.1 Regularity Conditions for Continuity

Since the constructions of the hard instances in the previous section all rely on very *flat* LL functions, our results underscore that *sharpness* of LL functions is essential to ensure the tractability of BLO. This observation inspires us to focus on more restricted function classes that possess sharpness to circumvent the ill-conditioned nature of BLO without LLSC. Below, we introduce two conditions that correspond to different degrees of sharpness.

Assumption 5.1 (Gradient Dominance)**.** *Suppose $g(x,y)$ is L-gradient Lipschitz jointly in (x,y), and there exists $\alpha > 0$ such that for any $x \in \mathbb{R}^{d_x}$, $y \in \mathcal{Y}$ we have $\mathcal{G}_{1/L}(y;x) \geq \alpha \operatorname{dist}(y, Y^*(x))$.*

Assumption 5.2 (Weak Sharp Minimum)**.** *Suppose $g(x,y)$ is L-Lipschitz in x, and there exists $\alpha > 0$ such that for any $x \in \mathbb{R}^{d_x}$, $y \in \mathcal{Y}$ we have $g(x,y) - g^*(x) \geq 2\alpha \operatorname{dist}(y, Y^*(x))$.*

Both conditions are widely used in convex optimization (Burke & Ferris, 1993; Drusvyatskiy & Lewis, 2018). They are milder conditions than LLSC by allowing $Y^*(x)$ to be non-singleton. Despite being more relaxed, we demonstrate below that either of them can lead to the continuity of $Y^*(x)$ and thus $\varphi(x)$. The continuity of $\varphi(x)$ is crucial for designing algorithms to optimize it.

Proposition 5.1. *Under Assumption 5.1 or 5.2, $Y^*(x)$ is (L/α)-Lipschitz.*
Furthermore, if $f(x,y)$ is C_f-Lipschitz, then $\varphi(x)$ is $(L/\alpha+1)C_f$-Lipschtz.

Therefore, the introduced conditions can avoid discontinuous instances such as Example 4.1. It is worth noting that these conditions fundamentally differ from LLSC, as $\varphi(x)$ can be nonsmooth under these conditions, as exemplified below. The potential nonsmoothness of $\varphi(x)$ further justifies the rationality of using Goldstein stationarity in Definition 3.8.

Example 5.1. *Let $f(x,y) = xy$, $g(x,y) = 0$ and $\mathcal{Y} = [-1,1]$. We obtain a BLO instance satisfying both Assumption 5.1 and 5.2. But the resulting $\varphi(x) = -|x|$ is nonsmooth and nonconvex.*

5.2 How to Verify the Conditions?

One may wonder how to verify the introduced conditions in applications. It is non-trivial as the value of $\operatorname{dist}(y, Y^*(x))$ is unknown. An easy case is Assumption 5.1 with $\mathcal{Y} = \mathbb{R}^{d_y}$, which reduces to the Polyak-Łojasiewicz condition (Polyak, 1963): $\|\nabla_y g(x,y)\|^2 \geq 2\alpha(g(x,y) - g^*(x))$ by Theorem 2 in Karimi et al. (2016). This inequality allows us to identify the following examples that fall into Assumption 5.1. Firstly, we can show that Assumption 5.1 strictly covers the LLSC condition.

Example 5.2. *If g is L-gradient Lipschitz and α-strongly convex, then it satisfies Assumption 5.1.*

Secondly, the following example that both AID and ITD fail to optimize satisfies Assumption 5.1.

Example 5.3. *Consider the hard BLO instance proposed by Liu et al. (2020):*

$$\min_{x \in \mathbb{R}, y \in Y^*(x)} (x - y_{[2]})^2 + (y_{[1]} - 1)^2, \quad Y^*(x) = \arg\min_{y \in \mathbb{R}^2} y_{[1]}^2 - 2xy_{[1]}.$$

The LL function satisfies Assumption 5.1 with $L = 2$ and $\alpha = 2$.

Thirdly, the BLO with least squares loss studied by Bishop et al. (2020) also satisfies Assumption 5.1. We leave more details of this model and its application in adversarial training in Section 7.1.

Example 5.4. *Consider the BLO with least squares loss:*

$$\min_{x\in\mathbb{R}^{d_x}, y\in Y^*(x)} \frac{1}{2n}\|Ax-y\|^2, \quad Y^*(x) = \arg\min_{y\in\mathbb{R}^n} \frac{1}{2n}\|Ax-y\|_M^2 + \frac{\lambda}{2n}\|y-b\|_M^2,$$

where $A \in \mathbb{R}^{n\times d_x}$, $b\in\mathbb{R}^n$ represents the features and labels of the n samples in the dataset, $\lambda > 0$ and M is a positive semi-definite matrix that induces the norm $\|z\|_M = \sqrt{z^\top Mz}$. The LL function satisfies Assumption 5.1 with $L = (\lambda+1)\sigma_{\max}(M)$ and $\alpha = (\lambda+1)\sigma_{\min}^{+}(M)$.

6 The Proposed Methods

In this section, we propose novel polynomial time algorithms for BLO under Assumption 5.1 and 5.2. In Section 6.1, we borrow ideas from switching gradient methods to overcome the difficulty of multiple LL minima. In Section 6.2 we propose a method motivated by zeroth-order optimization that can provably converge to a UL stationary point.

6.1 Finding an LL Optimal Solution via Switching Gradient Method

Algorithm 1 SGM $(x, y_0, K_0, K, \tau, \theta)$

```
$\mathcal{I} = \emptyset$,  $\hat{y}_0 = y_0$
for $k = 0, 1, \cdots, K_0 - 1$
    $\hat{y}_{k+1} = \mathcal{P}_{\mathcal{Y}}(\hat{y}_k - \tau\partial_y g(x, \hat{y}_k))$
end for
$\hat{g}^*(x) = g(x, \hat{y}_{K_0})$
for $k = 0, 1, \cdots, K - 1$
    if $g(x, y_k) - \hat{g}^*(x) \leq 2\theta$
        $y_{k+1} = \mathcal{P}_{\mathcal{Y}}(y_k - \tau\partial_y f(x, y_k))$
        $\mathcal{I} = \mathcal{I} \cup \{k\}$
    else
        $y_{k+1} = \mathcal{P}_{\mathcal{Y}}(y_k - \tau\partial_y g(x, y_k))$
end for
$y_{\mathrm{out}} = \frac{1}{|\mathcal{I}|}\sum_{k\in\mathcal{I}} y_k$
return $y_{\mathrm{out}}$
```

In Equation 4, the LL constraint $y \in Y^*(x)$ is equivalent to an inequality constraint $g(x,y) \leq g^*(x)$. Based on this observation, we generalize Polyak's Switching Gradient Method (Polyak, 1967) for functional constrained problems to Algorithm 1 when the following assumptions hold.

Assumption 6.1. *Suppose that*

a. both $f(x,y)$ and $g(x,y)$ are convex in y;

b. $\mathcal{Y}$ is compact with diameter R;

c. $f(x,y)$ is C_f-Lipschitz on $\mathbb{R}^{d_x}\times\mathcal{Y}$;

d. $g(x,\cdot\,)$ is C_g-Lipschitz on $\mathcal{Y}$ for any $x\in\mathbb{R}^{d_x}$;

e. either Assumption 5.1 or 5.2 holds for $g(x,y)$.

Under the above assumptions, we can prove the following result.

Theorem 6.1. *Fix an x. Under Assumption 6.1, Algorithm 1 with appropriate parameters can ouput a point y_{out} satisfying $|f(x,y_{\mathrm{out}}) - \varphi(x)| \leq \zeta$ and $g(x,y_{\mathrm{out}}) - g^*(x) \leq \zeta$ with $\mathcal{O}(\mathrm{poly}(1/\zeta))$ first-order oracle calls from g.*

The corresponding proof and specific parameters of the algorithm can be found Appendix G.

Algorithm 2 IGFM $(x_0, y_0, \eta, T, \delta, K_0, K, \tau, \theta)$

```
Require: Sub-routine $\mathcal{A}$ can estimate $\tilde{\varphi}(x) \approx \varphi(x)$ for any $x \in \mathbb{R}^{d_x}$
for $t = 0, 1, \cdots, T-1$
    Sample $u_t \in \mathbb{R}^{d_x}$ uniformly from the unit sphere in $\mathbb{R}^{d_x}$.
    Estimate $\tilde{\varphi}(x_t + \delta u_t)$ and $\tilde{\varphi}(x_t - \delta u_t)$ by sub-routine $\mathcal{A}$.
    $\hat{\nabla}_t = \frac{d_x}{2\delta}\left(\tilde{\varphi}(x_t + \delta u_t) - \tilde{\varphi}(x_t - \delta u_t)\right) u_t$
    $x_{t+1} = x_t - \eta \hat{\nabla}_t$
end for
return $x_{\rm out}$ uniformly chosen from $\{x_t\}_{t=0}^{T-1}$
```

6.2 Finding a UL Stationary Point via Zeroth-Order Method

Without LLSC, the hyper-gradient $\nabla \varphi(x)$ may not have an explicit form as Equation 3. To tackle this challenge, we propose the Inexact Gradient-Free Method (IGFM) in Algorithm 2. The algorithm is motivated by recent advances in nonsmooth nonconvex zeroth-order optimization (Lin et al., 2022). Our zeroth-order oracle $\tilde{\varphi}(x) \approx \varphi(x)$ is "inexact" since it is an approximation from a sub-routine $\mathcal{A}$. Below, we show that when $\mathcal{A}$ can guarantee sufficient approximation precision, the IGFM can provably find a Goldstein stationary point of a Lipschitz hyper-objective function $\varphi(x)$.

Assumption 6.2. *Suppose that*

a. $\varphi(x)$ is C_φ-Lipschitz.

b. $\mathcal{A}$ ensures $|\tilde{\varphi}(x) - \varphi(x)| \le \mathcal{O}(\delta \varepsilon^2 / (d_x C_\varphi))$ for any $x \in \mathbb{R}^{d_x}$.

Theorem 6.2. *Given any $\varepsilon \lesssim C_f$. Let $\Delta = \varphi(x_0) - \varphi^*$. Under Assumption 6.2, set*

$$T = \mathcal{O}\left(d_x^{3/2}\left(\frac{C_\varphi^4}{\varepsilon^4} + \frac{\Delta C_\varphi^3}{\delta \varepsilon^4}\right)\right), \ \eta = \Theta\left(\sqrt{\frac{\delta(\Delta + \delta C_\varphi)}{d_x^{3/2} C_\varphi^3 T}}\right). \tag{5}$$

Then Algorithm 2 can output a point $x_{\rm out}$ that satisfies $\mathbb{E} \min \{\|s\| : s \in \partial_\delta \varphi(x_{\rm out})\} \le \varepsilon$.

Now it remains to verify Assumption 6.2: Assumption 6.2a can be satisfied by Proposition 4.3, while Assumption 6.2b can be satisfied by Theorem 6.1. Therefore, we have the following result.

Corollary 6.1. *Suppose Assumption 6.1 holds. Set $\mathcal{A}$ as Algorithm 1. Then Algorithm 2 with appropriate parameters can output a (δ, ϵ)-Goldstein stationary point of $\varphi(x)$ in expectation within $\mathcal{O}(\mathrm{poly}(d_x, 1/\varepsilon, 1/\delta))$ zeroth-order and first-order oracle calls from f and g.*

To the best of our knowledge, it is the first theoretical analysis that shows the non-asymptotic convergence to a UL stationary point for BLO without LLSC.

7 Numerical Experiments

Table 2: MSE (mean ± std) achieved by different algorithms on the "abalone" dataset in adversarial training.

Method	MSE
AID	1.781 ± 0.418
ITD	0.982 ± 0.015
BGS	0.995 ± 0.259
BDA	0.976 ± 0.014
BOME	0.999 ± 0.140
IA-GM	0.992 ± 0.013
IGFM (Ours)	$\mathbf{0.936 \pm 0.015}$

Table 3: Test accuracy (%) achieved by different algorithms on the MNIST dataset under different corruption rates p in hyperparameter tuning.

Method	$p = 0.5$	$p = 0.3$	$p = 0.1$
AID	75.8	87.5	91.3
ITD	75.8	87.5	91.3
BGS	75.8	87.5	91.3
BDA	81.2	89.3	91.5
BOME	86.7	88.9	89.3
IA-GM	86.9	90.3	90.5
IGFM (Ours)	**88.4**	**91.0**	**91.8**

In this section, we compare IGFM with different baselines, including AID with conjugate gradient (Maclaurin et al., 2015), ITD (Ji et al., 2021), BGS (Arbel & Mairal, 2022), BDA (Liu et al., 2020), BOME (Liu et al., 2022), and IA-GM (Liu et al., 2021) in the following two different applications of BLO without LLSC.

7.1 Adversarial Training

Brückner & Scheffer (2011) proposed modeling adversarial training via BLO. In this model, the learner aims at finding the optimal parameter x, subject to data y being modified by an adversarial data provider. Like Bishop et al. (2020); Wang et al. (2021; 2022a), we use least squares loss for both f and g as Example 5.4. In the LL loss, we use a diagonal matrix M to assign different weights to each sample, and a ridge term $\|y-b\|_M^2$ to penalize the data provider when manipulating the original labels b. We set half the diagonal elements of M evenly in $[\sigma_{\min}^+, \sigma_{\max}]$ and the rest zero. We let $\lambda = 1$, $\sigma_{\max} = 1$ and $\sigma_{\min}^+ = 10^{-9}$. For BDA, we choose $s_u = s_l = 1$, $\alpha_k = \mu/(k+1)$ and tune μ from $\{0.1, 0.5, 0.9\}$ as Liu et al. (2020). For BOME, we choose the default option for ϕ_k and η from $\{0.9, 0.5, 0.1\}$ as Liu et al. (2022). For IGFM, we choose $\delta = 10^{-3}$ and tune θ from $\{10^{-1}, 10^{-2}, 10^{-3}\}$. For all algorithms, we tune the learning rates in $\{10^2, 10^1, 10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. We run all the algorithms for 500 UL iterations, with 10 LL iterations per UL iteration. Table 2 compares the mean squared error (MSE), measured by the value of $\varphi(x)$, achieved by the algorithms on the "abalone" dataset from LIBSVM (Chang & Lin, 2011). AID has poor performance because it requires taking the inverse of $\nabla_{yy}^2 g(x,y)$, which is ill-conditioned in this experiment. Among all the algorithms, the IGFM achieves the lowest mean value of MSE, and its variance is also maintained at a relatively low level.

7.2 Hyperparameter Tuning

We consider tuning the optimal ℓ_2 regularization for logistic regression to avoid overfitting a noisy training set $\mathcal{D}^{\rm tr}$, based on the performance on a clean validation set $\mathcal{D}^{\rm val}$. We let the UL variable x be the log-transformed regularization coefficient to avoid the constraint $x \geq 0$ (Pedregosa, 2016; Bertrand et al., 2020), and the LL variable y be the weight of the model. The problem can be formulated as BLO with:

$$\begin{aligned} f(x,y) &= \frac{1}{|\mathcal{D}^{\rm val}|} \sum_{(a_i,b_i)\in\mathcal{D}^{\rm val}} \ell(\langle a_i, y\rangle, b_i), \\ g(x,y) &= \frac{1}{|\mathcal{D}^{\rm tr}|} \sum_{(a_i,b_i)\in\mathcal{D}^{\rm tr}} \ell(\langle a_i, y\rangle, b_i) + \exp(x)\|y\|^2, \end{aligned}$$

where (a_i, b_i) is the i-th feature-label pair in the dataset, and $\ell(\,\cdot\,,\,\cdot\,)$ is the cross-entropy loss. We use the MNIST dataset (LeCun, 1998) in this experiment. We use 40,000 images for $\mathcal{D}^{\rm tr}$ and 20,000 images for $\mathcal{D}^{\rm val}$. We corrupt $\mathcal{D}^{\rm tr}$ by assigning random labels with probability p (Liu et al., 2022). We follow the same hyperparameter selection strategy as Section 7.1, and run all the algorithms with 100 UL iterations. Table 3 reports the accuracy evaluated on the testing set with 10,000 images under different levels of p. It can be seen that IGFM achieves the highest accuracy among all algorithms. Note that AID / ITD / BGS have similar performances since AID and ITD are proven to be consistent under LLSC (Ji et al., 2021) and BGS is a combination of them.

8 Conclusions and Discussions

This paper gives a comprehensive study of BLO without the typical LLSC assumption. We provide hardness results to show the intractability of this problem and introduce several key regularity conditions that can confer tractability. Novel algorithms with non-asymptotic convergence are proposed as well. Experiments on real-world datasets support our theoretical investigations.

Although this paper focuses primarily on the theoretical level, we expect our theory can shed light on efficient algorithm design for BLO applications in practice. We also hope our work can be a good starting point for non-asymptotic analysis for more challenging BLO problems, such as BLO with nonconvex LL functions or BLO with intertwined inequality constraints $h(x,y) \leq 0$.

REFERENCES

Michael Arbel and Julien Mairal. Non-convex bilevel games with critical point selection maps. In *NeurIPS*, 2022.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Amir Beck and Shoham Sabach. A first order method for finding minimal norm-like solutions of convex optimization problems. *Mathematical Programming*, 147(1):25–46, 2014.

Quentin Bertrand, Quentin Klopfenstein, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization. In *ICML*, 2020.

Nicholas Bishop, Long Tran-Thanh, and Enrico Gerding. Optimal learning from verified training data. In *NeurIPS*, 2020.

Jérôme Bolte, Tam Le, Edouard Pauwels, and Tony Silveti-Falls. Nonsmooth implicit differentiation for machine-learning and optimization. In *NeurIPS*, 2021.

Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *SIGKDD*, 2011.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

James V. Burke and Michael C. Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5):1340–1359, 1993.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology*, 2(3):1–27, 2011.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. A single-timescale stochastic bilevel optimization method. In *AISTATS*, 2022.

Frank H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Christian Clason. Nonsmooth analysis and optimization. *arXiv preprint arXiv:1708.04180*, 2017.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *ICML*, 2023.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *NeurIPS*, 2022.

Marina Danilova, Pavel Dvurechensky, Alexander Gasnikov, Eduard Gorbunov, Sergey Guminov, Dmitry Kamzolov, and Innokentiy Shibaev. Recent theoretical advances in non-convex optimization. *arXiv preprint arXiv:2012.06188*, 2020.

Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. In *NeurIPS*, 2022.

Stephan Dempe. *Foundations of bilevel programming*. Springer Science & Business Media, 2002.

Stephan Dempe and Alain Zemkoho. Bilevel optimization. *Springer optimization and its applications*, 161, 2020.

Stephan Dempe, Vyatcheslav V. Kalashnikov, and Nataliya Kalashnykova. Optimality conditions for bilevel programming problems. *Optimization with Multivalued Mappings: Theory, Applications, and Algorithms*, pp. 3–28, 2006.

Justin Domke. Generic methods for optimization-based modeling. In *AISTATS*, 2012.

Dmitriy Drusvyatskiy and Adrian S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *ICML*, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *ICML*, 2018.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

A. Goldstein. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13(1): 14–22, 1977.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *ICML*, 2020.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5149–5169, 2021.

Kaiyi Ji and Yingbin Liang. Lower bounds and accelerated algorithms for bilevel optimization. *JMLR*, 23:1–56, 2021.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *ICML*, 2021.

Kaiyi Ji, Mingrui Liu, Yingbin Liang, and Lei Ying. Will bilevel optimizers benefit from loops. In *NeurIPS*, 2022.

Ruichen Jiang, Nazanin Abolfazli, Aryan Mokhtari, and Erfan Yazdandoost Hamedani. A conditional gradient-based method for simple bilevel optimization with convex lower-level problem. In *AISTATS*, 2023.

Yibo Jiang and Victor Veitch. Invariant and transportable representations for anti-causal domain shifts. In *NeurIPS*, 2022.

Michael I. Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *COLT*, 2023.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, 2016.

Harshal D. Kaushik and Farzad Yousefian. A method with convergence rates for optimization problems with variational inequality constraints. *SIAM Journal on Optimization*, 2021.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In *NeurIPS*, 1999.

Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. In *NeurIPS*, 2021.

Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *arXiv preprint arXiv:2307.04504*, 2023.

Yann LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, 2020a.

Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. In *COLT*, 2020b.

Tianyi Lin, Zeyu Zheng, and Michael I. Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In *NeurIPS*, 2022.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. BOME! bilevel optimization made easy: A simple first-order approach. In *NeurIPS*, 2022.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *ICLR*, 2019.

Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *ICML*, 2020.

Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. *NeurIPS*, 2021.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *ICML*, 2015.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer, 1999.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *NeurIPS*, 2019.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *ICML*, 2016.

Quang Pham, Chenghao Liu, Doyen Sahoo, and HOI Steven. Contextual transformation networks for online continual learning. In *ICLR*, 2021.

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

Boris Teodorovich Polyak. A general method for solving extremal problems. In *Doklady Akademii Nauk*, volume 174, pp. 33–36. Russian Academy of Sciences, 1967.

Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *NeurIPS*, 2019.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

R. Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Shoham Sabach and Shimrit Shtern. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *AISTATS*, 2019.

Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifying some distributional robustness with principled adversarial training. In *ICLR*, 2018.

Daouda Sow, Kaiyi Ji, Ziwei Guan, and Yingbin Liang. A constrained optimization approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022.

Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of lipschitz functions. In *ICML*, 2022.

Jiali Wang, He Chen, Rujun Jiang, Xudong Li, and Zihao Li. Fast algorithms for stackelberg prediction game with least squares loss. In *ICML*, 2021.

Jiali Wang, Wen Huang, Rujun Jiang, Xudong Li, and Alex L. Wang. Solving stackelberg prediction game with least squares loss via spherically constrained least squares reformulation. In *ICML*, 2022a.

Xiaoxing Wang, Wenxuan Guo, Jianlin Su, Xiaokang Yang, and Junchi Yan. ZARTS: On zero-order optimization for neural architecture search. In *NeurIPS*, 2022b.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *NeurIPS*, 2021.

Jane J. Ye and DL Zhu. Optimality conditions for bilevel programming problems. *Optimization*, 33 (1):9–27, 1995.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonsmooth nonconvex functions. In *ICML*, 2020.

Miao Zhang, Steven W. Su, Shirui Pan, Xiaojun Chang, Ehsan M Abbasnejad, and Reza Haffari. iDARTS: Differentiable architecture search with stochastic implicit gradients. In *ICML*, 2021.

Xiao Zhou, Yong Lin, Renjie Pi, Weizhong Zhang, Renzhe Xu, Peng Cui, and Tong Zhang. Model agnostic sample reweighting for out-of-distribution learning. In *ICML*, 2022.

Barret Zoph and Quoc Le. Neural architecture search with reinforcement learning. In *ICLR*, 2016.

A Limitations of Value-Function Approach

In contrast to the hyper-objective approach adopted in this paper that pursues a UL stationary point such that $\|\nabla \varphi(x)\| \le \varepsilon$, existing non-asymptotic analysis (Liu et al., 2022; Sow et al., 2022) for BLO without LLSC relies on following value-function reformulation for Problem 1:

$$\min_{x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}} f(x,y) \quad \text{s.t.} \quad g(x,y) - g^*(x) \le 0. \tag{6}$$

These value-function approaches show convergence to the following KKT points.

Definition A.1 (KKT point). *Suppose that $g^*(x)$ is Clarke subdifferentiable. We call (x,y) an ε-KKT point of Problem 6 if there exists a scalar $\lambda \ge 0$ such that we have*

a. *(Stationary in x) $\|\nabla_x f(x,y) + \lambda(\nabla_x g(x,y) - \partial g^*(x))\| \le \varepsilon$.*

b. *(Stationary in y) $\|\nabla_y f(x,y) + \lambda \nabla_y g(x,y)\| \le \varepsilon$.*

c. *(Feasibility) $g(x,y) - g^*(x) \le \varepsilon$.*

d. *(Complementary Slackness) $|\lambda(g(x,y) - g^*(x))| \le \varepsilon$.*

We call (x,y) a KKT point if $\varepsilon = 0$.

Remark A.1. *In Definition A.1 we assume that $g^*(x)$ is Clarke differentiable. It can be easily satisfied under some mild conditions. For instance, when $g(x,y)$ is L-gradient Lipschitz, $g^*(x)$ is provably L-weakly concave, and thus Clarke differentiable (Rockafellar & Wets, 2009). In the unconstrained case that $\mathcal{Y} = \mathbb{R}^{d_y}$, under LLSC or more generally under Assumption 5.1, $g^*(x)$ is provably differentiable (Nouiehed et al., 2019) and the Clarke subdifferential $\partial g^*(x)$ reduces to the classical gradient $\nabla g^*(x)$.*

Unfortunately, classical constraint qualifications provably fail for the value-function-based reformulation (Ye & Zhu, 1995). For this reason, we can easily construct a BLO instance whose KKT points do not contain the optimal solution even under LLSC.

Example A.1. *Consider a BLO instance given by:*

$$\min_{x \in \mathbb{R}, y \in \mathbb{R}} -xy, \quad \text{s.t. } (x+y-2)^2 \le 0,$$

where the lower-level function is strongly convex in y. For this example:

a. *The stationary point of $\varphi(x)$ is exactly the global solution x^*.*

b. *However, the KKT points by Definition A.1 do not include any solution to this problem.*

Proof. We know that the lower-level constraint is $y = 2 - x$, so the problem is equivalent to $\min_{x \in \mathbb{R}} x^2 - 2x$ with the unique solution $(x^*, y^*) = (1,1)$. However, if we rewrite the problem by:

$$\min_{x \in \mathbb{R}, y \in \mathbb{R}} -xy, \quad \text{s.t. } (x+y-2)^2 \le 0.$$

The KKT condition is

$$\begin{cases} y - 2\lambda(x+y-2) = 0; \\ x - 2\lambda(x+y-2) = 0; \\ \lambda(x+y-2)^2 = 0; \\ (x+y-2)^2 \le 0, \quad \lambda \ge 0. \end{cases}$$

When $\lambda > 0$ there is no (x,y) that satisfies the KKT condition. When $\lambda = 0$, the KKT condition is only satisfied by $(x,y) = (0,0)$, but it is not the solution to this problem. □

One may argue that when relaxing the goal into finding an ε-KKT point, Slater's constraint qualification can be satisfied since we allow the constraint $g(x,y) - g^*(x) \le 0$ to be violated slightly. However, we give a concrete example indicating that an ε-KKT point may be far away from the solution set, even when the hyper-objective $\varphi(x)$ is strongly convex.

Example A.2. *Given $0 < \varepsilon \leq 1$. Suppose $\varphi(x)$ is μ-strongly convex with a unique solution x^*.*

a. Whenever a given point x satisfies $\|\nabla\varphi(x)\| \leq \varepsilon$, we have $\|x - x^*\| \leq \varepsilon/\mu$.

b. However, there exists a BLO instance with a convex lower-level function such that the resulting $\varphi(x)$ is strongly convex, but there is an infinite number of 2ε-stationary points (x, y) by Definition A.1 such that $\|x - x^*\| = 1$.

Proof. Below we prove the two parts separately.

a. Strong convexity ensures that $\mu\|x - x^*\| \leq \|\nabla\varphi(x)\|$.

b. Consider the bilevel problem given by:

$$\min_{x\in\mathbb{R},y\in\mathbb{R}} x^2 - 2\varepsilon xy, \quad \text{s.t. } y \in \arg\min_{y\in\mathbb{R}} \varepsilon^3 y^2,$$

where the lower-level problem is convex in y and the global solution is $x^* = 0$. It can be verified that $(x, y) = (1, \varepsilon^{-1})$ is an ε-KKT point with any multiplier satisfying $0 < \lambda \leq 1$ by

$$\begin{cases} g(x, y) - g^*(x) = \varepsilon^3 y^2 = \varepsilon; \\ |\nabla_x f(x, y) + \lambda(\nabla_x g(x, y) - \nabla g^*(x))| = 2(x - \varepsilon y) = 0; \\ |\nabla_y f(x, y) + \lambda\nabla_y g(x, y)| = 2(\varepsilon x - \lambda\varepsilon^3 y) \leq 2\varepsilon. \end{cases}$$

But we know that $\|x - x^*\| = 1$. □

B Proofs in Section 3

Proposition 3.1. *Suppose $f(x, \cdot)$ and $g(x, \cdot)$ are convex, and $\varphi(x)$ is locally Lipschitz. Then for any locally optimistic optimal solution (x^*, y^*), we have $\partial\varphi(x^*) = 0$, $f(x^*, y^*) = \varphi(x^*)$ and $g(x^*, y^*) = g^*(x^*)$.*

Proof. By the definition that $y^* \in Y^*(x)$, we know that $g(x^*, y^*) = g^*(x^*)$. When $g(x.\cdot)$ is convex, we know that $Y^*(x)$ is also a convex set for any given x. Then the problem $\min_{y\in Y^*(x)} f(x, y)$ is a convex problem with respect to y, where a local minimum is also a global minimum. This indicates that $\varphi(x^*) = f(x^*, y^*)$. Finally, the first-order necessary optimality condition for a local minimum of $\varphi(x)$ implies that $\partial\varphi(x^*) = 0$ (Theorem 8.4 by Clason (2017)).

□

C Proofs in Section 4.1

Proposition 4.1. *Given a pivot $\hat{y}$, there exists a BLO instance, where both $f(x, y)$ and $g(x, y)$ are convex in y, and the resulting hyper-objective $\varphi(x)$ is a quadratic function, but for any $\lambda > 0$ the regularized hyper-objective*

$$\varphi_\lambda(x) = \min_{y\in Y_\lambda^*(x)} f(x, y), \quad Y_\lambda^*(x) = \arg\min_{y\in\mathcal{Y}} g(x, y) + \lambda\|y - \hat{y}\|^2$$

is a linear function with $|\inf_{x\in\mathbb{R}^{d_x}} \varphi_\lambda(x) - \inf_{x\in\mathbb{R}^{d_x}} \varphi(x)| = \infty$.

Proof. We distinguish two different cases by whether we have $\hat{y}_{[1]} = 0$.

When $\hat{y}_{[1]} \neq 0$, we consider the problem given by

$$\min_{x\in\mathbb{R},y\in Y^*(x)} y_{[1]}^2 - 2xy_{[1]}, \quad Y^*(x) = \arg\min_{y\in\mathbb{R}^2} (y_{[2]} - \hat{y}_{[2]})^2.$$

After adding regularization, we have $Y_\lambda^*(x) = \{\hat{y}\}$ and $\varphi_\lambda(x) = \hat{y}_{[1]}^2 - 2x\hat{y}_{[1]}$.

When $\hat{y}_{[1]} = 0$, we instead consider the problem given by

$$\min_{x\in\mathbb{R},y\in Y^*(x)} (y_{[1]} + 1)^2 - 2x(y_{[1]} + 1), \quad Y^*(x) = \arg\min_{y\in\mathbb{R}^2} (y_{[2]} - \hat{y}_{[2]})^2.$$

And after adding regularization we have $Y_\lambda^*(x) = \{0\}$ and $\varphi_\lambda(x) = 1 - 2x$.

However, for both the two cases the original hyper-objective is the quadratic function $\varphi(x) = -x^2$. □

D PROOFS IN SECTION 4.2

Our lower bound is based on the following first-order zero-chain (Nesterov, 2018).

Definition D.1 (Zero-Chain)**.** *We call function $h(z) : \mathbb{R}^q \to \mathbb{R}$ a first-order zero-chain if for any sequence $\{z_k\}_{k\geq 1}$ satisfying*

$$z_i \in \mathrm{Span}\{\partial h(z_0), \cdots, \partial h(z_{i-1})\}, \quad i \geq 1; \quad z_0 = 0 \tag{7}$$

it holds that $z_{i,[j]} = 0$, $i + 1 \leq j \leq q$.

We remark that in the construction of a zero chain, we can always assume $z_0 = 0$ without loss of generality. Otherwise, we can translate the function to $h(z - z_0)$. Below, we introduce the convex zero-chain from Nesterov (2018), Section 2.1.2.

Since the subgradients may contain more than one element, we also say $h(z)$is zero-chain whenever there exists some adversarial subgradient oracle. This would also provide a valid lower bound (Nesterov, 2018).

Definition D.2 (Gradient Lipschitz Worse-Case Zero-Chain)**.** *Consider the family of functions:*

$$h_q(z) = \frac{1}{8}(z_{[1]} - 1)^2 + \frac{1}{8}\sum_{j=1}^{q-1}\left(z_{[j+1]} - z_{[j]}\right)^2.$$

The following properties hold for any $h_q(z)$ with $q \in \mathbb{N}^+$:

a. It is a first-order zero-chain.

b. It has a unique minimizer $z^* = \mathbf{1}$.

c. It is 1-gradient Lipschitz.

In bilevel problems, it is crucial to find a point y that is close to $Y^*(x)$, instead of just achieving a small optimality gap $g(x, y) - g^*(x)$. However, it is difficult for any first-order algorithms to "locate" the minimizers of the function class in Definition D.2.

Proposition 4.2. *Fix an x. For any $K \in \mathbb{N}^+$, there exists $d_y \in \mathbb{N}^+$, such that for any $y_0 \in \mathbb{R}^{d_y}$, there exists a pair of functions $f(x, \cdot), g(x, \cdot)$ that are both convex and 1-gradient Lipschitz , for any first-order algorithm $\mathcal{A}$ which initializes from $y_0 \in \mathcal{Y}$ with $\mathrm{dist}\,(y_0, y^*(x)) \leq 1$ and generates a sequence of test points $\{y_k\}_{k=0}^K$ with*

$$y_k \in y_0 + \mathrm{Span}\{\nabla_y f(x, y_0), \nabla_y g(x, y_0), \cdots, \nabla_y f(x, y_{k-1}), \nabla_y g(x, y_{k-1})\}, \quad k \geq 1,$$

it holds that $|f(x, y_k) - \varphi(x)| \geq 1/4$, where $y^*(x)$ is the unique solution to $\min_{y\in Y^*(x)} f(x, y)$.

Proof. Without loss of generality, we assume $y_0 = 0$. Let $d_y = q = 2K$, $\sigma = 1/\sqrt{q}$ and

$$f(x, y) = \frac{1}{2}\sum_{j=K+1}^{q} y_{[j]}^2, \quad g(x, y) = \sigma^2 h_q(y/\sigma),$$

where $h_q(y)$ follows Definition D.2. It is clear from the construction that both $f(x, \cdot), g(x, \cdot)$ are convex and 1-gradient Lipschitz. Moreover, both of them are zero-chains. Then the property of zero-chain leads to

$$y_{k,[j]} = 0, \quad \forall k + 1 \leq j \leq q, \quad 0 \leq k \leq K.$$

Therefore $f(x, y_k)$ remains zero for all $0 \leq k \leq K$.

However, we know that $Y^*(x) = \{\sigma \mathbf{1}\}$. Therefore,

$$\varphi(x) = \frac{1}{2} \sum_{j=K+1}^{q} \sigma^2 = \frac{K\sigma^2}{2} = \frac{1}{4},$$

which indicates that any first-order algorithm $\mathcal{A}$ has a constant sub-optimality gap. □

Next, we prove similar a lower bound also holds for Lipschitz nonsmooth convex lower-level functions, using the following function class, which appears in Nesterov (2018), Section 3.2.1.

Definition D.3 (Lipschitz Zero-Chain)**.** *Consider the family of functions:*

$$h_q(z) = \frac{\sqrt{q}}{2+\sqrt{q}} \max_{1 \le j \le q} z_{[j]} + \frac{1}{2\left(2+\sqrt{q}\right)} \|z\|^2,$$

The following properties hold for any $h_q(z)$ with $q \in \mathbb{N}^+$:

a. It is a first-order zero-chain.

b. It has a unique minimizer $z^* = -\mathbf{1}/\sqrt{q}$.

c. It is 1-Lipschitz in the unit Euclidean ball $\mathbb{B}(z^*) \triangleq \{z : \|z - z^*\| \le 1\}$.

Analogous to Proposition 4.2, we can show the following result.

Proposition D.1. *Fix an x. For any $K \in \mathbb{N}^+$, there exists $d_y \in \mathbb{N}^+$, such that for any $y_0 \in \mathbb{R}^{d_y}$, there exist there exists a pair of functions $f(x,\cdot), g(x,\cdot)$ that are both convex and 1-Lipschitz on $\mathbb{B}(y^*(x))$, such that for any first-order algorithm $\mathcal{A}$ which initializes from $y_0 \in \mathbb{B}(y^*(x))$, and generates a sequence of test points $\{y_k\}_{k=0}^K$ with*

$$y_k \in y_0 + \mathrm{Span}\{\partial_y f(x,y_0), \partial_y g(x,y_0), \cdots, \partial_y f(x,y_{k-1}), \partial_y g(x,y_{k-1})\}, \quad k \ge 1,$$

there exists some subgradients sequence $\{\partial_y f(x,y_0), \partial_y g(x,y_0), \cdots, \partial_y g(x,y_{k-1})\}$ to make $|f(x,y_k) - \varphi(x)| \ge 1/4$ for all k, where $y^*(x)$ is the unique solution to $\min_{y \in Y^*(x)} f(x,y)$.

Proof. Without loss of generality, we assume $y_0 = 0$. Let $d_y = q = 2K$ and

$$f(x,y) = \sum_{j=K+1}^{q} \psi(y_{[j]}), \quad g(x,y) = h_q(y),$$

where $h_q(y)$ follows Definition D.3 and $\psi(y)$ the Huber function defined by

$$\psi(y) = \begin{cases} \beta y - \frac{1}{2}y^2, & y \ge \beta; \\ \frac{1}{2}y^2, & -\beta < y < \beta; \\ -\beta y + \frac{1}{2}y^2, & y \le -\beta. \end{cases}$$

Since $|\psi'(y)| \le \beta$, we know $f(x,\cdot)$ is $(\sqrt{q}\beta)$-Lipschitz since

$$\begin{aligned} &\left| \sum_{j=K+1}^{q} \psi(y_{[j]}) - \sum_{j=K+1}^{q} \psi(y'_{[j]}) \right| \\ &\le \sum_{j=K+1}^{q} \left| \psi(y_{[j]}) - \psi(y'_{[j]}) \right| \\ &\le \beta \sum_{j=K+1}^{q} \left| y_{[j]} - y'_{[j]} \right| \\ &\le \beta\sqrt{q}\|y - y'\| \end{aligned}$$

Let $\beta = 1/\sqrt{q}$ then $f(x,\cdot)$ is 1-Lipschitz. And $g(x,\cdot)$ is 1-Lipschitz on $\mathbb{B}(y^*(x))$.

Note that f always returns a zero subgradient at the origin, while g is a zero-chain. We have

$$y_{k,[j]} = 0, \quad \forall k+1 \leq j \leq q, \quad 0 \leq k \leq K.$$

Therefore $f(x, y_k)$ remains zero for all $0 \leq k \leq K$.

However, we know that $Y^*(x) = \{-\mathbf{1}/\sqrt{q}\}$. So it can be calculated that

$$\varphi(x) = \sum_{j=K+1}^{q} \psi(-1/\sqrt{q}) = -\frac{K}{2q} = -\frac{1}{4},$$

which indicates that any first-order algorithm $\mathcal{A}$ has a constant sub-optimality gap. □

We remark that projection onto the ball centered at the origin $\mathbb{B}(0)$ will not produce additional nonzero entries. Therefore, the possible projection operation in the algorithm will not distort the zero-chain structure.

E Proofs in Section 4.3

Proposition 4.3. *Suppose the solution mapping $Y^*(x)$ is non-empty and compact for any $x \in \mathbb{R}^{d_x}$.*

a. If $f(x, y)$ and $Y^*(x)$ are locally Lipschitz , then $\varphi(x)$ is locally Lipschitz.

b. Conversely, if $\varphi(x)$ is locally Lipschitz for any locally Lipschitz function $f(x, y)$, then $Y^*(x)$ is locally Lipschitz.

c. If $f(x, y)$ is C_f-Lipschitz and $Y^*(x)$ is κ-Lipschitz, then $\varphi(x)$ is C_φ-Lipschitz with coefficient $C_\varphi = (\kappa + 1)C_f$.

d. Conversely, if $\varphi(x)$ is C_φ-Lipschitz for any C_f-Lipschitz function $f(x, y)$, then $Y^*(x)$ is κ-Lipschitz with coefficient $\kappa = C_\varphi/C_f$.

Proof. Note that we can replace sup and inf with max and min in Definition 3.3 due to the compactness of $Y^*(x)$. Below we prove each part of the proposition, separately.

a. Since $Y^*(x_1), Y^*(x_2)$ are nonempty compact sets, we can pick

$$y_1 \in \arg\min_{y \in Y^*(x_1)} f(x_1, y), \quad y_2 \in \arg\min_{y \in Y^*(x_2)} f(x_2, y).$$

Then the Lipschitz continuity of $Y^*(x)$ implies there exist $y_1' \in Y^*(x_1)$ and $y_2' \in Y^*(x_2)$ such that

$$\varphi(x_1) - \varphi(x_2) \leq f(x_1, y_1') - f(x_2, y_2) \leq C_f \left(\|x_1 - x_2\| + \|y_2 - y_1'\|\right) \leq (\kappa + 1)C_f\|x_1 - x_2\|,$$
$$\varphi(x_2) - \varphi(x_1) \leq f(x_2, y_2') - f(x_1, y_1) \leq C_f \left(\|x_1 - x_2\| + \|y_1 - y_2'\|\right) \leq (\kappa + 1)C_f\|x_1 - x_2\|,$$

This establishes the Lipschitz continuity of φ.

b. It suffices to bound the following term for any x_1, x_2:

$$\max\left\{\underbrace{\max_{y_2 \in Y^*(x_2)} \min_{y_1 \in Y^*(x_1)} \|y_1 - y_2\|}_{\text{(I)}}, \quad \underbrace{\max_{y_1 \in Y^*(x_1)} \min_{y_2 \in Y^*(x_2)} \|y_1 - y_2\|}_{\text{(II)}}\right\}. \tag{8}$$

Without loss of generality, we assume $C_f = 1$, otherwise we can scale $f(x, y)$ by C_f to prove the result. We let $f(x, y) = -\min_{y_1 \in Y^*(x_1)} \|y - y_1\|$, then

$$\text{(I)} = \varphi(x_1) - \varphi(x_2) \leq C_\varphi\|x_1 - x_2\|.$$

Next, we let $f(x, y) = \max_{y_1 \in Y^*(x_1)} \|y - y_1\|$, then

$$\text{(II)} \leq \varphi(x_2) - \varphi(x_1) \leq C_\varphi\|x_1 - x_2\|.$$

Together, recalling the definition of (I) and (II) in Equation 8, we know that

$$\operatorname{dist}(Y^*(x_1), Y^*(x_2)) \le C_\varphi \|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathbb{R}^d.$$

Proposition 4.3c and Proposition 4.3d replace the global Lipschitz continuity with local Lipschitz continuity. The proofs are similar, with additional care for the local argument.

c. We use $\mathcal{N}_\delta(\,\cdot\,)$ to denote the open neighbourhood ball with radius δ. For a vector z, we define $\mathcal{N}_\delta(z) \triangleq \{z' : \|z' - z\| < \delta\}$. For a set S, we define $\mathcal{N}_\delta(S) \triangleq \{z' : \operatorname{dist}(z', S) < \delta\}$. For a given $x_1 \in \mathbb{R}^d$ and any $y \in Y^*(x_1)$, the local Lipschitz continuity of $f(\,\cdot\,,\,\cdot\,)$ implies that there exists $\delta_y > 0$ and $L_y > 0$ such that $f(\,\cdot\,,\,\cdot\,)$ is L_y-Lipschitz in $\mathcal{N}_{\delta_y}(x_1) \times \mathcal{N}_{\delta_y}(y)$. Note that the set $S \triangleq \cup_y \left\{\mathcal{N}_{\delta_y}(x_1) \times \mathcal{N}_{\delta_y}(y)\right\}$ forms an open cover of the set $x_1 \times Y^*(x_1)$. The compactness of set $Y^*(x_1)$ guarantees the existence of a finite subcover $\cup_{k=0}^{n} \left\{\mathcal{N}_{\delta_{y_k}}(x_1) \times \mathcal{N}_{\delta_{y_k}}(y_k)\right\}$. Therefore, we can conclude that there exists $\delta_1 > 0$, such that $f(\,\cdot\,,\,\cdot\,)$ is L_1-Lipschitz in the neighborhood $\mathcal{N}_{\delta_1}(x_1) \times \mathcal{N}_{\delta_1}(Y^*(x_1))$, where $L_1 = \max_k L_{y_k}$.

Next, the local Lipschitz continuity of $Y^*(\,\cdot\,)$ implies the existence of $\delta_2 > 0$ and $L_2 > 0$ such that $Y^*(\,\cdot\,)$ is L_2-Lipschitz in $\mathcal{N}_{\delta_2}(x_1)$. Take $\delta = \min\{\delta_1, \delta_2, \delta_1/L_2\}$. The choice of δ ensures $(x_2, y_2) \in \mathcal{N}_{\delta_1}(x_1) \times \mathcal{N}_{\delta_1}(Y^*(x_1))$ for any $x_2 \in \mathcal{N}_\delta(x_1)$ and $y_2 \in Y^*(x_2)$. For any $x_2 \in \mathcal{N}_\delta(x_1)$, we pick

$$y_1 \in \operatorname*{arg\,min}_{y \in Y^*(x_1)} f(x_1, y), \quad y_2 \in \operatorname*{arg\,min}_{y \in Y^*(x_2)} f(x_2, y).$$

The Lipschitz continuity of $f(\,\cdot\,,\,\cdot\,)$ in $\mathcal{N}_{\delta_1}(x_1) \times \mathcal{N}_{\delta_1}(Y^*(x_1))$ and the Lipschitz continuity of $Y^*(\,\cdot\,)$ in $\mathcal{N}_{\delta_2}(x_1)$ implies there exist $y_1' \in Y^*(x_1)$ and $y_2' \in Y^*(x_2)$ such that

$$\varphi(x_1) - \varphi(x_2) \le f(x_1, y_1') - f(x_2, y_2) \le L_1 (\|x_1 - x_2\| + \|y_2 - y_1'\|) \le (L_2 + 1) L_1 \|x_1 - x_2\|.$$
$$\varphi(x_2) - \varphi(x_1) \le f(x_2, y_2') - f(x_1, y_1) \le L_1 (\|x_1 - x_2\| + \|y_1 - y_2'\|) \le (L_2 + 1) L_1 \|x_1 - x_2\|.$$

hold for any $x_2 \in \mathcal{N}_\delta(x_1)$, implying the locally Lipschitz property of $\varphi(\,\cdot\,)$.

d. We again use the function $f(x, y)$ in the proof of **b.** to bound (I) and (II) defined in Equation 8. Let $f(x, y) = -\min_{y_1 \in Y^*(x_1)} \|y - y_1\|$, then there exist $\delta_1 > 0$ and $L_1 > 0$ such that

$$\text{(I)} = \varphi(x_1) - \varphi(x_2) \le L_1 \|x_1 - x_2\|, \quad \forall \|x_1 - x_2\| \le \delta_1.$$

Let $f(x, y) = \max_{y_1 \in Y^*(x_1)} \|y - y_1\|$, then there exist $\delta_2 > 0$ and $L_2 > 0$ such that

$$\text{(II)} \le \varphi(x_2) - \varphi(x_1) \le L_2 \|x_1 - x_2\|, \quad \forall \|x_1 - x_2\| \le \delta_2.$$

Together, taking $\delta = \min\{\delta_1, \delta_2\}$ and $L = \max\{L_1, L_2\}$ and recalling the definition of (I) and (II) in Equation 8, we can show that there exists some $\delta > 0$ such that it holds

$$\operatorname{dist}(Y^*(x_1), Y^*(x_2)) \le L \|x_1 - x_2\|, , \quad \forall \|x_1 - x_2\| \le \delta,$$

which implies the local Lipschitz property of $Y^*(\,\cdot\,)$.

□

F Proofs in Section 5

Proposition 5.1. *Under Assumption 5.1 or 5.2, $Y^*(x)$ is (L/α)-Lipschitz.*
Furthermore, if $f(x, y)$ is C_f-Lipschitz, then $\varphi(x)$ is $(L/\alpha + 1) C_f$-Lipschtz.

Proof. We show that $Y^*(x)$ is Lipschitz, and then $\varphi(x)$ is also Lipschitz by Proposition 4.3.

Under Assumption 5.1, for any $y_1 \in Y^*(x_1)$, there exists $y_2 \in Y^*(x_2)$ such that

$$\begin{aligned}
&\alpha \|y_1 - y_2\| \\
&\le \|\mathcal{G}_{1/L}(y_1; x_2) - \mathcal{G}_{1/L}(y_1; x_1)\| \\
&= L \left\| \mathcal{P}_{\mathcal{Y}} \left(y_1 - \frac{1}{L} \nabla_y g(x_2, y_1) \right) - \mathcal{P}_{\mathcal{Y}} \left(y_1 - \frac{1}{L} \nabla_y g(x_1, y_1) \right) \right\|
\end{aligned}$$

$$\begin{aligned}&\leq \|\nabla_y g(x_2, y_1) - \nabla_y g(x_1, y_1)\| \\ &\leq L\|x_1 - x_2\|,\end{aligned}$$

where we use $\mathcal{G}_{1/L}(y_1; x_1) = 0$ (Drusvyatskiy & Lewis, 2018) and Assumption 5.1 in the second line; the third line follows from the definition of the generalized gradient; the fourth line uses the non-expansiveness of projection operator by Corollary 2.2.3 in Nesterov (2018); and the last line uses the smoothness property of the lower-level function.

Under Assumption 5.2, for any $y_1 \in Y^*(x_1)$, there exists $y_2 \in Y^*(x_2)$ such that

$$\begin{aligned}&2\alpha\|y_1 - y_2\| \\ \leq\ & g(x_2, y_1) - g(x_2, y_2) \\ \leq\ & g(x_1, y_1) - g(x_1, y_2) + 2L\|x_1 - x_2\| \\ \leq\ & 2L\|x_1 - x_2\|,\end{aligned}$$

where the last line uses $g(x_1, y_2) \leq g(x_1, y_2)$. □

G Proofs in Section 6.1

First of all, we prove that the proposed Switching (sub)Gradient Method (SGM) in Algorithm 1 can find an LL optimal solution under the following Hölderian error bound condition.

Assumption G.1. *We suppose the lower-level function $g(x, \cdot)$ satisfies the r-th order Hölderian error bound condition on set $\mathcal{Y}$ with some coefficient $\nu > 0$, that is,*

$$\frac{\nu}{r}\operatorname{dist}(y, Y^*(x))^r \leq g(x, y) - g^*(x), \quad \forall y \in \mathcal{Y}.$$

Note that this condition is also used by Jiang et al. (2023) and they show the following result.

Lemma G.1 (Proposition 1 (i) in Jiang et al. (2023)). *Suppose that Assumption G.1 holds, $f(x, \cdot)$ is convex and $f(x, y)$ is C_f-Lipschitz. If a point y satisfies*

$$f(x, y) - \varphi(x) \leq \zeta, \quad g(x, y) - g^*(x) \leq \frac{\nu}{r}\left(\frac{\zeta}{C_f}\right)^r, \tag{9}$$

then we have $|f(x, y) - \varphi(x)| \leq \zeta$.

Equation 9 can be achieved by the SGM, then we can show the following result for finding an LL optimal solution the Hölderian error bound condition.

Theorem G.1. *Under Assumption 6.1 and G.1, if we let*

$$\theta = \min\left\{\zeta, \frac{\nu}{4r}\left(\frac{\zeta}{C_f}\right)^r\right\}, \quad K_0 = K = \left\lceil \frac{4R^2 \max\{C_f^2, C_g^2\}}{\theta^2} \right\rceil, \quad \tau = \frac{R}{\max\{C_f, C_g\}\sqrt{K}}. \tag{10}$$

then Algorithm 1 can output a point y_{out} satisfying $|f(x, y_{\text{out}}) - \varphi(x)| \leq \zeta$ within $\mathcal{O}\left(\frac{r^2 \max\{C_f^2, C_g^2\} C_f^{2r} R^2}{\nu^2 \zeta^{2r}}\right)$ first-order oracle complexity.

Proof. Combine Lemma G.1 and Lemma G.3. □

Next, we prove Lemma G.3, which relies on the following standard lemma for subgradient descent.

Lemma G.2 (Subgradient Descent). *Suppose h is a L-Lipschitz convex function. For any $y, z \in \mathcal{Y}$, if we let $y^+ = \mathcal{P}(y - \tau\partial h(y))$, then it holds that*

$$h(y) - h(z) \leq \frac{1}{2\tau}(\|y - z\|^2 - \|y^+ - z\|^2) + \frac{\tau L^2}{2}.$$

Proof. See Theorem 3.2 in Bubeck et al. (2015). □

Using this lemma, we then show the following result.

Lemma G.3. *Under the setting of Theorem G.1, the output of Algorithm 1 satisfies:*

$$f(x, y_{\rm out}) - \varphi(x) \le \theta, \quad g(x, y_{\rm out}) - g^*(x) \le 4\theta.$$

Proof. By Theorem 3.2 in Bubeck et al. (2015), the initialization step ensures $\hat{g}^*(x) - g^*(x) \le 2\theta$. Pick any $y^*(x) \in \arg\min_{y \in Y^*(x)} f(x, y)$ and denote $C = \max\{C_f, C_g\}$.

According to Lemma G.2 we obtain

$$f(x, y_k) - \varphi(x) \le \frac{1}{2\tau}(\|y_k - y^*(x)\|^2 - \|y_{k+1} - y^*(x)\|^2) + \frac{\tau C^2}{2}, \quad k \in \mathcal{I};$$
$$g(x, y_k) - g^*(x) \le \frac{1}{2\tau}(\|y_k - y^*(x)\|^2 - \|y_{k+1} - y^*(x)\|^2) + \frac{\tau C^2}{2}, \quad k \notin \mathcal{I};$$

Combing them together yields

$$\frac{1}{K}\sum_{k \in \mathcal{I}} f(x, y_k) - \varphi(x) + \frac{1}{K}\sum_{k \notin \mathcal{I}} g(x, y_k) - g^*(x) \le \frac{R^2}{2\tau K} + \frac{\tau C^2}{2} = \frac{RC}{\sqrt{K}}. \tag{11}$$

With Equation 11 in hand, it suffices to show the result.

Firstly, we show that $\mathcal{I} \neq \emptyset$, and thus $y_{\rm out}$ is well-defined. Otherwise, we would have the following contradiction:

$$2\theta \le \frac{1}{K}\sum_{k=0}^{K-1} g(x, y_k) - \hat{g}^*(x) \le \frac{1}{K}\sum_{k=0}^{K-1} g(x, y_k) - g^*(x) \le \frac{RC}{\sqrt{K}} \le \frac{\theta}{2}.$$

Secondly, we show that the output will not violate the constraint too much by:

$$g(x, y_{\rm out}) - g^*(x) \le \frac{1}{|\mathcal{I}|}\sum_{k \in \mathcal{I}}(g(x, y_k) - \hat{g}^*(x)) + (\hat{g}^*(x) - g^*(x)) \le 4\theta.$$

Thirdly, we show that $f(x, y_{\rm out}) - \varphi(x) \le \theta$. It is trivial when $\sum_{k \in \mathcal{I}} f(x, y_k) - \varphi(x) \le 0$ since it is an immediate result of Jensen's inequality. Therefore we can only focus on the case when $\sum_{k \in \mathcal{I}} f(x, y_k) - \varphi(x) > 0$. In this case, we can show that $|\mathcal{I}| \ge K/2$, otherwise we would have

$$\theta < \frac{1}{K}\sum_{k \notin \mathcal{I}} g(x, y_k) - \hat{g}^*(x) \le \frac{1}{K}\sum_{k \notin \mathcal{I}} g(x, y_k) - g^*(x) \le \frac{RC}{\sqrt{K}} \le \frac{\theta}{2},$$

which also leads to a contradiction. Hence we must have $|\mathcal{I}| \ge K/2$, therefore, we obtain

$$f(x, y_{\rm out}) - \varphi(x) \le \frac{1}{|\mathcal{I}|}\sum_{k \in \mathcal{I}} f(x, y_k) - \varphi(x) \le \frac{2}{K}\sum_{k \in \mathcal{I}} f(x, y_k) - \varphi(x) \le \frac{2RC}{\sqrt{K}} \le \theta.$$

This completes our proof. □

We want to use Theorem G.1 to prove Theorem 6.1. Their only difference between them is the assumption. The following proposition shows that both Assumption 5.1 and 5.2 imply Assumption G.1 when $g(x, y)$ is convex in y. Therefore, the function class studied in Theorem 6.1 is contained in the function class studied in Theorem G.1.

Proposition G.1. *If $g(x, \cdot)$ is convex, then either Assumption 5.1 or 5.2 implies Assumption G.1.*

Proof. According to Corollary 3.6 in Drusvyatskiy & Lewis (2018), Assumption 5.1 implies Assumption G.1 with any $\nu < \alpha$ under the convexity of $g(x, \cdot)$ For Assumption 5.2, it is clear that it is equivalent to Assumption G.1 with $r = 1$. □

Now we can easily prove Theorem 6.1.

Theorem 6.1. *Fix an x. Under Assumption 6.1, Algorithm 1 with appropriate parameters can ouput a point $y_{\rm out}$ satisfying $|f(x, y_{\rm out}) - \varphi(x)| \le \zeta$ and $g(x, y_{\rm out}) - g^*(x) \le \zeta$ with $\mathcal{O}(\mathrm{poly}(1/\zeta))$ first-order oracle calls from g.*

Proof. Combine Theorem G.1 and Proposition G.1.

□

H Proofs in Section 6.2

Theorem 6.2. *Given any $\varepsilon \lesssim C_f$. Let $\Delta = \varphi(x_0) - \varphi^*$. Under Assumption 6.2, set*

$$T = \mathcal{O}\left(d_x^{3/2}\left(\frac{C_\varphi^4}{\varepsilon^4} + \frac{\Delta C_\varphi^3}{\delta\varepsilon^4}\right)\right), \ \eta = \Theta\left(\sqrt{\frac{\delta(\Delta + \delta C_\varphi)}{d_x^{3/2} C_\varphi^3 T}}\right). \tag{5}$$

Then Algorithm 2 can output a point x_{out} that satisfies $\mathbb{E}\min\{\|s\| : s \in \partial_\delta \varphi(x_{\mathrm{out}})\} \leq \varepsilon$.

Proof. First of all, we let

$$\varphi_\delta = \mathbb{E}_{v\sim\mathbb{P}_v}[\varphi(x + \delta v)],$$

where $\mathbb{P}_v$ is a uniform distribution on a unit ball in ℓ_2-norm. Then, we define

$$\nabla_t = \frac{d_x}{2\delta}(\varphi(x_t + \delta u_t) - \varphi(x_t - \delta u_t))u_t. \tag{12}$$

According to Lemma D.1 in Lin et al. (2022), ∇_t satisfies the following properties:

$$\mathbb{E}_{u_t}[\nabla_t \mid x_t] = \nabla\varphi_\delta(x_t), \quad \mathbb{E}_{u_t}\left[\|\nabla_t\|^2 \mid x_t\right] \leq 16\sqrt{2\pi} d_x C_\varphi^2,$$

Then we know that

$$\mathbb{E}_{u_t}\left[\|\nabla_t - \hat{\nabla}_t\| \mid x_t\right] \leq \frac{d_x\zeta}{\delta}\mathbb{E}_{u_t}\|u_t\| = \frac{d_x\zeta}{\delta} \leq \frac{c_4\varepsilon^2}{C_\varphi} \tag{13}$$

and

$$\begin{aligned}
&\mathbb{E}_{u_t}\left[\|\hat{\nabla}_t\|^2 \mid x_t\right] \\
&\leq 2\mathbb{E}_{u_t}\left[\|\nabla_t\|^2 \mid x_t\right] + 2\mathbb{E}_{u_t}\left[\|\nabla_t - \hat{\nabla}_t\|^2 \mid x_t\right] \\
&\leq 2\mathbb{E}_{u_t}\left[\|\nabla_t\|^2 \mid x_t\right] + \frac{2d_x^2\zeta^2}{\delta^2}\mathbb{E}_{u_t}\|u_t\|^2 \\
&\leq 32\sqrt{2\pi}d_x C_\varphi^2 + \frac{2d_x^2\zeta^2}{\delta^2} \\
&\leq c_1 d_x C_\varphi^2
\end{aligned} \tag{14}$$

for some positive constant $c_1, c_4 > 0$. Then we use the results of Equation 13, Equation 14 as well as the standard analysis of gradient descent to obtain

$$\begin{aligned}
\mathbb{E}\left[\varphi_\delta(x_{t+1}) \mid x_t\right] &\leq \varphi_\delta(x_t) - \eta\langle\, \nabla\varphi_\delta(x_t),\, \mathbb{E}[\hat{\nabla}_t \mid x_t]\,\rangle + \frac{c_2\eta^2 C_\varphi\sqrt{d_x}}{2\delta}\mathbb{E}\left[\|\hat{\nabla}_t\|^2 \mid x_t\right] \\
&\leq \varphi_\delta(x_t) - \eta\|\nabla\varphi_\delta(x_t)\|^2 + \frac{\eta C_\varphi d_x\zeta}{\delta} + \frac{c_2\eta^2 C_\varphi\sqrt{d_x}}{2\delta}\mathbb{E}\left[\|\hat{\nabla}_t\|^2 \mid x_t\right] \\
&\leq \varphi_\delta(x_t) - \eta\|\nabla\varphi_\delta(x_t)\|^2 + \frac{c_3\eta^2 C_\varphi^3 d_x^{3/2}}{\delta} + \eta c_4\varepsilon^2,
\end{aligned}$$

where we use Proposition 2.3 in Lin et al. (2022) that φ_δ is differentiable and C_φ-Lipschitz with the $(c_2 C_\varphi\sqrt{d_x}/\delta)$-Lipschitz gradient where $c_2 > 0$ is a positive constant and we define $c_3 = 2c_1c_2$.

Telescoping for $t = 0, 1, \cdots, T$, we obtain

$$\mathbb{E}\|\nabla\varphi_\delta(x_{\mathrm{out}})\|^2 \leq \frac{\Delta + \delta C_\varphi}{\eta T} + \frac{c_3\eta C_\varphi^3 d^{3/2}}{\delta} + c_4\varepsilon^2,$$

where we use $|\varphi_\delta(x) - \varphi(x)| \leq \delta C_\varphi$ for any $x \in \mathbb{R}^{d_x}$ by Proposition 2.3 in Lin et al. (2022).

Lastly, plugging the value of η, T with a sufficiently small constant c_4 and noting that $\nabla\varphi(x_{\mathrm{out}}) \in \partial_\delta\varphi(x_{\mathrm{out}})$ by Theorem 3.1 in Lin et al. (2022), we arrive at the conclusion. □