

Improved Analysis of Penalty-Based Methods for Bilevel Optimization with Coupled Constraints

Liuyuan Jiang^{*‡}, Quan Xiao^{†‡}, Tianyi Chen^{†‡}

[‡]Dept. of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute

^{*}Dept. of Electrical and Computer Engineering, University of Rochester

[†]Dept. of Electrical and Computer Engineering, Cornell Tech, Cornell University

Abstract—Bi-objective optimization arises in various applications, often leading to bilevel optimization (BLO) formulations with coupled constraints. To solve BLO via gradient-based approaches, implicit gradient methods resort to the Hessian inverse to estimate the descent direction for the upper-level variable, which is computationally costly. Penalty-based approaches offer an attractive alternative by reformulating the problem as a single-level problem, allowing the use of only first-order information. However, existing penalty-based methods suffer from the challenging optimization landscape (large smoothness constant), which limits the convergence rate to $\mathcal{O}(\epsilon^{-1.5})$. This work revisited the penalty-based formulation that ensures an $\mathcal{O}(1)$ -smooth objective. We achieve this by analyzing the 2nd-order directional derivative under both non-coupled and coupled constraints. Consequently, our approach improves the iteration complexity of the recent Penalty-Based Gradient Descent (PBGD) method [20] from $\mathcal{O}(\epsilon^{-1.5})$ to $\mathcal{O}(\epsilon^{-1})$, matching the rate of gradient descent applied on smooth objectives. Our results apply to bilevel optimization with general nonlinear coupled constraints, enhancing the efficiency of penalty-based methods in BLO. The Appendix of this work, which includes the theoretical details and experimental results, is available at this GitHub.

Index Terms—bilevel optimization, penalty, first order, Hessian

I. INTRODUCTION

Bi-objective optimization, which seeks to optimize two potentially conflicting objectives simultaneously, is a fundamental problem in decision-making across various domains, including representation learning [1], reinforcement learning [21], financial pricing [22], and transportation network [19].

Many bi-objective problems exhibit a hierarchical structure [1], [21], where one objective seeks to optimize $f(x, y)$, while the other aims at choosing y as $y_g^*(x) = \arg \min_y g(x, y)$. Additionally, many problems [19], [22] impose feasibility constraints, e.g. $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $c(x, y) \leq 0$. This naturally lead to a BiLevel Optimization (BLO) formulation:

$$\min_{x \in \mathcal{X}} \phi(x) := f(x, y_g^*(x)) \quad \text{s.t.} \quad y_g^*(x) := \arg \min_{y \in \mathcal{Y}(x)} g(x, y) \\ \text{where} \quad \mathcal{Y}(x) := \{y \in \mathcal{Y} : c(x, y) \leq 0\}. \quad (1)$$

Here, we call $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}$ and $g : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}$ respectively upper-level (UL) and lower-level (LL) objectives; $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ is the UL domain constraint; $\mathcal{Y}(x) \subseteq \mathbb{R}^{d_y}$ is the LL

constraint including domain constraint $\mathcal{Y}$ independent from x and coupled inequality constraints $c : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}^{d_c}$.

In this way, the BLO problem (1) solves the bi-objective problem by finding optimal x^* over $\phi(x)$ and its associated $y_g^*(x^*)$ minimizing $g(x, \cdot)$ under constraints. Using gradient-based methods, the key challenge lies in determining a proper descent direction for x . To address this, Implicit Gradient Descent (IGD) methods (e.g., [5], [7]–[9], [13]) approximate $\frac{\partial}{\partial x} y_g^*(x)$ via the inversion of hessian $\nabla_{yy} g(x, y_g^*(x))$, which is computationally costly and is limited to tackling only $\mathcal{Y} = \mathbb{R}^{d_y}$, e.g. in [23], [24]. Penalty-based methods, e.g. [11], [14], [15], [20], [27], offer an alternative by penalizing the LL objective optimality gap into the UL via a large penalty constant γ :

$$H_\gamma(x, y) := f(x, y) + \gamma(g(x, y) - \min_{y_g \in \mathcal{Y}(x)} g(x, y_g)). \quad (2)$$

Under mild conditions, it was established that the local solutions to (2) are within $\mathcal{O}(\epsilon)$ -squared-distance of those to (1) when choosing $\gamma = \Omega(\epsilon^{-0.5})$ [20]. Moreover, the value function $v(x) = \min_{y_g \in \mathcal{Y}(x)} g(x, y_g)$ is $l_{v,1}$ -smooth, implying $H_\gamma(x, y)$ is $l_{H,1} = \mathcal{O}(\gamma)$ -smooth. This enables solving (1) via implementing Projected Gradient Descent (PGD) on (2). However, the choice of $\gamma = \Omega(\epsilon^{-0.5})$ requires the step size $\eta = \mathcal{O}(\epsilon^{0.5})$ to satisfy condition $\eta \leq l_{H,1}^{-1}$ in the PGD algorithm. This dampens the algorithm complexity to $\mathcal{O}(\eta^{-1} \epsilon^{-1}) = \mathcal{O}(\epsilon^{-1.5})$. This prompts the question:

(Q): Can we solve the penalty problem (2) with the same iteration complexity of gradient descent by showing a formulation with smoothness constant independent of γ ?

We answer this affirmatively via decoupling x from y :

$$F_\gamma(x) := \min_{y \in \mathcal{Y}(x)} H_\gamma(x, y) \\ = \gamma \underbrace{\min_{y_g \in \mathcal{Y}(x)} \left(\frac{1}{\gamma} f(x, y_g) + g(x, y_g) \right)}_{=: v_\gamma(x)} - \gamma \underbrace{\min_{y_g \in \mathcal{Y}(x)} g(x, y_g)}_{=: v(x)}. \quad (3)$$

To analyze the smoothness constant $l_{F,1}$ of $F_\gamma(x)$, we examine the second order directional derivative $D_{dd}^2(F(x))$, since $l_{F,1}$ serves as an upper bounds for $\|D_{dd}^2(F(x))\|$. This follows from the analysis of the value functions $v_\gamma(x)$ and $v(x)$. When the LL constraint is absent, i.e. $\mathcal{Y}(x) = \mathbb{R}^{d_y}$ and $c(x, y) = 0$, a closed-form Hessian expression of $F_\gamma(x)$ was concluded [3] and $F_\gamma(x)$ was estimated to be $\mathcal{O}(1)$ -smooth [4] based on LL

The work was supported by NSF project 2401297, NSF CAREER project 2532349, Cisco Research Award, and the IBM-Rensselaer Future of Computing Research Collaboration.

TABLE I: Comparison of Methods*

Method	LL Constraint	$l_{F,1}$ (or $l_{H,1}$)	Complexity
JNT-PBGD	$y \in \mathcal{Y} \ \& \ c(y) \leq 0$	$\mathcal{O}(\gamma)$	$\tilde{\mathcal{O}}(\epsilon^{-1.5})$
Prox-F ² SA	$c(y) \leq 0$	$\mathcal{O}(\gamma)$	$\tilde{\mathcal{O}}(\epsilon^{-1.5})$
BLOCC	$y \in \mathcal{Y} \ \& \ c(x, y) \leq 0$ $A(x)y + B(x) \leq 0$	$\mathcal{O}(\gamma)$	$\tilde{\mathcal{O}}(\epsilon^{-1.5})$ $\tilde{\mathcal{O}}(\epsilon^{-2.5})$
F ² SA	unconstrained	$\mathcal{O}(1)$	$\tilde{\mathcal{O}}(\epsilon^{-1})$
Ours	$\mathcal{Y} \ \& \ c(y) \leq 0$	$\mathcal{O}(1)$	$\tilde{\mathcal{O}}(\epsilon^{-1})$
	$\mathcal{Y} \ \& \ c(x, y) \leq 0$		$\tilde{\mathcal{O}}(\epsilon^{-2})$
	$A(x)y + B(x) \leq 0$		$\tilde{\mathcal{O}}(\epsilon^{-1})$

*We compare our results with JNT-PBGD [20], Prox-F²SA [15], BLOCC [11], and improved analysis of F²SA [3], [4]. The convergence metric is the squared (generalized) gradient norm. We use $\tilde{\mathcal{O}}$ in short for $\mathcal{O}(\ln(\epsilon^{-1}))$

stationarity. However, introducing constraints complicates the analysis, as $\nabla_y g(x, y_g^*(x)) = 0$ no longer holds, requiring us to address constraint-induced discontinuities, a challenge not addressed in existing literature.

A. Contributions

Our work is the first to tackle the challenge in **(Q)** considering coupled constraints. We highlight key contributions using **C1)**, **C2)**, etc. In Section III-A, we begin the analysis from the non-coupled-constraint case, i.e. $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(y) \leq 0\}$. **C1)** we observe an alternative to the stationary condition in which the directional derivative of $y_g^*(x)$ is orthogonal to $\nabla_y g(x, y_g^*(x))$. In this way, with the strong convexity and some Lipschitz conditions of $g(x, \cdot)$, **C2)** we bridge the connection between the 2nd-order directional derivative of $v(x)$ and $v_\gamma(x)$ in Lemma 3 and therefore conclude that $F_\gamma(x)$ in (3) is $l_{F,1}$ -Lipschitz-smooth with $l_{F,1} = \mathcal{O}(1)$. In Section III-B, we revisited an alternating version of Penalty-Based Gradient Descent (PBGD) method for $\min_{x \in \mathcal{X}} F_\gamma(x)$, ALT-PBGD in Algorithm 1, which alternates between minimizing $H(x, y)$ over $y \in \mathcal{Y}$ and $F_\gamma(x)$ over $x \in \mathcal{X}$. **C3)** ALT-PBGD achieves $\tilde{\mathcal{O}}(\epsilon^{-1})$ complexity, with its outer loop matching the complexity of gradient descent. It improves the $\tilde{\mathcal{O}}(\epsilon^{-1.5})$ complexity of JNT-PBGD method jointly minimizing $H_\gamma(x, y)$ over $(x, y) \in \mathcal{X} \times \mathcal{Y}$ in existing literature [20]. **C4)** Section IV extends the results to the coupled constrained case $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(x, y) \leq 0\}$. In this way, we establish $\mathcal{O}(1)$ -smoothness for $F_\gamma(x)$ and improve the iteration complexity of BLOCC, a PBGD method for BLO with Coupled Constraints [11], by $\mathcal{O}(\epsilon^{-0.5})$. Numerical experiments are provided in Appendix [10].

B. Prior art

BLO has a rich history, with early work dating back to [2]. Recent advances focus on efficient gradient-based methods with finite-time guarantees. *IGD methods*, introduced by [18], approximate the hypergradient $\frac{\partial}{\partial x} y_g^*(x)$ using the implicit function theorem, primarily under the strongly convex LL assumption [5], [7]–[9], [13]. However, IGD methods are computationally expensive due to the need for second order calculation. Alternatively, *Penalty-Based methods* reformulate BLO as a single-level problem with penalty terms, which avoids Hessian computations and is fully first-order. Dating

back to [26], these methods have regained significant popularity recently [14], [16], [17], [20], [27]. Moreover, motivated by real-world applications, recent research has increasingly focused on BLO problems with LL constraints [11], [12], [15], [20], [23]–[25], and penalty-based methods [11], [24], demonstrate their effectiveness in handling both function constraints $c(x, y) \leq 0$ and domain constraints $y \in \mathcal{Y}$ with low algorithm complexity.

When applying penalty-based methods, the smoothness of the penalty reformulation is crucial, as the step size is bounded by the inverse of the smoothness constant. For non-coupled LL constraints (e.g., $\mathcal{Y}$ or $c(y) \leq 0$), [14], [20] achieve an $\mathcal{O}(\epsilon^{-0.5})$ -smoothness for the penalty reformulation. Similarly, [11] extends this to coupled constraints $c(x, y) \leq 0$ and domain constraints $y \in \mathcal{Y}$, leading to a step size bound of $\mathcal{O}(\epsilon^{0.5})$, which in turn dampens iteration complexity. [15] derives a closed-form expression for $\nabla^2 v(x)$ under $c(y) \leq 0$, but the smoothness constant remains at $\mathcal{O}(\gamma)$. For unconstrained BLO, [4] achieves $\mathcal{O}(1)$ smoothness via decoupled penalty reformulation $F_\gamma(x)$. However, results for constrained LL problems, especially with domain $\mathcal{Y}$ and coupled inequality constraints $c(x, y) \leq 0$, remain limited. Table I compares prior works on penalty methods and smoothness analysis.

II. PRELIMINARY OF THE PENALTY REFORMULATION

This section explores preliminary properties for penalty reformulation $F_\gamma(x)$. Before proceeding, we outline the assumptions, with definitions provided in Appendix [10, Sec. II].

Assumption 1 (Upper level). Assume differentiable $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}$ is (1) $l_{f,0}$ -Lipschitz in $y \in \mathcal{Y}$, (2) $l_{f,1}$ -smooth in $(x, y) \in \mathcal{X} \times \mathcal{Y}$, (3) locally-Lipschitz in $x \in \mathcal{X}$.

Assumption 2 (Lower level). Assume differentiable $g : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}$ is (1) μ_g -strongly-convex in $y \in \mathcal{Y}$, (2) $l_{g,1}$ -smooth in $(x, y) \in \mathcal{X} \times \mathcal{Y}$, (3) locally-Lipschitz in $x \in \mathcal{X}$.

Assumption 3 (Constraints). Assume (1) $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ are closed and convex; (2) differentiable $c : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}^{d_c}$ is convex in $y \in \mathcal{Y}$, $l_{c,1}$ -smooth in $(x, y) \in \mathcal{X} \times \mathcal{Y}$, satisfies the Linear Constraint Qualification (LICQ) condition in $y \in \mathcal{Y}$ at optimal points, (3) and locally-Lipschitz in x .

The differentiability and Lipschitz continuity conditions for f , g , and c in Assumptions 1, 2, and 3 are standard [5], [7]–[9], [11], [14]. The strong convexity of the LL problem is conventional [3], [5], [7], [11] and still presents challenges due to the imposed constraints. Moreover, assuming $c(x, y)$ convex in y is mild and traditional [11], [12], [23], [25]. The convexity and closure of $\mathcal{X}$ and $\mathcal{Y}$ are standard, and the LICQ is a common assumption in constrained BLO [11], [15], [24].

With these conditions, $F_\gamma(x)$ is a good approximation to $\phi(x)$ in (1) with distance controlled by γ^{-1} and solving $F_\gamma(x)$ is equivalent to solving to find ϵ -suboptimal $\phi(x)$.

Lemma 1. Suppose Assumption 1.1-2, 2.1-2, and 3 hold. The ϵ -suboptimal local solutions in distance square metric for ϵ -approximation problem of (1):

$$\min_{x \in \mathcal{X}, y \in \mathcal{Y}(x)} f(x, y) \quad \text{s.t.} \quad \|y - y_g^*(x)\|^2 \leq \epsilon, \quad (4)$$

are ϵ -suboptimal local solutions for $\min_{x \in \mathcal{X}} F_\gamma(x)$ in (3) with $\gamma = \mathcal{O}(\epsilon^{-0.5})$ and $\gamma > \frac{l_{f,1}}{\alpha_g}$. Additionally, there is

$$\|y_g^*(x) - y_\gamma^*(x)\|^2 \leq \mathcal{O}(l_{f,0}\mu_g^{-1}\gamma^{-1}), \quad (5)$$

where $y_g^*(x)$ is in (1), and

$$y_\gamma^*(x) := \arg \min_{y \in \mathcal{Y}(x)} \gamma^{-1}f(x, y) + g(x, y). \quad (6)$$

The proof of Lemma 1 follows from [11, Theorem 1] and [20] directly. Here, g is strongly convex in y and f is smooth, and $\gamma^{-1}f + g$ is strongly convex in y when $\gamma \geq \frac{l_{f,1}}{\mu_g}$ as $l_{f,1}$ -smoothness ensures a lower bound for negative curvature of f . Moreover, $F_\gamma(x) = \gamma(v_\gamma(x) - v(x))$ features favorable properties such as differentiability and smoothness, as do the value functions.

Lemma 2 (Derivative of $v(x)$ [11, Lemma 2]). *Suppose Assumption 1, 2, 3 hold. For $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(x, y) \leq 0\}$, the value function $v(x) = \min_{y \in \mathcal{Y}(x)} g(x, y)$ is differentiable:*

$$\nabla v(x) = \nabla_x g(x, y_g^*(x)) + \langle \lambda_g^*(x), \nabla_x c(x, y_g^*(x)) \rangle, \quad (7)$$

where $\lambda_g^*(x)$ is the unique Lagrangian multiplier.

The lemma 2 is the cornerstone of the implementation of a gradient descent-based algorithm to solve the reformulation $F_\gamma(x)$ or $H_\gamma(x)$, such as in [11], [14], [20].

III. IMPROVED CONVERGENCE RATE UNDER NON-COUPLED CONSTRAINT

In this section, we start by considering the non-coupled constraint $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(y) \leq 0\}$ independent from x . Section III-A provides a dedicated analysis of the smoothness of $F_\gamma(x)$. In Section III-B, we revisited ALT-PBGD and demonstrated that it is an optimal algorithm that matches the convergence complexity of the gradient descent.

A. Tighter smoothness estimate of $F_\gamma(x)$

Existing literature [15], [20] investigates the joint minimization of (x, y) for $H_\gamma(x, y)$ in (2), whose smoothness modulus is of order $\mathcal{O}(\gamma)$ [20]. This leads to a prior estimate of the smoothness modulus for $F_\gamma(x)$ as $l_{F,1} = \mathcal{O}(\gamma)$. However, empirical evidence, e.g. Example 1, shows that although $\nabla_x H(x, y)$ will be scaled up by γ , $\nabla F_\gamma(x)$ remains at a constant value. As in Figure 1, larger γ results in steeper gradients for $\nabla_x H(x, y)$ while it hardly affects $\nabla F_\gamma(x)$.

Example 1. With $\mathcal{X} = \mathcal{Y}(x) = [0, 3]$, consider the BLO problem in (1) with the objectives as follows

$$f(x, y) = \frac{e^{-y+1}}{2 + \cos(4x)} + \frac{1}{2} \ln((4x-2)^2 + 1) + x^2$$

$$g(x, y) = 2(y-x)^2 + \frac{x}{2} \sin^2(x+y).$$

This motivates a re-examination of the smoothness properties of $F_\gamma(x)$. To analyze the smoothness constant $l_{F,1}$, we consider the second-order directional derivative $D_{dd}^2(F(x))$, as $l_{F,1}$ provides an upper bound for $\|D_{dd}^2(F(x))\|$. Specifically, recalling that $F_\gamma(x) = \gamma(v_\gamma(x) - v(x))$, as given in (3), we are led to

analyze the second-order properties of the value functions. In the unconstrained case, when assuming LL strongly-convexity, the lower level stationarity $\nabla_y g(x, y_g^*(x)) = 0$ gives

$$0 = \lim_{r \downarrow 0} \frac{1}{r} (\nabla_y g(x + rd, y_g^*(x + rd)) - \nabla_y g(x, y_g^*(x)))$$

$$= \nabla_{xy} g(x, y_g^*(x))^\top d + \nabla_{yy} g(x, y_g^*(x)) \frac{\partial}{\partial x} y_g^*(x) d \quad (8)$$

by Taylor's expansion. Therefore, prior arts e.g. [5], [7] obtain

$$\frac{\partial}{\partial x} y_g^*(x) = \nabla_{yy} g(x, y_g^*(x))^{-1} \nabla_{xy} g(x, y_g^*(x)). \quad (9)$$

This enables finding $\nabla^2 v(x)$ and its counterpart $\nabla^2 v_\gamma(x)$ such as in [4]. However, when involving the LL constraint, $\nabla_y g(x, y_g^*(x)) = 0$ does not hold in general. We address this by observing an alternative. Under $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(y) \leq 0\}$,

$$\left\langle \nabla_y g(x, y_g^*(x)), \lim_{r \downarrow 0} \frac{y_g^*(x + rd) - y_g^*(x)}{r} \right\rangle = 0 \quad (10)$$

holds for all unit direction $d \in \mathbb{R}^{d_x}$, as summarized in Lemma 6 in Appendix [10, Sec. III-A]. This enables the analysis of the second-order directional derivative of value functions by constructing an alternative to (8).

Lemma 3. Consider $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(y) \leq 0\}$. Suppose Assumption 1.1-2, 2.1-2, 3 hold. Fix any $\delta > 0$. there exists some finite γ^* such that for any x and unit direction $d \in \mathbb{R}^{d_x}$, there exists an index set $\mathcal{I} \subseteq [d_y]$ such that the second-order directional derivatives of $v(x)$ and $v_\gamma(x)$ are

$$D_{dd}^2(v(x)) = d^\top \left(A(x) - B(x) \begin{bmatrix} C(x)_{[\mathcal{I}, \mathcal{I}]}^{-1} B(x)_{[\cdot, \mathcal{I}]}^\top \\ 0 \end{bmatrix} \right) d + \mathcal{O}(\delta),$$

$$D_{dd}^2(v_\gamma(x)) = d^\top \left(A_\gamma(x) - B_\gamma(x) \begin{bmatrix} C_\gamma(x)_{[\mathcal{I}, \mathcal{I}]}^{-1} B_\gamma(x)_{[\cdot, \mathcal{I}]}^\top \\ 0 \end{bmatrix} \right) d + \mathcal{O}(\delta) \quad (11)$$

for all $\gamma > \gamma^*$ for the same $\mathcal{I}$, where $A(x) = \nabla_{xx} g(x, y_\gamma^*(x))$, $B(x) = \nabla_{xy} g(x, y_\gamma^*(x))$, $C(x) = \nabla_{yy} g(x, y_\gamma^*(x))$, $A_\gamma(x) = \gamma^{-1} \nabla_{xx} f(x, y_\gamma(x)) + \nabla_{xx} g(x, y_\gamma^*(x))$, $B_\gamma(x) = \gamma^{-1} \nabla_{xy} f(x, y_\gamma(x)) + \nabla_{xy} g(x, y_\gamma^*(x))$, $C_\gamma(x) = \gamma^{-1} \nabla_{yy} f(x, y_\gamma(x)) + \nabla_{yy} g(x, y_\gamma^*(x))$.

The proof of Lemma 3 is in Appendix [10, Sec. III-A]. Building on this, we seek to provide a tighter estimate for $l_{F,1}$ with the following conventional assumption [4], [14].

Assumption 4. Assume f, g are twice differentiable on $\mathcal{X} \times \mathcal{Y}$, and $\nabla^2 f, \nabla^2 g$ are respectively $l_{f,2}, l_{g,2}$ -Lipschitz in $y \in \mathcal{Y}$.

Theorem 1. Suppose $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(y) \leq 0\}$, and Assumption 1.1-2, 2.1-2, 3, 4 hold. Fix any $\delta > 0$, there exists some finite $\gamma^* > 0$ such that for any x and unit direction $d \in \mathbb{R}^{d_x}$, the directional derivative

$$\|D_{dd}^2(F_\gamma(x))\| \leq l_{F,1} = C_1 C_0 + \frac{1}{\gamma} C_2 C_0^2 + \frac{1}{\gamma^2} C_3 C_0^3 + \mathcal{O}(\delta),$$

for all $\gamma > \gamma^*$, where $C_1, C_2, C_3, C_4 = \mathcal{O}(1)$.

The proof for Theorem 1 is in Appendix [10, III-B]. In other word, the smoothness $l_{F,1} = \mathcal{O}(1)$ is not scalable with γ . This is consistent with the observation in Figure 1.

Algorithm 1 ALT-PBGD

1: **inputs:** initial point x_0 ; stepsize η ; counters T ; inner Min Solver.
2: **for** $t = 0, 1, \dots, T - 1$ **do**
3: update y_t^g as (12) by Min Solver.
4: update y_t^γ as (13) by Min Solver.
5: update $x_{t+1} = \text{Proj}_{\mathcal{X}}(x_t - \eta g_t)$ where g_t is in (14).
6: **end for**
7: **outputs:** (x_T, y_T^g)

B. ALT-PBGD: an improved PBGD method

In this section, we revisit the PBGD [20] method and demonstrate the effectiveness of its alternate version, ALT-PBGD, which updates y and x sequentially rather than jointly optimizing over (x, y) . At each iteration t , ALT-PBGD updates

$$y_t^g \approx \arg \min_{y \in \mathcal{Y}(x)} g(x, y), \quad (12)$$

$$y_t^\gamma \approx \arg \min_{y \in \mathcal{Y}(x)} \gamma^{-1} f(x, y) + g(x, y) \quad (13)$$

to ϵ -suboptimal points in distance metrics. Following Lemma 2 where the Lagrangian term is not involved in this setting, we can access the estimate of $\nabla F_\gamma(x_t) = \gamma(\nabla v_\gamma(x) - v(x))$ as

$$g_t = \nabla_x f(x, y_t^\gamma) + \gamma(\nabla_x g(x, y_t^\gamma) - \nabla_x g(x, y_t^g)), \quad (14)$$

and update $x_{t+1} = \text{Proj}_{\mathcal{X}}(x_t - \eta g_t)$ with $\eta \leq l_{F,1}^{-1}$. We outline the oracle in Algorithm 1 and present the complexity analysis in Proposition 2 with proof in Appendix [10, III-C].

Proposition 2. Consider $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(y) \leq 0\}$. Suppose Assumption 1, 2, 3, 4 hold. For $\gamma \geq \frac{l_{f,1}}{\mu_g}$, Algorithm 1 with $\eta = \mathcal{O}(1) \leq l_{F,1}^{-1}$ is achieved for $\mathcal{O}(\epsilon^{-1})$ outer-loop complexity for $\|G_\eta(x)\|^2 < \epsilon$, where $G_\eta(x) = \frac{x - \text{Proj}_{\mathcal{X}}(x - \eta \nabla F_\gamma(x))}{\eta}$.

The generalized gradient metric, $G_\eta(x)$, is common in constrained problems [5], [6], [15]. This Proposition enables setting large γ , e.g. $\gamma = \mathcal{O}(\epsilon^{-0.5})$ to bridge the equivalence in (4). Additionally, when PGD is chosen as the Min Solver, Algorithm 1 is of $\mathcal{O}(\epsilon^{-1} \ln(\epsilon^{-1})) = \tilde{\mathcal{O}}(\epsilon^{-1})$ overall complexity, as PGD converges linearly. This matches the optimal complexity of PGD for single-level optimization. This result highlights the advantage of minimizing $F_\gamma(x)$ over jointly minimizing $H_\gamma(x, y)$ in JNT-PBGD methods [20], since $H_\gamma(x, y)$ has a smoothness modulus $l_{H,1} = \mathcal{O}(\gamma)$, requiring $\eta = \mathcal{O}(\epsilon^{0.5})$ and leading to $\tilde{\mathcal{O}}(\epsilon^{-1.5})$ complexity, as also empirically corroborated in Figure 2.

IV. EXTENSION TO COUPLED CONSTRAINTS SETTING

This section addresses the general BLO problem with coupled inequality constraints $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(x, y) \leq 0\}$ in (1). As illustrated in Lemma 2, in the coupled constraint setting, $\nabla v(x)$ can be achieved by finding the solution $y_g^*(x)$ and the corresponding unique Lagrangian multiplier $\lambda_g^*(x)$. Therefore, the PBGD algorithm for solving Bi-Level Optimization with

Algorithm 2 BLOCC [11]

1: **inputs:** initial point x_0 ; stepsize η ; counters T ; inner MaxMin Solver.
2: **for** $t = 0, 1, \dots, T - 1$ **do**
3: update (λ_t^g, y_t^g) as (16) by MaxMin Solver.
4: update $(\lambda_t^\gamma, y_t^\gamma)$ as (15) by MaxMin Solver.
5: update $x_{t+1} = \text{Proj}_{\mathcal{X}}(x_t - \eta g_t)$ where g_t is in (17).
6: **end for**
7: **outputs:** (x_T, y_T^γ)

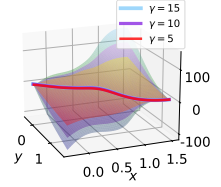


Fig. 1: $\nabla_x H_\gamma(x, y)$ for Example 1 with different γ . The lines represent $\nabla F_\gamma(x)$, showing its smaller variations.

Coupled Constraint, BLOCC [11], was developed similarly to ALT-PBGD. At each iteration t , find the ϵ -suboptimal solutions

$$(\lambda_{t+1}^g, y_{t+1}^g) \approx \arg \max_{\lambda \in \mathbb{R}^{d_c}} \min_{y \in \mathcal{Y}} \underbrace{g(x_t, y) + \langle \lambda, c(x_t, y) \rangle}_{=: L_g(x_t, y, \lambda)}, \quad (15)$$

$$(\lambda_{t+1}^\gamma, y_{t+1}^\gamma) \approx \arg \max_{\lambda \in \mathbb{R}^{d_c}} \min_{y \in \mathcal{Y}} \underbrace{\frac{1}{\gamma} f(x_t, y) + L_g(x_t, y, \lambda)}_{=: L_\gamma(x_t, y, \lambda)} \quad (16)$$

where $L_g(x, y, \lambda)$, and $L_\gamma(x, y, \lambda)$ are the Lagrangians for the two constrained problems in $F_\gamma(x)$ in (3). By Lemma 2, the estimate of $\nabla F_\gamma(x_t)$ can be achieved by finding $\nabla v(x_t)$ and $\nabla v_\gamma(x_t)$ through L_g and L_γ , i.e.

$$g_t = \gamma \nabla_x L_\gamma(x_t, y_t^\gamma, \lambda_t^\gamma) - \gamma \nabla_x L_g(x_t, y_t^g, \lambda_t^g). \quad (17)$$

Then, it updates $x_{t+1} = \text{Proj}_{\mathcal{X}}(x_t - \eta g_t)$ where step size $\eta \leq l_{F,1}^{-1}$. The algorithm is summarized in Algorithm 2.

[11] estimates the smoothness modulus of $F_\gamma(x)$ as $l_{F,1} = \mathcal{O}(\gamma)$, implying a choice $\eta = \mathcal{O}(\gamma^{-1})$. However, this estimate is not tight, as in the non-CC case. To address this, we provide a generalized version of Lemma 3 for the coupled constrained setting $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(x, y) \leq 0\}$ in Lemma 10 in Appendix [10, IV-A] under mild additional Assumption 5.

Assumption 5. The domain $\mathcal{Y}$ is smooth on the boundary and c is twice differentiable with $\nabla^2 c$ being $l_{c,1}$ -Lipschitz y .

In this way, the generalized version of Lemma 3 for the coupled constrained setting $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(x, y) \leq 0\}$ can be established, as per Lemma 10 in Appendix [10, IV-A]. This similarly help in concluding $\mathcal{O}(1)$ -smoothness of $F_\gamma(x)$.

Theorem 3. Consider $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(x, y) \leq 0\}$. Suppose Assumption 1, 2, 3, 4, 5 hold. Then, there exists finite $\gamma^* > 0$ such that $F_\gamma(x)$ is $l_{F,1} = \mathcal{O}(1)$ -smooth for all $\gamma > \gamma^*$.

The proof for Theorem 3 follows directly from Lemma 10 and is presented at the end of Appendix [10, IV-A]. This is a generalization to Theorem 1. It allows for $\eta = \mathcal{O}(1)$ stepsize choice of running BLOCC in the coupled constraint setting and results in a reduced complexity. As corroborated in Figure 3, increasing γ does not require decrease in η .

Proposition 4. Suppose all assumptions in Theorem 3 hold. For $\gamma \geq \frac{l_{f,1}}{\mu_g}$, Algorithm 1 with $\eta = \mathcal{O}(1) \leq l_{F,1}^{-1}$ is achieved for $\mathcal{O}(\epsilon^{-1})$ outer-loop complexity for $\|G_\eta(x)\|^2 < \epsilon$.

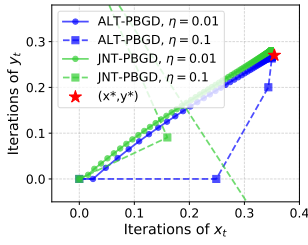


Fig. 2: Iterations of solving Example 1 via ALT-PBGD (Algorithm 1) and Joint-(JNT-)PBGD on $\min_{x,y} H_\gamma(x, y)$ [20, V-PBGD].

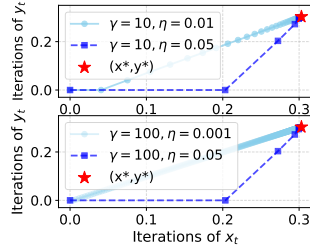


Fig. 3: Iterations of solving Example 2 in Appendix [10] via BLOCC (Algorithm 2) on $F_\gamma(x)$ with $\gamma = 10, 100$ and varying step-sizes.

Remark 1. The MaxMin Solver can be the accelerated version of Algorithm 2 in [11], therefore achieving $\mathcal{O}(\epsilon^{-2})$ overall complexity. For $\mathcal{Y} = \mathbb{R}^{d_y}$ and $c(x, y) = A(x)y + B(x)$ linear in y , the MaxMin Solver can be the fully single-loop version of Algorithm 2 in [11], achieving $\tilde{\mathcal{O}}(\epsilon^{-1})$ complexity.

The proof of Proposition 4 follows directly from [11], hence omitted. Here, $\eta = \mathcal{O}(1)$ choice leads to improved rate $T = \mathcal{O}(\epsilon^{-1})$, compared with $T = \mathcal{O}(\gamma\epsilon^{-1}) = (\epsilon^{-1.5})$ in [11].

V. CONCLUSION

This work tackles BLO with coupled constraints by using a penalty-based formulation that decouples UL and LL variables. By analyzing the Hessians of associated value functions, we establish that the reformulated objective maintains $\mathcal{O}(1)$ -smoothness under both non-coupled domain constraints and coupled inequality constraints. This enables to establish an improved iteration complexity of $\mathcal{O}(\epsilon^{-1})$ for the ALT-PBGD method, matching the optimal rate of standard gradient descent. Our results extend to BLO with general nonlinear constraints, offering a more efficient and scalable framework for solving bi-objective optimization problems. We provide numerical experiments in Appendix [10]. This advancement significantly corroborates the practicality of penalty-based methods in applications requiring constrained BLO.

REFERENCES

- [1] Sanjeev Arora, Simon Du, Sham Kakade, Yuping Luo, and Nikunj Saunshi. Provable representation learning for imitation learning via bi-level optimization. In *International Conference on Machine Learning*, pages 367–376. PMLR, 2020.
- [2] Jerome Bracken and James T McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973.
- [3] Lesi Chen, Yaohua Ma, and Jingzhao Zhang. Near-optimal nonconvex-strongly-convex bilevel optimization with fully first-order oracles. *arXiv preprint arXiv:2306.14853*, 2023.
- [4] Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hypergradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 947–980. PMLR, 2024.
- [5] Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *Advances in Neural Information Processing Systems*, Virtual, 2021.
- [6] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

- [7] Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.
- [8] 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.
- [9] Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pages 4882–4892. PMLR, 2021.
- [10] Liuyuan Jiang, Quan Xiao, and Tianyi Chen. Appendix for improved analysis of penalty-based methods for bilevel optimization with coupled constraints, 2025. <https://github.com/Liuyuan999/Improved-Analysis-of-Penalty-Based-Method-for-BLO-with-CC>.
- [11] Liuyuan Jiang, Quan Xiao, Victor M Tenorio, Fernando Real-Rojas, Antonio Marques, and Tianyi Chen. A primal-dual-assisted penalty approach to bilevel optimization with coupled constraints. In *Advances in Neural Information Processing Systems*, 2024.
- [12] Prashant Khanduri, Ioannis Tsaknakis, Yihua Zhang, Jia Liu, Sijia Liu, Jiawei Zhang, and Mingyi Hong. Linearly constrained bilevel optimization: A smoothed implicit gradient approach. In *International Conference on Machine Learning*, pages 16291–16325, 2023.
- [13] Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. In *Advances in Neural Information Processing Systems*, Virtual, 2021.
- [14] Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages 18083–18113, 2023.
- [15] Jeongyeol Kwon, Dohyun Kwon, Steve Wright, and Robert Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *International Conference on Learning Representations*, 2024.
- [16] Songtao Lu. SIm: A smoothed first-order lagrangian method for structured constrained nonconvex optimization. *Advances in Neural Information Processing Systems*, 36, 2023.
- [17] Akshay Mehra and Jihun Hamm. Penalty method for inversion-free deep bilevel optimization. In *Asian conference on machine learning*, pages 347–362, 2021.
- [18] Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *Proc. International Conference on Machine Learning*, New York City, NY, 2016.
- [19] Maria João Santos, Eduardo Curcio, Pedro Amorim, Margarida Carvalho, and Alexandra Marques. A bilevel approach for the collaborative transportation planning problem. *International Journal of Production Economics*, 233:108004, 2021.
- [20] Han Shen, Quan Xiao, and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, Honolulu, HI, 2023.
- [21] Mao Tan, Zhuocen Dai, Yongxin Su, Caixue Chen, Ling Wang, and Jie Chen. Bi-level optimization of charging scheduling of a battery swap station based on deep reinforcement learning. *Engineering Applications of Artificial Intelligence*, 118:105557, 2023.
- [22] Congying Wei, Qiwei Wu, Jian Xu, Yang Wang, and Yuanzhang Sun. Bi-level retail pricing scheme considering price-based demand response of multi-energy buildings. *International Journal of Electrical Power & Energy Systems*, 139:108007, 2022.
- [23] Quan Xiao, Han Shen, Wotao Yin, and Tianyi Chen. Alternating implicit projected sgd and its efficient variants for equality-constrained bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, 2023.
- [24] Siyuan Xu and Minghui Zhu. Efficient gradient approximation method for constrained bilevel optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
- [25] Wei Yao, Chengming Yu, Shangzhi Zeng, and Jin Zhang. Constrained bi-level optimization: Proximal lagrangian value function approach and hessian-free algorithm. *arXiv preprint arXiv:2401.16164*, 2024.
- [26] Jane J Ye, Daoli Zhu, and Qiji Jim Zhu. Exact penalization and necessary optimality conditions for generalized bilevel programming problems. *SIAM Journal on optimization*, 7(2):481–507, 1997.
- [27] Mao Ye, Bo Liu, Stephen Wright, Peter Stone, and Qiang Liu. Bome! bilevel optimization made easy: A simple first-order approach. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, 2022.