A Generic First-Order Algorithmic Framework for Bi-Level Programming Beyond Lower-Level Singleton

Risheng Liu [1 2] **Pan Mu** [1 2] **Xiaoming Yuan** [3] **Shangzhi Zeng** [3] **Jin Zhang** [4]

Abstract

In recent years, a variety of gradient-based bi-level optimization methods have been developed for learning tasks. However, theoretical guarantees of these existing approaches often heavily rely on the simplification that for each fixed upper-level variable, the lower-level solution must be a singleton (a.k.a., Lower-Level Singleton, LLS). In this work, by formulating bi-level models from the optimistic viewpoint and aggregating hierarchical objective information, we establish Bi-level Descent Aggregation (BDA), a flexible and modularized algorithmic framework for bi-level programming. Theoretically, we derive a new methodology to prove the convergence of BDA without the LLS condition. Furthermore, we improve the convergence properties of conventional first-order bi-level schemes (under the LLS simplification) based on our proof recipe. Extensive experiments justify our theoretical results and demonstrate the superiority of the proposed BDA for different tasks, including hyper-parameter optimization and meta learning.

1. Introduction

Bi-Level Programs (BLPs) are mathematical programs with optimization problems in their constraints and recently have been recognized as powerful theoretical tools for a variety of machine learning applications. Mathematically, BLPs can be (re)formulated as the following optimization problem:

$$\min_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathbb{R}^m} F(\mathbf{x},\mathbf{y}),\ s.t.\ \mathbf{y}\in\mathcal{S}(\mathbf{x}), \tag{1}$$

where the Upper-Level (UL) objective F is a jointly continuous function, the UL constraint $\mathcal{X}$ is a compact set, and the set-valued mapping $\mathcal{S}(\mathbf{x})$ indicates the parameterized solution set of the Lower-Level (LL) subproblem. Without loss of generality, we consider the following LL subproblem:

$$\mathcal{S}(\mathbf{x}) = \arg\min_{\mathbf{y}} f(\mathbf{x},\mathbf{y}), \tag{2}$$

where f is another jointly continuous function. Indeed, the BLPs model formulated in Eqs. (1)-(2) is a hierarchical optimization problem with two coupled variables $(\mathbf{x},\mathbf{y})\in\mathbb{R}^n\times\mathbb{R}^m$. Specifically, given the UL variable $\mathbf{x}$ from the feasible set $\mathcal{X}$ (i.e., $\mathbf{x}\in\mathcal{X}$), the LL variable $\mathbf{y}$ is an optimal solution of the LL subproblem governed by $\mathbf{x}$ (i.e., $\mathbf{y}\in\mathcal{S}(\mathbf{x})$). Due to the hierarchical structure and the complicated dependency between UL and LL variables, solving the above BLPs problem is challenging in general, especially when the LL solution set $\mathcal{S}(\mathbf{x})$ in Eq. (2) is not a singleton (Jeroslow, 1985; Dempe, 2018). In this work, we always call the condition that $\mathcal{S}(\mathbf{x})$ is a singleton as Lower-Level Singleton (or LLS for short).

1.1. Related Work

Although early works on BLPs can date back to the nineteen seventies (Dempe, 2018), it was not until the last decade that a large amount of bi-level optimization models were established to formulate specific machine learning problems, include meta learning (Franceschi et al., 2018; Rajeswaran et al., 2019; Zügner & Günnemann, 2019), hyper-parameter optimization (Franceschi et al., 2017; Okuno et al., 2018; MacKay et al., 2019), reinforcement learning (Yang et al., 2019), generative adversarial learning (Pfau & Vinyals, 2016), and image processing (Kunisch & Pock, 2013; De los Reyes et al., 2017), just to name a few.

A large number of optimization techniques have been developed to solve BLPs in Eqs. (1)-(2). For example, the works in (Kunapuli et al., 2008; Moore, 2010; Okuno et al., 2018) aimed to reformulate the original BLPs in Eqs. (1)-(2) as a single-level optimization problem based on the first-order optimality conditions. However, these approaches involve too many auxiliary variables, thus are not applicable for complex machine learning tasks.

[1]DUT-RU International School of Information Science and Engineering, Dalian University of Technology. [2]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province. [3]Department of Mathematics, The University of Hong Kong. [4]SUSTech International Center for Mathematics and Department of Mathematics, Southern University of Science and Technology. Correspondence to: Jin Zhang <zhangj9@sustech.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

Table 1. Comparing the convergence results (together with properties required by the UL and LL subproblems) between BDA and the existing bi-level FOMs in different scenarios (i.e., BLPs with and without LLS condition). Here $\xrightarrow{s}$ and $\xrightarrow{u}$ represent the subsequential and uniform convergence, respectively. The superscript * denotes that it is the true optimal variables/values.

Alg.		LLS	w/o LLS
Existing bi-level FOMs	UL	$F(\mathbf{x},\cdot)$ is Lipschitz continuous.	Not Available
	LL	$\{\mathbf{y}_K(\mathbf{x})\}$ is uniformly bounded on $\mathcal{X}$, $\mathbf{y}_K(\mathbf{x}) \xrightarrow{u} \mathbf{y}^*(\mathbf{x})$.	
	Main results: $\mathbf{x}_K \xrightarrow{s} \mathbf{x}^*$, $\inf_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) \to \inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x})$.		
BDA	UL	$F(\mathbf{x},\cdot)$ is Lipschitz continuous.	$F(\mathbf{x},\cdot)$ is Lipschitz continuous, L_F-smooth, and σ-strongly convex.
	LL	$\{\mathbf{y}_K(\mathbf{x})\}$ is uniformly bounded on $\mathcal{X}$, $f(\mathbf{x},\mathbf{y}_K(\mathbf{x})) \xrightarrow{u} f^*(\mathbf{x})$, $f(\mathbf{x},\mathbf{y})$ is level-bounded in $\mathbf{y}$ locally uniformly in $\mathbf{x}\in\mathcal{X}$.	$f(\mathbf{x},\cdot)$ is L_f-smooth and convex, $\mathcal{S}(\mathbf{x})$ is continuous.
	Main results: $\mathbf{x}_K \xrightarrow{s} \mathbf{x}^*$, $\inf_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) \to \inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x})$.		

Recently, gradient-based First-Order Methods (FOMs) have also been investigated to solve BLPs. The key idea underlying these approaches is to hierarchically calculate gradients of UL and LL objectives. Specifically, the works in (Maclaurin et al., 2015; Franceschi et al., 2017; 2018) first calculate gradient representations of the LL objective and then perform either reverse or forward gradient computations (a.k.a., automatic differentiation, based on the LL gradients) for the UL subproblem. It is known that the reverse mode is related to the back-propagation through time while the forward mode actually appears to the standard chain rule (Franceschi et al., 2017). In fact, similar ideas have also been used in (Jenni & Favaro, 2018; Zügner & Günnemann, 2019; Rajeswaran et al., 2019), but with different specific implementations. In (Shaban et al., 2019), a truncated back-propagation scheme is adopted to improve the scale issue for the LL gradient updating. Furthermore, the works in (Lorraine & Duvenaud, 2018; MacKay et al., 2019) trained a so-called hyper-network to map LL gradients for their hierarchical optimization.

Although widely used in different machine learning applications, theoretical properties of these bi-level FOMs are still not convincing (summarized in Table 1). Indeed, all of these methods require the LLS constraint in Eq. (2) to simplify their optimization process and theoretical analysis. For example, to satisfy such restrictive condition, existing works (Franceschi et al., 2018; Shaban et al., 2019) have to enforce a (local) strong convexity assumption to their LL subproblem, which is actually too tough to be satisfied in real-world complex tasks.

1.2. Our Contributions

This work proposes Bi-level Descent Aggregation (BDA), a generic bi-level first-order algorithmic framework that is flexible and modularized to handle BLPs in Eqs. (1)-(2). Unlike the above existing bi-level FOMs, which require the LLS assumption on Eq. (2) and separate the original model into two single-level subproblems, our BDA investigates BLPs from the optimistic viewpoint and develop a new hierarchical optimization scheme, which consists of a single-level optimization formulation for the UL variable $\mathbf{x}$ and a simple bi-level optimization formulation for the LL variable $\mathbf{y}$. Theoretically, we establish a general proof recipe to analyze the convergence behaviors of these bi-level FOMs. We prove that the convergence of BDA can be strictly guaranteed in the absence of the restrictive LLS condition. Furthermore, we demonstrate that the strong convexity of the LL objective (required in previous theoretical analysis (Franceschi et al., 2018)) is actually non-essential for these existing LLS-based bi-level FOMs, such as (Domke, 2012; Maclaurin et al., 2015; Franceschi et al., 2017; 2018; Shaban et al., 2019). Table 1 compares the convergence results of BDA and the existing approaches. It can be seen that in LLS scenario, BDA and the existing methods share the same requirements for the UL subproblem. However, for the LL subproblem, assumptions required in previous approaches are essentially more restrictive than that in BDA. More importantly, when solving BLPs without LLS, no theoretical results can be obtained for these classical methods. Fortunately, BDA can still obtain the same convergence properties as that in LLS scenario. The contributions can be summarized as:

- A counter-example (i.e., Example 1) explicitly indicates the importance of the LLS condition for the existing bi-level FOMs. In particular, we investigate their iteration behaviors and reach the conclusion that using these approaches in the absence of the LLS condition may lead to incorrect solutions.
- By formulating BLPs in Eqs. (1)-(2) from the viewpoint of optimistic bi-level, BDA provides a generic bi-level algorithmic framework. Embedded with a specific gradient-aggregation-based iterative module, BDA is applicable to a variety of learning tasks.
- A general proof recipe is established to analyze the convergence behaviors of bi-level FOMs. We strictly

prove the convergence of BDA without the LLS assumption. Furthermore, we revisit and improve the convergence properties of the existing bi-level FOMs in the LLS scenario.

2. First-Order Methods for BLPs

2.1. Solution Strategies with Lower-Level Singleton

As aforementioned, a number of FOMs have been proposed to solve BLPs in Eqs. (1)-(2). However, these existing methods all rely on the uniqueness of $\mathcal{S}(\mathbf{x})$ (i.e., LLS assumption). That is, rather than considering the original BLPs in Eqs. (1)-(2), they actually solve the following simplification:

$$\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x},\mathbf{y}),\ s.t.\ \mathbf{y}=\arg\min_{\mathbf{y}} f(\mathbf{x},\mathbf{y}), \tag{3}$$

where the LL subproblem only has one single solution for a given $\mathbf{x}$. By considering $\mathbf{y}$ as a function of $\mathbf{x}$, the idea behind these approaches is to take a gradient-based first-order scheme (e.g, gradient descent, stochastic gradient descent, or their variations) on the LL subproblem. Therefore, with the initialization point $\mathbf{y}_0$, a sequence $\{\mathbf{y}_k\}_{k=0}^K$ parameterized by $\mathbf{x}$ can be generated, e.g.,

$$\mathbf{y}_{k+1}=\mathbf{y}_k-s_l\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}_k),\ k=0,\cdots,K-1, \tag{4}$$

where $s_l>0$ is an appropriately chosen step size. Then by considering $\mathbf{y}_K(\mathbf{x})$ (i.e., the output of Eq. (4) for a given $\mathbf{x}$) as an approximated optimal solution to the LL subproblem, we can incorporate $\mathbf{y}_K(\mathbf{x})$ into the UL objective and obtain a single-level approximation model, i.e., $\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x},\mathbf{y}_K(\mathbf{x}))$. Finally, by unrolling the iterative update scheme in Eq. (4), we can calculate the derivative of $F(\mathbf{x},\mathbf{y}_K(\mathbf{x}))$ (w.r.t. $\mathbf{x}$) to optimize Eq. (3) by automatic differentiation techniques (Franceschi et al., 2017; Baydin et al., 2017).

2.2. Fundamental Issues and Counter-Example

It can be observed that the LLS condition fairly matters for the validation of the existing bi-level FOMs. However, such singleton assumption on the solution set of the LL subproblem is actually too restrictive to be satisfied, especially in real-world applications. In this subsection, we design an interesting counter-example (Example 1 below) to illustrate such invalidation of these conventional gradient-based bi-level schemes in the absence of the LLS condition.

Example 1. *(Counter-Example) With $\mathbf{x}\in[-100,100]$ and $\mathbf{y}\in\mathbb{R}^2$, we consider the following BLPs problem:*

$$\begin{aligned}&\min_{\mathbf{x}\in[-100,100]} \tfrac{1}{2}(\mathbf{x}-[\mathbf{y}]_2)^2+\tfrac{1}{2}([\mathbf{y}]_1-1)^2,\\ &s.t.\ \mathbf{y}\in\arg\min_{\mathbf{y}\in\mathbb{R}^2}\tfrac{1}{2}[\mathbf{y}]_1^2-\mathbf{x}[\mathbf{y}]_1,\end{aligned} \tag{5}$$

where $[\cdot]_i$ denotes the i-th element of the vector. By simple calculation, we know that the optimal solution of Eq. (5) *is $\mathbf{x}^*=1,\mathbf{y}^*=(1,1)$. However, if adopting the existing gradient-based scheme in Eq.* (4) *with initialization $\mathbf{y}_0=(0,0)$ and varying step size $s_l^k\in(0,1)$, we have that $[\mathbf{y}_K]_1=(1-\prod_{k=0}^{K-1}(1-s_l^k))\mathbf{x}$ and $[\mathbf{y}_K]_2=0$. Then the approximated problem of Eq.* (5) *amounts to*

$$\min_{\mathbf{x}\in[-100,100]} F(\mathbf{x},\mathbf{y}_K)=\frac{1}{2}\mathbf{x}^2+\frac{1}{2}((1-\prod_{k=0}^{K-1}(1-s_l^k))\mathbf{x}-1)^2.$$

By defining $\varphi_K(\mathbf{x})=F(\mathbf{x},\mathbf{y}_K)$, we have

$$\mathbf{x}_K^*=\arg\min_{\mathbf{x}\in[-100,100]}\phi_K(\mathbf{x})=\frac{(1-\prod_{k=0}^{K-1}(1-s_l^k))}{1+(1-\prod_{k=0}^{K-1}(1-s_l^k))^2}.$$

It is easy to check that

$$0\le\liminf_{K\to\infty}\prod_{k=0}^{K-1}(1-s_l^k)\le\limsup_{K\to\infty}\prod_{k=0}^{K-1}(1-s_l^k)\le 1,$$

then we have $\limsup_{K\to\infty}\frac{(1-\prod_{k=0}^{K-1}(1-s_l^k))}{1+(1-\prod_{k=0}^{K-1}(1-s_l^k))^2}\le\frac{1}{2}$. So $\mathbf{x}_K^*$ cannot converge to the true solution (i.e., $\mathbf{x}^*=1$).

Remark 1. *The UL objective F is indeed a function of both the UL variable $\mathbf{x}$ and the LL variable $\mathbf{y}$. Conventional bi-level FOMs only use the gradient information of the LL subproblem to update $\mathbf{y}$. Thanks to the LLS assumption, for fixed UL variable $\mathbf{x}$, the LL solution $\mathbf{y}$ can be uniquely determined. Thus the sequence $\{\mathbf{y}_k\}_{k=0}^K$ could converge to the true optimal solution, that minimizes both the LL and UL objectives. However, when LLS is absent, $\{\mathbf{y}_k\}_{k=0}^K$ may easily fail to converge to the true solution. Therefore, $\mathbf{x}_K^*$ may tend to be incorrect limiting points. Fortunately, we will demonstrate in Sections 3 and 5 that the example in Eq.* (5) *can be efficiently solved by our proposed BDA.*

3. Bi-level Descent Aggregation

In contrast to previous works, which only consider simplified BLPs with the LLS assumption in Eq. (3), we propose a new algorithmic framework, named Bi-level Descent Aggregation (BDA), to handle more generic BLPs in Eqs. (1)-(2).

3.1. Optimistic Bi-level Algorithmic Framework

In fact, the situation becomes intricate if the LL subproblem is not uniquely solvable for each $\mathbf{x}\in\mathcal{X}$. In this work, we consider BLPs from the optimistic bi-level viewpoint[1], thus for any given $\mathbf{x}$, we expect to choose the LL solution $\mathbf{y}\in\mathcal{S}(\mathbf{x})$ that can also lead to the best objective function value for the UL objective (i.e., $F(\mathbf{x},\cdot)$). Inspired by this observation, we can reformulate Eqs. (1)-(2) as

$$\min_{\mathbf{x}\in\mathcal{X}}\varphi(\mathbf{x}),\ \text{with}\ \varphi(\mathbf{x})=\inf_{\mathbf{y}\in\mathcal{S}(\mathbf{x})}F(\mathbf{x},\mathbf{y}). \tag{6}$$

[1]For more theoretical details of optimistic BLPs, we refer to (Dempe, 2018) and the references therein.

Such reformulation reduces BLPs to a single-level problem $\min_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x})$ w.r.t. the UL variable $\mathbf{x}$. While for any given $\mathbf{x}$, φ actually turns out to be the value function of a simple bi-level problem w.r.t. the LL variable $\mathbf{y}$, i.e.,

$$\min_{\mathbf{y}} F(\mathbf{x},\mathbf{y}),\ s.t.\ \mathbf{y}\in\mathcal{S}(\mathbf{x}),\ \text{(with fixed } \mathbf{x}). \tag{7}$$

Based on the above analysis, we actually could update $\mathbf{y}$ by

$$\mathbf{y}_{k+1}(\mathbf{x}) = \mathcal{T}_k(\mathbf{x},\mathbf{y}_k(\mathbf{x})),\ k=0,\cdots,K-1, \tag{8}$$

where $\mathcal{T}_k(\mathbf{x},\cdot)$ stands for a schematic iterative module originated from a certain simple bi-level solution strategy on Eq. (7) with a fixed UL variable $\mathbf{x}$.[2] Let $\mathbf{y}_0 = \mathcal{T}_0(\mathbf{x})$ be the initialization of the above scheme and denote $\mathbf{y}_K(\mathbf{x})$ as the output of Eq. (8) after K iterations (including the initial calculation $\mathcal{T}_0$). Then we can replace $\varphi(\mathbf{x})$ by $F(\mathbf{x},\mathbf{y}_K(\mathbf{x}))$ and obtain the following approximation of Eq. (6):

$$\min_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) = F(\mathbf{x},\mathbf{y}_K(\mathbf{x})). \tag{9}$$

With the above procedure, the BLPs in Eqs. (1)-(2) is approximated by a sequence of standard single-level optimization problems. For each approximation subproblem in Eq. (9), its descent direction is actually implicitly representable in terms of a certain simple bi-level solution strategy (i.e., Eq. (8)). Therefore, these existing automatic differentiation techniques all can be involved to achieve optimal solutions to Eq. (9) (Franceschi et al., 2017; Baydin et al., 2017).

3.2. Aggregated Iteration Modules

Now optimizing BLPs in Eqs. (1)-(2) reduces to the problem of designing proper $\mathcal{T}_k$ for Eq. (8). As discussed above, $\mathcal{T}_k$ is related to both the UL and LL objectives. So it is natural to aggregate the descent information of these two subproblems to design $\mathcal{T}_k$. Specifically, for a given $\mathbf{x}$, the descent directions of the UL and LL objectives can be defined as

$$\mathbf{d}_k^F(\mathbf{x}) = s_u\nabla_{\mathbf{y}}F(\mathbf{x},\mathbf{y}_k),$$
$$\mathbf{d}_k^f(\mathbf{x}) = s_l\nabla_{\mathbf{y}}f(\mathbf{x},\mathbf{y}_k),$$

where $s_u, s_l > 0$ are their step size parameters. Then we formulate $\mathcal{T}_k$ as the following first-order descent scheme:

$$\mathcal{T}_k(\mathbf{x},\mathbf{y}_k(\mathbf{x})) = \mathbf{y}_k - \Big(\alpha_k\mathbf{d}_k^F(\mathbf{x}) + (1-\alpha_k)\mathbf{d}_k^f(\mathbf{x})\Big), \tag{10}$$

where $\alpha_k\in(0,1)$ denotes the aggregation parameter.

Remark 2. *In this part, we just introduce a gradient aggregation based $\mathcal{T}_k$ to handle the simple bi-level subproblem in Eq.* (7)*. Indeed, our theoretical analysis in Section 4 will demonstrate that BDA algorithmic framework is flexible enough to incorporate a variety of numerical schemes. For example, in Supplemental Material, we also design an appropriate $\mathcal{T}_k$ to handle BLPs with nonsmooth LL objective while its convergence is still strictly guaranteed within our framework.*

[2] It can be seen that $\mathcal{T}_k$ actually should integrate the information from both the UL and LL subproblems in Eqs. (1)-(2). We will discuss specific choices of $\mathcal{T}_k$ in the following subsection.

4. Theoretical Investigations

In this section, we first derive a general convergence proof recipe together with two elementary properties to systematically investigate the convergence behaviors of bi-level FOMs (Section 4.1). Following this roadmap, the convergence of our BDA can successfully get rid of depending upon the LLS condition (Section 4.2). We also improve the convergence results for the existing bi-level FOMs in the LLS scenario (Section 4.3). To avoid triviality, hereafter we always assume that $\mathcal{S}(\mathbf{x})$ is nonempty for any $\mathbf{x}\in\mathcal{X}$. Please notice that all the proofs of our theoretical results are stated in the Supplemental Material.

4.1. A General Proof Recipe

We first state some definitions, which are necessary for our analysis.[3] A series of continuity properties for set-valued mappings and functions can be defined as follows.

Definition 1. *A set-valued mapping $\mathcal{S}(\mathbf{x}):\mathbb{R}^n\rightrightarrows\mathbb{R}^m$ is Outer Semi-Continuous (OSC) at $\bar{\mathbf{x}}$ if $\limsup_{\mathbf{x}\to\bar{\mathbf{x}}}\mathcal{S}(\mathbf{x})\subseteq\mathcal{S}(\bar{\mathbf{x}})$ and Inner Semi-Continuous (ISC) at $\bar{\mathbf{x}}$ if $\liminf_{\mathbf{x}\to\bar{\mathbf{x}}}\mathcal{S}(\mathbf{x})\supseteq\mathcal{S}(\bar{\mathbf{x}})$. $\mathcal{S}(\mathbf{x})$ is called continuous at $\bar{\mathbf{x}}$ if it is both OSC and ISC at $\bar{\mathbf{x}}$, as expressed by $\lim_{\mathbf{x}\to\bar{\mathbf{x}}}\mathcal{S}(\mathbf{x}) = \mathcal{S}(\bar{\mathbf{x}})$. Here $\limsup_{\mathbf{x}\to\bar{\mathbf{x}}}\mathcal{S}(\mathbf{x})$ and $\liminf_{\mathbf{x}\to\bar{\mathbf{x}}}\mathcal{S}(\mathbf{x})$ are defined as*

$$\limsup_{\mathbf{x}\to\bar{\mathbf{x}}}\mathcal{S}(\mathbf{x}) = \{\mathbf{y}|\ \exists\mathbf{x}^\nu\to\bar{\mathbf{x}},\exists\mathbf{y}^\nu\to\mathbf{y},\mathbf{y}^\nu\in\mathcal{S}(\mathbf{x}^\nu)\},$$
$$\liminf_{\mathbf{x}\to\bar{\mathbf{x}}}\mathcal{S}(\mathbf{x}) = \{\mathbf{y}|\ \forall\mathbf{x}^\nu\to\bar{\mathbf{x}},\exists\mathbf{y}^\nu\to\mathbf{y},\mathbf{y}^\nu\in\mathcal{S}(\mathbf{x}^\nu)\},$$

where $\nu\in\mathbb{N}$.

Definition 2. *A function $\varphi(\mathbf{x}):\mathbb{R}^n\to\mathbb{R}$ is Upper Semi-Continuous (USC) at $\bar{\mathbf{x}}$ if $\limsup_{\mathbf{x}\to\bar{\mathbf{x}}}\varphi(\mathbf{x})\le\varphi(\bar{\mathbf{x}})$, or equivalently $\limsup_{\mathbf{x}\to\bar{\mathbf{x}}}\varphi(\mathbf{x}) = \varphi(\bar{\mathbf{x}})$, and USC on $\mathbb{R}^n$ if this holds for every $\bar{\mathbf{x}}\in\mathbb{R}^n$. Similarly, $\varphi(\mathbf{x})$ is Lower Semi-Continuous (LSC) at $\bar{\mathbf{x}}$ if $\liminf_{\mathbf{x}\to\bar{\mathbf{x}}}\varphi(\mathbf{x})\ge\varphi(\bar{\mathbf{x}})$, or equivalently $\liminf_{\mathbf{x}\to\bar{\mathbf{x}}}\varphi(\mathbf{x}) = \varphi(\bar{\mathbf{x}})$, and LSC on $\mathbb{R}^n$ if this holds for every $\bar{\mathbf{x}}\in\mathbb{R}^n$. Here $\limsup_{\mathbf{x}\to\bar{\mathbf{x}}}\varphi(\mathbf{x})$ and $\liminf_{\mathbf{x}\to\bar{\mathbf{x}}}\varphi(\mathbf{x})$ are respectively defined as*

$$\limsup_{\mathbf{x}\to\bar{\mathbf{x}}}\varphi(\mathbf{x}) = \lim_{\delta\to0}\Big[\sup_{\mathbf{x}\in\mathbb{B}_\delta(\bar{\mathbf{x}})}\varphi(\mathbf{x})\Big]$$

and

$$\liminf_{\mathbf{x}\to\bar{\mathbf{x}}}\varphi(\mathbf{x}) = \lim_{\delta\to0}\big[\inf_{\mathbf{x}\in\mathbb{B}_\delta(\bar{\mathbf{x}})}\varphi(\mathbf{x})\big]$$

where $\mathbb{B}_\delta(\bar{\mathbf{x}}) = \{\mathbf{x}|\mathtt{dist}(\mathbf{x},\bar{\mathbf{x}})\le\delta\}$.

[3] Please also refer to (Rockafellar & Wets, 2009) for more details on these variational analysis properties.

Then for a given function $f(\mathbf{x}, \mathbf{y})$, we state the property that it is level-bounded in $\mathbf{x}$ locally uniform in $\mathbf{y}$ in the following definition.

Definition 3. *Given a function $f(\mathbf{x}, \mathbf{y}) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, if for a point $\bar{\mathbf{x}} \in \mathcal{X} \subseteq \mathbb{R}^n$ and $c \in \mathbb{R}$, there exist $\delta > 0$ along with a bounded set $\mathcal{B} \in \mathbb{R}^m$, such that*

$$\{\mathbf{y} \in \mathbb{R}^m \mid f(\mathbf{x}, \mathbf{y}) \leq c\} \subseteq \mathcal{B}, \ \forall \mathbf{x} \in \mathbb{B}_\delta(\bar{\mathbf{x}}) \cap \mathcal{X},$$

then we call $f(\mathbf{x}, \mathbf{y})$ is level-bounded in $\mathbf{y}$ locally uniformly in $\bar{\mathbf{x}} \in \mathcal{X}$. If the above property holds for each $\bar{\mathbf{x}} \in \mathcal{X}$, we further call $f(\mathbf{x}, \mathbf{y})$ level-bounded in $\mathbf{y}$ locally uniformly in $\mathbf{x} \in \mathcal{X}$.

Now we are ready to establish the general proof recipe, which describes the main steps to achieve the converge guarantees for our bi-level updating scheme (stated in Eqs. (8)-(9), with a schematic $\mathcal{T}_k$). Basically, our proof methodology consists of two main steps:

(1) **LL solution set property:** For any $\epsilon > 0$, there exists $k(\epsilon) > 0$ such that whenever $K > k(\epsilon)$,

$$\sup_{\mathbf{x} \in \mathcal{X}} \mathtt{dist}(\mathbf{y}_K(\mathbf{x}), \mathcal{S}(\mathbf{x})) \leq \epsilon.$$

(2) **UL objective convergence property:** $\varphi(\mathbf{x})$ is LSC on $\mathcal{X}$, and for each $\mathbf{x} \in \mathcal{X}$,

$$\lim_{K \to \infty} \varphi_K(\mathbf{x}) \to \varphi(\mathbf{x}).$$

Equipped with the above two properties, we can establish our general convergence results in the following theorem for the schematic bi-level scheme in Eqs. (8)-(9).

Theorem 1. *Suppose both the above LL solution set and UL objective convergence properties hold and let $\mathbf{x}_K \in \arg\min_{\mathbf{x} \in \mathcal{X}} \varphi_K(\mathbf{x})$. Then we have*

(1) Any limit point $\bar{\mathbf{x}}$ of the sequence $\{\mathbf{x}_K\}$ satisfies that $\bar{\mathbf{x}} \in \arg\min_{\mathbf{x} \in \mathcal{X}} \varphi(\mathbf{x})$.

(2) $\inf_{\mathbf{x} \in \mathcal{X}} \varphi_K(\mathbf{x}) \to \inf_{\mathbf{x} \in \mathcal{X}} \varphi(\mathbf{x})$ as $K \to \infty$.

Remark 3. *Indeed, if $\mathbf{x}_K$ is a local minimum of $\varphi_K(\mathbf{x})$ with uniform neighborhood modulus $\delta > 0$, we can still have that any limit point $\bar{\mathbf{x}}$ of the sequence $\{\mathbf{x}_K\}$ is a local minimum of $\varphi(\mathbf{x})$. Please see our Supplemental Material for more details on this issue.*

4.2. Convergence Properties of BDA

The objective here is to demonstrate that our BDA meets these two elementary properties required by Theorem 1. Before proving the convergence results for BDA, we first take the following as our blanket assumption.

Assumption 1. *For any $\mathbf{x} \in \mathcal{X}$, $F(\mathbf{x}, \cdot) : \mathbb{R}^m \to \mathbb{R}$ is L_0-Lipschitz continuous, L_F-smooth, and σ-strongly convex, $f(\mathbf{x}, \cdot) : \mathbb{R}^m \to \mathbb{R}$ is L_f-smooth and convex.*

Please notice that Assumption 1 is quite standard for BLPs in machine learning areas (Franceschi et al., 2018; Shaban et al., 2019). As can be seen, it is satisfied for all the applications considered in this work. We first present some necessary variational analysis preliminaries. Denoting

$$\tilde{\mathcal{S}}(\mathbf{x}) = \arg\min_{\mathbf{y} \in \mathcal{S}(\mathbf{x})} F(\mathbf{x}, \mathbf{y}),$$

under Assumption 1, we can quickly obtain that $\tilde{\mathcal{S}}(\mathbf{x})$ is nonempty and unique for any $\mathbf{x} \in \mathcal{X}$. Moreover, we can derive the boundedness of $\tilde{\mathcal{S}}(\mathbf{x})$ in the following lemma.

Lemma 1. *Suppose $F(\mathbf{x}, \mathbf{y})$ is level-bounded in $\mathbf{y}$ locally uniformly in $\mathbf{x} \in \mathcal{X}$. If $\mathcal{S}(\mathbf{x})$ is ISC on $\mathcal{X}$, then $\cup_{\mathbf{x} \in \mathcal{X}} \tilde{\mathcal{S}}(\mathbf{x})$ is bounded.*

Thanks to the continuity of $f(\mathbf{x}, \mathbf{y})$, we further have the following result.

Lemma 2. *Denote $f^*(\mathbf{x}) = \min_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$. If $f(\mathbf{x}, \mathbf{y})$ is continuous on $\mathcal{X} \times \mathbb{R}^m$, then $f^*(\mathbf{x})$ is USC on $\mathcal{X}$.*

Now we are ready to establish our fundamental LL solution set and UL objective convergence properties required in Theorem 1. In the following proposition, we first derive the convergence of $\{\mathbf{y}_K(\mathbf{x})\}$ in the light of the general fact stated in (Sabach & Shtern, 2017).

Proposition 1. *Suppose Assumption 1 is satisfied and let $\{\mathbf{y}_K\}$ be defined as in Eq. (10), $s_l \in (0, 1/L_f]$, $s_u \in (0, 2/(L_F + \sigma)]$,*

$$\alpha_k = \min\{2\gamma/k(1-\beta), 1-\varepsilon\},$$

with $k \geq 1$, $\varepsilon > 0$, $\gamma \in (0, 1]$, and

$$\beta = \sqrt{1 - 2s_u \sigma L_F/(\sigma + L_F)}.$$

Denote $\tilde{\mathbf{y}}_K(\mathbf{x}) = \mathbf{y}_K(\mathbf{x}) - s_l \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_K(\mathbf{x}))$, and

$$C_{\mathbf{y}^*(\mathbf{x})} = \max\left\{\|\mathbf{y}_0 - \mathbf{y}^*(\mathbf{x})\|, \frac{s_u}{1-\beta}\|\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\|\right\},$$

with $\mathbf{y}^*(\mathbf{x}) \in \tilde{\mathcal{S}}(\mathbf{x})$ and $\mathbf{x} \in \mathcal{X}$. Then we have

$$\|\mathbf{y}_K(\mathbf{x}) - \mathbf{y}^*(\mathbf{x})\| \leq C_{\mathbf{y}^*(\mathbf{x})},$$
$$\|\mathbf{y}_K(\mathbf{x}) - \tilde{\mathbf{y}}_K(\mathbf{x})\| \leq \frac{2C_{\mathbf{y}^*(\mathbf{x})}(J+2)}{K(1-\beta)},$$
$$f(\mathbf{x}, \tilde{\mathbf{y}}_K(\mathbf{x})) - f^*(\mathbf{x}) \leq \frac{2C^2_{\mathbf{y}^*(\mathbf{x})}(J+2)}{K(1-\beta)s_l},$$

where $J = \lfloor 2/(1-\beta) \rfloor$. Furthermore, for any $\mathbf{x} \in \mathcal{X}$, $\{\mathbf{y}_K(\mathbf{x})\}$ converges to $\tilde{\mathcal{S}}(\mathbf{x})$ as $K \to \infty$.

Proposition 1, together with Lemma 1, shows that $\{\tilde{\mathbf{y}}_K(\mathbf{x})\}$ is a bounded sequence and $\{f(\mathbf{x}, \tilde{\mathbf{y}}_K(\mathbf{x}))\}$ uniformly converges. We next prove the uniform convergence of $\{\tilde{\mathbf{y}}_K(\mathbf{x})\}$ towards the solution set $\mathcal{S}(\mathbf{x})$ through the uniform convergence of $\{f(\mathbf{x}, \tilde{\mathbf{y}}_K(\mathbf{x}))\}$.

Proposition 2. *Let $\mathcal{Y} \subseteq \mathbb{R}^m$ be a bounded set and $\epsilon > 0$. If $\mathcal{S}(\mathbf{x})$ is ISC on $\mathcal{X}$, then there exists $\delta > 0$ such that for any $\mathbf{y} \in \mathcal{Y}$,*

$$\sup_{\mathbf{x}\in\mathcal{X}} \operatorname{dist}(\mathbf{y}, \mathcal{S}(\mathbf{x})) \leq \epsilon,$$

in case $\sup_{\mathbf{x}\in\mathcal{X}} \{f(\mathbf{x}, \mathbf{y}) - f^*(\mathbf{x})\} \leq \delta$ is satisfied.

Combining Lemmas 1 and 2, together with Proposition 2, the *LL solution set* property required in Theorem 1 can be eventually derived. Let us now prove the LSC property of φ on $\mathcal{X}$ in the following proposition.

Proposition 3. *Suppose $F(\mathbf{x}, \mathbf{y})$ is level-bounded in $\mathbf{y}$ locally uniformly in $\mathbf{x} \in \mathcal{X}$. If $\mathcal{S}(\mathbf{x})$ is OSC at $\mathbf{x} \in \mathcal{X}$, then $\varphi(\mathbf{x})$ is LSC at $\mathbf{x} \in \mathcal{X}$.*

Then the *UL objective convergence* property required in Theorem 1 can be obtained subsequently based on Proposition 3, In summary, we present the main convergence results of BDA in the following theorem.

Theorem 2. *Suppose Assumption 1 is satisfied and let $\{\mathbf{y}_K\}$ be defined as in Eq. (10), $s_l \in (0, 1/L_f]$, $s_u \in (0, 2/(L_F + \sigma)]$,*

$$\alpha_k = \min\{2\gamma/k(1-\beta), 1-\varepsilon\},$$

with $k \geq 1$, $\varepsilon > 0$, $\gamma \in (0, 1]$, and

$$\beta = \sqrt{1 - 2s_u\sigma L_F/(\sigma + L_F)}.$$

Assume further that $\mathcal{S}(\mathbf{x})$ is continuous on $\mathcal{X}$. Then we have that both the LL solution set and UL objective convergence properties hold.

Remark 4. *Our proposed theoretical results are indeed general enough for BLPs in different application scenarios. For example, when the LL objective takes a nonsmooth form, e.g., $h = f + g$ with smooth f and nonsmooth g, we can adopt the proximal operation based iteration module (Beck, 2017) to construct $\mathcal{T}_k$ within our BDA framework. The convergence proofs are highly similar to that in Theorem 2. More details on such extension can be found in our Supplemental Material.*

4.3. Improving Existing LLS Theories

Although with the LLS simplification on BLPs in Eqs. (1)-(2), the theoretical properties of the existing bi-level FOMs are still not very convincing. Their convergence proofs in essence depend on the strong convexity of the LL objective, which may restrict the use of these approaches in complex machine learning applications. In this subsection, by weakening the required assumptions, we improve the convergence results in (Franceschi et al., 2018; Shaban et al., 2019) for these conventional bi-level FOMs in the LLS scenario. Specifically, we first introduce an assumption on the LL objective $f(\mathbf{x}, \mathbf{y})$, which is needed for our analysis in this subsection.

Assumption 2. *$f(\mathbf{x}, \mathbf{y}) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is level-bounded in $\mathbf{y}$ locally uniformly in $\mathbf{x} \in \mathcal{X}$.*

In fact, Assumption 2 is mild and satisfied by a large number of bi-level FOMs, when the LL subproblem is convex but not necessarily strongly convex. In contrast, theoretical results in existing literature (Franceschi et al., 2018; Shaban et al., 2019) require the more restrictive (local) strong convexity property on the LL objective to meets the LLS condition.

Under Assumption 2, the following lemma verifies the continuity of $\mathcal{S}(\mathbf{x})$ in the LLS scenario.

Lemma 3. *Suppose $\mathcal{S}(\mathbf{x})$ is single-valued on $\mathcal{X}$ and Assumption 2 is satisfied. Then $\mathcal{S}(\mathbf{x})$ is continuous on $\mathcal{X}$.*

As can be seen from the proof of Theorem 3 in our Supplemental Material, Lemma 3 and the uniform convergence of $\{f(\mathbf{x}, \mathbf{y}_K(\mathbf{x}))\}$ actually imply the *LL solution set* and *UL objective convergence* properties. Hence Theorem 1 is applicable, which inspires an improved version of the convergence results for the existing bi-level FOMs as follows.

Theorem 3. *Suppose $\mathcal{S}(\mathbf{x})$ is single-valued on $\mathcal{X}$ and Assumption 2 is satisfied, $\{\mathbf{y}_K(\mathbf{x})\}$ is uniformly bounded on $\mathcal{X}$, and $\{f(\mathbf{x}, \mathbf{y}_K(\mathbf{x}))\}$ converges uniformly to $f^*(\mathbf{x})$ on $\mathcal{X}$ as $K \to \infty$. Then we have that both the LL solution set and UL objective convergence properties hold.*

Remark 5. *Theorem 3 actually improves the converge results in (Franceschi et al., 2018). In fact, the uniform convergence assumption of $\{\mathbf{y}_K(\mathbf{x})\}$ towards $\mathbf{y}^*(\mathbf{x})$ required in (Franceschi et al., 2018) is essentially based on the strong convexity assumption (see Remark 3.3 of (Franceschi et al., 2018)). Instead of assuming such strong convexity, we only need to assume a weaker condition that $\{f(\mathbf{x}, \mathbf{y}_K(\mathbf{x}))\}$ converges uniformly to $f^*(\mathbf{x})$ on $\mathcal{X}$ as $K \to \infty$.*

It is natural for us to illustrate our improvement in terms of concrete applications. Specifically, we take the gradient-based bi-level scheme summarized in Section 2.1 (which has been used in (Franceschi et al., 2018; Jenni & Favaro, 2018; Shaban et al., 2019; Zügner & Günnemann, 2019; Rajeswaran et al., 2019)). In the following two propositions, we assume that $f(\mathbf{x}, \cdot) : \mathbb{R}^m \to \mathbb{R}$ is L_f-smooth and convex, and $s_l \leq 1/L_f$.

Inspired by Theorems 10.21 and 10.23 in (Beck, 2017), we first derive the following proposition.

Proposition 4. *Let $\{\mathbf{y}_K\}$ be defined as in Eq. (4). Then it holds that*

$$\|\mathbf{y}_K(\mathbf{x}) - \mathbf{y}^*(\mathbf{x})\| \leq \|\mathbf{y}_0 - \mathbf{y}^*(\mathbf{x})\|,$$

and

$$f(\mathbf{y}_K(\mathbf{x})) - f^*(\mathbf{x}) \leq \frac{\|\mathbf{y}_0 - \mathbf{y}^*(\mathbf{x})\|^2}{2s_l K},$$

with $\mathbf{y}^*(\mathbf{x}) \in \mathcal{S}(\mathbf{x})$ *and* $\mathbf{x} \in \mathcal{X}$.

Then in the following proposition we can immediately verify our required assumption on $\{f(\mathbf{x}, \mathbf{y}_K(\mathbf{x}))\}$ in the absence of the strong convexity property on the LL objective.

Proposition 5. *Suppose that $\mathcal{S}(\mathbf{x})$ is single-valued on $\mathcal{X}$ and Assumption 2 is satisfied. Let $\{\mathbf{y}_K\}$ be defined as in Eq.* (4)*. Then $\{\mathbf{y}_K(\mathbf{x})\}$ is uniformly bounded on $\mathcal{X}$ and $\{f(\mathbf{x}, \mathbf{y}_K(\mathbf{x}))\}$ converges uniformly to $f^*(\mathbf{x})$ on $\mathcal{X}$ as $K \to \infty$.*

Remark 6. *When the LL subproblem is convex, but not necessarily strongly convex, a large number of gradient-based methods, including accelerated gradient methods such as FISTA (Beck & Teboulle, 2009) and block coordinate descent method (Tseng, 2001), automatically meet our assumption, i.e., the uniform convergence of optimal values $\{f(\mathbf{x}, \mathbf{y}_K(\mathbf{x}))\}$ towards $f^*(\mathbf{x})$ on $\mathcal{X}$.*

5. Experimental Results

In this section, we first verify the theoretical findings and then evaluate the performance of our proposed method on different problems, such as hyper-parameter optimization and meta learning. We conducted these experiments on a PC with Intel Core i7-7700 CPU (3.6 GHz), 32GB RAM and an NVIDIA GeForce RTX 2060 6GB GPU.

5.1. Synthetic BLPs

Our theoretical findings are investigated based on the synthetic BLPs described in Section 2.2. As stated above, this deterministic bi-level formulation satisfies all the assumptions required in Section 4, but it cannot meet the LLS condition considered in (Finn et al., 2017; Franceschi et al., 2017; 2018; Shaban et al., 2019). Here, we fix the learning rate parameters $s_u = 0.7$ and $s_l = 0.2$ in this experiment.

In Figure 1, we plotted numerical results of BDA and one of the most representative bi-level FOMs (i.e., Reverse Hyper-Gradient (RHG) (Franceschi et al., 2017; 2018)) with different initialization points. We considered different numerical metrics, such as $|F - F^*|$, $|f - f^*|$, $\|\mathbf{x} - \mathbf{x}^*\|^2/\|\mathbf{x}^*\|^2$, and $\|\mathbf{y} - \mathbf{y}^*\|^2/\|\mathbf{y}^*\|^2$, for evaluations. It can be observed that RHG is always hard to obtain correct solution, even start from different initialization points. This is mainly because that the solution set of the LL subproblem in Eq. (5) is not a singleton, which does not satisfy the fundamental assumption of RHG. In contrast, our BDA aggregated the UL and LL information to perform the LL updating, thus we are able to obtain true optimal solution in all these scenarios. The initialization actually only slightly affected on the convergence speed of our iterative sequences.

Figure 2 further plotted the convergence behaviors of B-DA and RHG with different LL iterations (i.e., K). We observed that the results of RHG cannot be improved by increasing K. But for BDA, the three iterative sequences (with $K = 8, 16, 64$) are always converged and the numerical performance can be improved by performing relatively more LL iterations. In the above two figures, we set $\alpha_k = 0.5/k$, $k = 1, \cdots, K$.

Figure 3 evaluated the convergence behaviors of BDA with different choices of α_k. By setting $\alpha_k = 0$, we was unable to use the UL information to guide the LL updating, thus it is hard to obtain proper feasible solutions for the UL subproblem. When choosing a fixed α_k in $(0, 1)$ (e.g., $\alpha_k = 0.5$), the numerical performance can be improved but the convergence speed was still slow. Fortunately, we followed our theoretical findings and introduced an adaptive strategy to incorporate UL information into LL iterations, leading to nice convergence behaviors for both UL and LL variables.

5.2. Hyper-parameter Optimization

Hyper-parameter optimization aims choosing a set of optimal hyper-parameters for a given machine learning task. In this experiment, we consider a specific hyper-parameter optimization example, known as data hyper-cleaning (Franceschi et al., 2017; Shaban et al., 2019), to evaluate our proposed bi-level algorithm. In this task, we need to train a linear classifier on a given image set, but part of the training labels are corrupted. Following (Franceschi et al., 2017; Shaban et al., 2019), here we consider softmax regression (with parameters $\mathbf{y}$) as our classifier and introduce hyper-parameters $\mathbf{x}$ to weight samples for training.

Specifically, let $\ell(\mathbf{y}; \mathbf{u}_i, \mathbf{v}_i)$ be the cross-entropy function with the classification parameter $\mathbf{y}$ and data pairs $(\mathbf{u}_i, \mathbf{v}_i)$ and denote $\mathcal{D}_{\mathtt{tr}}$ and $\mathcal{D}_{\mathtt{val}}$ as the training and validation sets, respectively. Then we can define the LL objective as the following weighted training loss:

$$f(\mathbf{x}, \mathbf{y}) = \sum_{(\mathbf{u}_i, \mathbf{v}_i) \in \mathcal{D}_{\mathtt{tr}}} [\sigma(\mathbf{x})]_i \ell(\mathbf{y}; \mathbf{u}_i, \mathbf{v}_i),$$

where $\mathbf{x}$ is the hyper-parameter vector to penalize the objective for different training samples. Here $\sigma(\mathbf{x})$ denotes the element-wise sigmoid function on $\mathbf{x}$ and is used to constrain the weights in the range $[0, 1]$. For the UL subproblem, we define the objective as the cross-entropy loss with ℓ_2 regularization on the validation set, i.e.,

$$F(\mathbf{x}, \mathbf{y}) = \sum_{(\mathbf{u}_i, \mathbf{v}_i) \in \mathcal{D}_{\mathtt{val}}} \ell(\mathbf{y}(\mathbf{x}); \mathbf{u}_i, \mathbf{v}_i) + \lambda \|\mathbf{y}(\mathbf{x})\|^2,$$

where $\lambda > 0$ is the trade-off parameter and fixed as 10^{-4}.

We applied our BDA together with the baselines, RHG and Truncated RHG (T-RHG) (Shaban et al., 2019), to

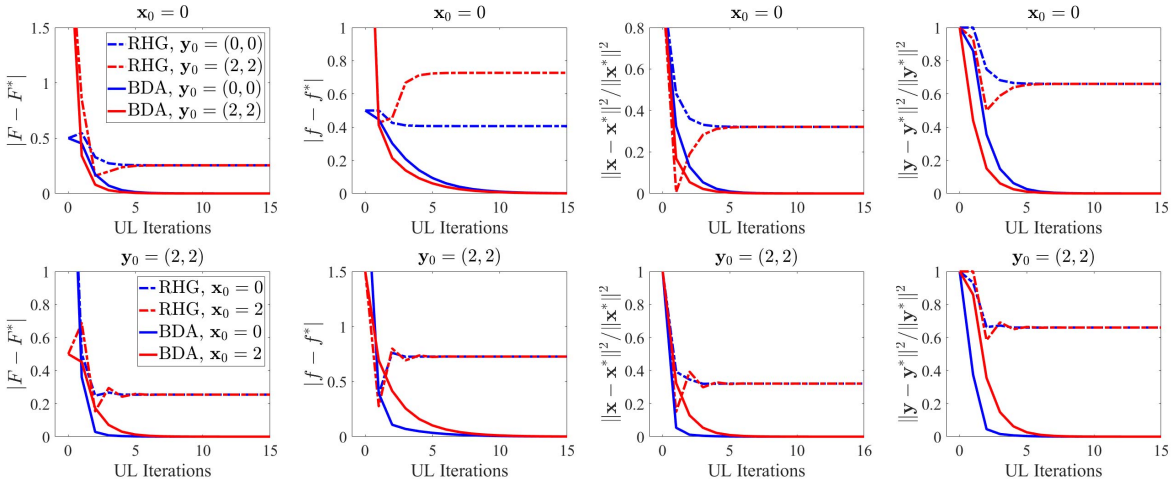

Figure 1. Illustrating the numerical performance of first-order BLPs algorithms with different initialization points. Top row: fix $\mathbf{x}_0 = 0$ and vary $\mathbf{y}_0 = (0,0), (2,2)$. Bottom row: fix $\mathbf{y}_0 = (2,2)$ and vary $\mathbf{x}_0 = 0, 2$. We fix $K = 16$ for UL iterations. The dashed and solid curves denote the results of RHG and BDA, respectively. The legend is only plotted in the first subfigure.

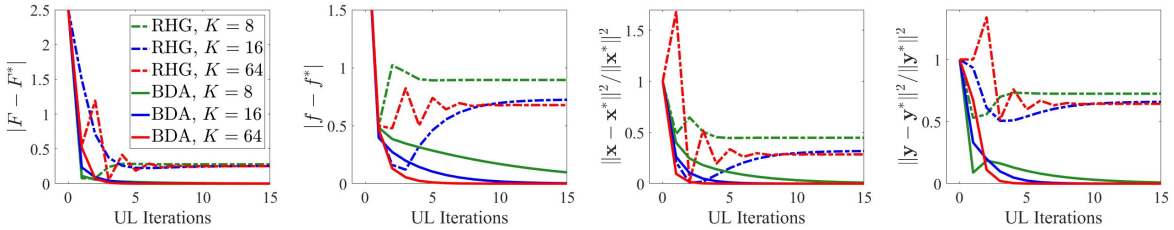

Figure 2. Illustrating the numerical performance of first-order BLPs algorithms with different LL iterations (i.e., $K = 8, 16, 64$). We fix initialization as $\mathbf{x}_0 = 0$ and $\mathbf{y}_0 = (2,2)$. The dashed and solid curves denotes the results of RHG and BDA, respectively. The legend is only plotted in the first subfigure.

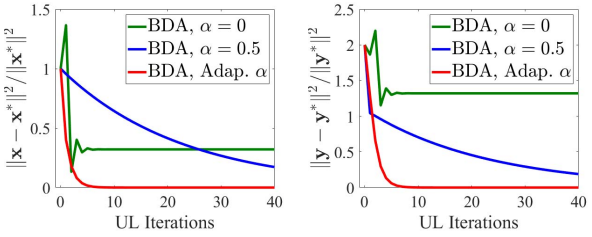

Figure 3. Illustrating the numerical performance of BDA with fixed α_k (e.g., $\alpha_k = 0, 0.5$) and adaptive α_k (e.g., $\{\alpha_k = 0.9/k\}$, denoted as "Adap. α"). The initialization and LL iterations are fixed as $\mathbf{x}_0 = 0$, $\mathbf{y}_0 = (2,2)$, and $K = 16$, respectively.

Table 2. Data hyper-cleaning accuracy of the compared methods with different number of LL iterations (i.e., $K = 50, 100, 200, 400, 800$) on MNIST.

Method	No. of LL Iterations (K)				
	50	100	200	400	800
RHG	88.96	89.73	90.13	90.19	90.15
T-RHG	87.90	88.28	88.50	88.52	89.99
BDA	**89.12**	**90.12**	**90.57**	**90.81**	**90.86**

solve the above BLPs problem on MNIST database (LeCun et al., 1998). Both the training and the validation sets consist of 7000 class-balanced samples and the remaining 56000 samples are used as the test set. We adopted the architectures used in RHG as the feature extractor for all the compared methods. For T-RHG, we chose 25-step truncated back-propagation to guarantee its convergence. Table 2 reported the averaged accuracy for all these compared methods with different number of LL iterations (i.e., $K = 50, 100, 200, 400, 800$). We observed that RHG outperformed T-RHG. While BAD consistently achieved the highest accuracy. Our theoretical results suggested that most of the improvements in BDA should come from the aggregations of the UL and LL information. The results also showed that more LL iterations are able to improve the final performances in most cases.

Table 3. The averaged few-shot classification accuracy on Omniglot ($N = 5, 20$ and $M = 1, 5$).

Method	5-way		20-way	
	1-shot	5-shot	1-shot	5-shot
MAML	98.70	**99.91**	95.80	98.90
Meta-SGD	97.97	98.96	93.98	98.40
Reptile	97.68	99.48	89.43	97.12
RHG	98.60	99.50	95.50	98.40
T-RHG	98.74	99.52	95.82	98.95
BDA	**99.04**	99.62	**96.50**	**99.10**

5.3. Meta Learning

The aim of meta learning is to learn an algorithm that should work well on novel tasks. In particular, we consider the few-shot learning problem (Vinyals et al., 2016; Qiao et al., 2018), where each task is a N-way classification and it is to learn the hyper-parameter $\mathbf{x}$ such that each task can be solved only with M training samples (i.e., N-way M-shot).

Following the experimental protocol used in recent works, we separate the network architecture into two parts: the cross-task intermediate representation layers (parameterized by $\mathbf{x}$) outputs the meta features and the multinomial logistic regression layer (parameterized by $\mathbf{y}^j$) as our ground classifier for the j-th task. We also collect a meta training data set $\mathcal{D} = \{\mathcal{D}^j\}$, where $\mathcal{D}^j = \mathcal{D}^j_{\mathtt{tr}} \cup \mathcal{D}^j_{\mathtt{val}}$ is linked to the j-th task. Then for the j-th task, we consider the cross-entropy function $\ell(\mathbf{x}, \mathbf{y}^j; \mathcal{D}^j_{\mathtt{tr}})$ as the task-specific loss and thus the LL objective can be defined as

$$f(\mathbf{x}, \{\mathbf{y}^j\}) = \sum_j \ell(\mathbf{x}, \mathbf{y}^j; \mathcal{D}^j_{\mathtt{tr}}).$$

As for the UL objective, we also utilize cross-entropy function but define it based on $\{\mathcal{D}^j_{\mathtt{val}}\}$ as

$$F(\mathbf{x}, \{\mathbf{y}^j\}) = \sum_j \ell(\mathbf{x}, \mathbf{y}^j; \mathcal{D}^j_{\mathtt{val}}).$$

Our experiments are conducted on two widely used benchmarks, i.e., Ominglot (Lake et al., 2015), which contains 1623 hand written characters from 50 alphabets and Mini-ImageNet (Vinyals et al., 2016), which is a subset of ImageNet (Deng et al., 2009) and includes 60000 downsampled images from 100 different classes. We followed the experimental protocol used in MAML (Finn et al., 2017) and compared our BDA to several state-of-the-art approaches, such as MAML (Finn et al., 2017), Meta-SGD (Li et al., 2018), Reptile (Nichol et al., 2018), RHG, and T-RHG.

It can be seen in Table 3 that BDA compared well to these methods and achieved the highest classification accuracy except in the 5-way 5-shot task. In this case, practical performance of BDA was slightly worse than MAML. We further conducted experiments on the more challenging MiniImageNet data set. In the second column of Table 4, we reported the averaged accuracy of three first-order BLPs based methods (i.e., RHG, T-RHG and BDA). Again, the performance of BDA is better than RHG and T-RHG. In the rightmost two columns, we also compared the number of averaged UL iterations when they achieved almost the same accuracy ($\approx 44\%$). These results showed that BDA needed the fewest iterations to achieve such accuracy.

Table 4. The few-shot classification performances on MiniImageNet ($N = 5$ and $M = 1$). The second column reported the averaged accuracy after converged. The rightmost two columns compared the UL Iterations (denoted as "UL Iter."), when achieving almost the same accuracy ($\approx 44\%$). Here "Ave. $\pm$ Var. (Acc.)" denotes the averaged accuracy and the corresponding variance.

Method	Acc.	Ave. $\pm$ Var. (Acc.)	UL Iter.
RHG	48.89	44.46 ± 0.78	3300
T-RHG	47.67	44.21 ± 0.78	3700
BDA	**49.08**	44.24 ± 0.79	**2500**

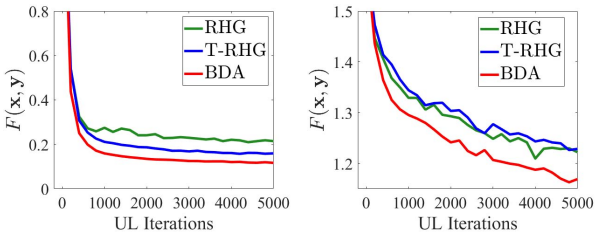

Figure 4. Illustrating the validation loss (i.e., UL objectives $F(\mathbf{x}, \mathbf{y})$) for three BLPs based methods on few-shot classification task. The curves in left and right subfigures are based on 5-way 1-shot results in Tables 3 and 4, respectively.

6. Conclusions

The proposed BDA is a generic first-order algorithmic scheme to address BLPs. We first designed a counter-example to indicate that the existing bi-level FOMs in the absence of the LLS condition may lead to incorrect solutions. Considering BLPs from the optimistic bi-level viewpoint, BDA could reformulate the original models in Eqs. (1)-(2) as the composition of a single-level subproblem (w.r.t. $\mathbf{x}$) and a simple bi-level subproblem (w.r.t., $\mathbf{y}$). We established a general proof recipe for bi-level FOMs and proved the convergence of BDA without the LLS assumption. As a nontrivial byproduct, we further improved convergence results for those existing schemes. Extensive evaluations showed the superiority of BDA for different applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 61922019, 61672125, 61733002, 61772105 and 11971220), LiaoNing Revitalization Talents Program (XLYC1807088), the Fundamental Research Funds for the Central Universities and the Natural Science Foundation of Guangdong Province 2019A1515011152. This work was also supported by the General Research Fund 12302318 from Hong Kong Research Grants Council.

References

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. Automatic differentiation in machine learning: a survey. *The Journal of Machine Learning Research*, 18 (1):5595–5637, 2017.

Beck, A. *First-order methods in optimization*. SIAM, 2017.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

De los Reyes, J. C., Schönlieb, C.-B., and Valkonen, T. Bilevel parameter learning for higher-order total variation regularisation models. *Journal of Mathematical Imaging and Vision*, 57(1):1–25, 2017.

Dempe, S. *Bilevel optimization: theory, algorithms and applications*. TU Bergakademie Freiberg Mining Academy and Technical University, 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Domke, J. Generic methods for optimization-based modeling. In *AISTATS*, pp. 318–326, 2012.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pp. 1126–1135, 2017.

Franceschi, L., Donini, M., Frasconi, P., and Pontil, M. Forward and reverse gradient-based hyperparameter optimization. In *ICML*, pp. 1165–1173, 2017.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *ICML*, pp. 1563–1572, 2018.

Jenni, S. and Favaro, P. Deep bilevel learning. In *ECCV*, pp. 618–633, 2018.

Jeroslow, R. G. The polynomial hierarchy and a simple model for competitive analysis. *Mathematical Programming*, 32(2):146–164, 1985.

Kunapuli, G., Bennett, K. P., Hu, J., and Pang, J.-S. Classification model selection via bilevel programming. *Optimization Methods & Software*, 23(4):475–489, 2008.

Kunisch, K. and Pock, T. A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 6(2):938–983, 2013.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, Z., Zhou, F., Chen, F., and Li, H. Meta-sgd: Learning to learn quickly for few-shot learning. In *ICML*, 2018.

Lorraine, J. and Duvenaud, D. Stochastic hyperparameter optimization through hypernetworks. *CoRR, abs/1802.09419*, 2018.

MacKay, M., Vicol, P., Lorraine, J., Duvenaud, D., and Grosse, R. Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions. *ICLR*, 2019.

Maclaurin, D., Duvenaud, D., and Adams, R. Gradient-based hyperparameter optimization through reversible learning. In *ICML*, pp. 2113–2122, 2015.

Moore, G. M. *Bilevel programming algorithms for machine learning model selection*. Rensselaer Polytechnic Institute, 2010.

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms. *CoRR, abs/1803.02999*, 2018.

Okuno, T., Takeda, A., and Kawana, A. Hyperparameter learning via bilevel nonsmooth optimization. *CoRR, abs/1806.01520*, 2018.

Pfau, D. and Vinyals, O. Connecting generative adversarial networks and actor-critic methods. In *NeurIPS Workshop on Adversarial Training*, 2016.

Qiao, S., Liu, C., Shen, W., and Yuille, A. L. Few-shot image recognition by predicting parameters from activations. In *CVPR*, pp. 7229–7238, 2018.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. In *NeurIPS*, pp. 113–124, 2019.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*. Springer Science & Business Media, 2009.

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

Shaban, A., Cheng, C.-A., Hatch, N., and Boots, B. Truncated back-propagation for bilevel optimization. In *AISTATS*, pp. 1723–1732, 2019.

Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. In *NeurIPS*, pp. 3630–3638, 2016.

Yang, Z., Chen, Y., Hong, M., and Wang, Z. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In *NeurIPS*, pp. 8351–8363, 2019.

Zügner, D. and Günnemann, S. Adversarial attacks on graph neural networks via meta learning. *ICLR*, 2019.