A General Descent Aggregation Framework for Gradient-based Bi-level Optimization

Risheng Liu, Member, IEEE, Pan Mu, Xiaoming Yuan, Shangzhi Zeng and Jin Zhang

Abstract—In recent years, a variety of gradient-based methods have been developed to solve Bi-Level Optimization (BLO) problems in machine learning and computer vision areas. However, the theoretical correctness and practical effectiveness of these existing approaches always rely on some restrictive conditions (e.g., Lower-Level Singleton, LLS), which could hardly be satisfied in real-world applications. Moreover, previous literature only proves theoretical results based on their specific iteration strategies, thus lack a general recipe to uniformly analyze the convergence behaviors of different gradient-based BLOs. In this work, we formulate BLOs from an optimistic bi-level viewpoint and establish a new gradient-based algorithmic framework, named Bi-level Descent Aggregation (BDA), to partially address the above issues. Specifically, BDA provides a modularized structure to hierarchically aggregate both the upper- and lower-level subproblems to generate our bi-level iterative dynamics. Theoretically, we establish a general convergence analysis template and derive a new proof recipe to investigate the essential theoretical properties of gradient-based BLO methods. Furthermore, this work systematically explores the convergence behavior of BDA in different optimization scenarios, i.e., considering various solution qualities (i.e., global/local/stationary solution) returned from solving approximation subproblems. Extensive experiments justify our theoretical results and demonstrate the superiority of the proposed algorithm for hyper-parameter optimization and meta-learning tasks. Source code is available at https://github.com/vis-opt-group/BDA.

Index Terms—Bi-level optimization, gradient-based method, descent aggregation, hyper-parameter optimization, meta-learning.

1 INTRODUCTION

Bi-Level Optimization (BLO) are a class of mathematical programs with optimization problems in their constraints. Recently, thanks to the powerful modeling capabilities, BLO have been recognized as important tools for a variety of machine learning and computer vision applications [1], [2], [3], [4], [5]. Mathematically, BLO can be formulated as

$$\min_{x \in X, y \in Y} F(x, y), \quad \text{s.t.} \quad y \in S(x) := \arg\min_y f(x, y),$$  \hspace{1cm} (1)

where the Upper-Level (UL) objective $F$ and the Lower-Level (LL) objective $f$ both are jointly continuous functions, the UL constraint $X$ is a compact set, the set-valued mapping $S(x)$ indicates the solution set of the LL subproblem parameterized by $x$, and $Y \subseteq \text{dom} F$ is a compact convex set. Indeed, the BLO model in Eq. (1) is a hierarchical optimization problem with two coupled variables $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ which need to be optimized simultaneously. This makes the computation of an optimal solution a challenging task [6]. To overcome such an unpleasant situation, from an optimistic BLO viewpoint, we decouple Eq. (1) as optimizing UL variable $x$ and LL variable $y$ separately. Specifically, for any given $x$, we expect that the LL solution $y \in S(x)$ also leads to the best UL objective value (i.e., $F(x, y)$) simultaneously. For this purpose, following the optimistic BLO idea, we incorporate some taste of hierarchy regarding the LL variable $y$, and Eq. (1) is thus reformulated as

$$\min_{x \in X} \phi(x), \quad \text{with} \quad \phi(x) := \inf_{y \in Y \cap S(x)} F(x, y).$$  \hspace{1cm} (2)

Actually, the above stated optimistic viewpoint is general and has received extensive attentions in BLO literature [8], [9], [10], [11]. Such reformulation reduces BLO to a single-level problem $\min_{x \in X} \phi(x)$ w.r.t. the UL variable $x$. Although early works on BLO can date back to the nineteen seventies [6], it was not until the last decade that a large amount of bi-level optimization models were established to capture vision and machine learning applications, including meta learning [12], [2], [13], hyper-parameter optimization [14], [15], [3], reinforcement learning [16], neural architecture search [17], [18], [19], [20] and image processing [4], [21], [22], [23], [5], and etc.

Due to the hierarchical structure and the sophisticated dependency between UL and LL variables, solving BLO is challenging in general, especially when the LL solution set $S(x)$ is not a singleton [24], [6]. Actually, the most straightforward idea in existing learning and vision literature is to assume that $S(x)$ is a singleton. Formally, we call the BLO model is with the Lower-Level Singleton (LLS) condition if

1. For more theoretical details of optimistic BLO, we refer to [6], [7] and the references therein.

---

Manuscript received April 19, 2020; revised August 26, 2015.

This work was supported by the National Key R&D Program of China (Grant No. 2018YFB1003200), and the National Natural Science Foundation of China (NSFC) (Grant No. 61802487).
∀x ∈ X. Under this condition, a variety of Gradient-based Bi-level Methods (GBMs) have been developed to solve BLOs in different machine learning and computer vision applications.

The key idea behind these existing GBMs is to solve BLOs with an approximated Best Response (BR) Jacobian (i.e., the gradient of the best response mapping w.r.t. the UL variable x). From this perspective, we can roughly categorize existing GBMs into two groups, i.e., explicit and implicit BR methods. For explicit BR methods, the BR gradients are obtained by automatic differentiation [25] through iterations of the LL gradient descent. This explicit structure mainly includes: recurrence-based [26], [14], [12], [27], initialization-based [28], [29] and proxy-based scheme [30], [3]. Specifically, recurrence-based BR 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. In [28], [29], known for its simplicity and state of the art performance, initialization-based structure estimated a good initialization of model parameters for the fast adaptation to new tasks purely by a gradient-based search. For proxy-based scheme [30], [3], a so-called hyper-network is trained to map LL gradients for their hierarchical optimization. These explicit methods only rely on the gradient information of the LL subproblem to update LL variable that cannot cover the UL descent information. On the other hand, implicit BR methods ([2], [31], [32], [33], [34] and [35]) are designed based on the observation that it is possible to replace the LL subproblem by an implicit equation. These implicit methods derive their BR gradients but involve computing a Hessian matrix and its inverse, which could be computationally expensive and unstable for large-scale problems.

Note that, for existing methods within the explicit BR category, the explicit BR approximation by optimization iteration dynamics raises an issue regarding approximation quality. In fact, without the LLS assumption, the dynamics procedures of existing methods, in general may not be good approximations. This is because in this case, the optimization dynamics converge to some minimizers of the LL objective but not necessarily to the one that also minimizes the UL objective. This unpleasant situation was noticed by both the machine learning and the optimization communities; see, e.g., [12, Section 3]. In theory, research on the theoretical convergence is still in its infancy (as summarized in Table 1). Indeed, all the mentioned GBMs require the LLS condition in LL subproblem to simplify their optimization processes and gain theoretical guarantees. For example, the works in [12], [27] enforce the strong convexity assumption to the LL subproblem. Unfortunately, it has been demonstrated that such LLS assumption is too restrictive to be satisfied in most real-world learning and vision applications. Further, these existing methods only concern the convergence towards stationary or global/local minimum, thus lack comprehensive convergence analyses.

In response to these limitations, this work proposes a novel framework termed Bi-level Descent Aggregation (BDA). Specifically, we propose a gradient type method for solving BLOs by aggregating UL and LL objectives. Theoretically, this work provides a general proof recipe as a basic template for the convergence analysis. In particular, in the absence of LLS, the BDA convergence was strictly guaranteed as long as the embedded inner simple bi-level dynamics meet the so-called UL objective convergence property and LL objective convergence property; see Section 4 for details. Specifically, we construct dynamics for optimizing the inner simple bi-level subproblem and hence achieve a justified good approximation. By using some variational analysis techniques sophisticatedly, the new optimization dynamics are shown to meet UL objective convergence property and LL objective convergence property without imposing any strong convexity assumptions in either UL or LL subproblems. Thanks to the new proof recipe, we provide the convergence results, which are classified by global/local solution cases returned from solving the approximation subproblems (i.e., min_x ϕ_K(x)). Besides, if solving an approximation subproblems (i.e., min_x ϕ_K(x)) returns (approximate) stationarity, we demonstrate the stationarity convergence result under the designed algorithm scheme (i.e., BDA). Moreover, as can be seen in Table 1, a striking feature of our study is that all the sufficient conditions we use to meet the desired convergence are easily verifiable for practical learning applications. We designed a high-dimensional counter-example with a series of complex experiments to verify our theoretical investigations and explore the intrinsic principles of the proposed algorithms. Extensive experiments also show the superiority of our method for different tasks, including hyper-parameter optimization and meta learning. We summarize the contributions of this work as follows.

- By designing a gradient-aggregation strategy to formulate the inner simple bi-level dynamics, we provide a new algorithmic framework to handle the LLS issue, which has been widely witnessed, but related research is still missing among existing gradient-based BLO approaches.
- We establish a general convergence analysis template together with an associated proof recipe for BDA. This new proof technique enhances our understanding of the essence of gradient-based method’s convergence, hence helps to eliminate the UL strong convexity assumption, which is required in [1].
- We provide a comprehensive theoretical convergence analysis of the developed algorithm. Focusing on different solution qualities (namely, global/local/stationary solutions), we elaborate the convergence properties respectively, thus significantly extend results in [1].
- As a nontrivial byproduct, the iterative gradient-aggregation dynamics (i.e., Eq. (9)) are of independent interest in convex optimization. They can be identified as a new iterative optimization scheme for solving the simple bi-level problem without the UL strong convexity.

2 Gradient-based BLOs: A Brief Review

As for the BLO model in Eq. (2), it is worthwhile noting that the LL solution set S(x) may have multiple solutions for every (or some) fixed x. However, it is challenging to solve BLO, especially when the LL solution set is not a singleton. Thus, in learning and vision application scenarios, the most...
straightforward idea of designing GBMs is to enforce the singleton assumption on $S(x)$. With such LLS condition, the BLO model in Eq. (2) actually can be simplified as follows:

$$\min_{x \in \mathcal{X}} \varphi(x) := F(x, y^*(x)), \ s.t. \ y^* = \arg\min_{y \in Y} f(x, y).$$

(3)

Thus the optimization task reduces to solve a single-level problem (i.e., $\min_{x \in \mathcal{X}} \varphi(x) = F(x, y^*(x))$) with an optimal LL variable $y^*$. In this way, the gradient of $\varphi$ (w.r.t., x) can be written as

$$\nabla \varphi(x) = \nabla_x F(x, y^*(x)) + \left( \frac{\partial y^*(x)}{\partial x} \right)^T \nabla_x F(x, y^*).$$

(4)

where “grad.” denotes the abbreviation of gradient and $(\cdot)^T$ means the transpose operation. In existing GBMs, they actually first numerically approximate $y^*$ by $y_K$ and thus define $\varphi_K(x) = F(x, y_K(x))$. Then the UL variable x can be updated based on the following practical formulation

$$\nabla \varphi_K(x) = \nabla_x F(x, y_K(x)) + \left( \frac{\partial y_K(x)}{\partial x} \right)^T \nabla_x F(x, y_K(x)).$$

(5)

In particular, a variety of techniques [14], [12], [27], [28], [29], [30], [3] have been developed to explicitly formulate $y_K$ using dynamic systems. For example, by enforcing the LLS assumption on the BLO problem and considering $x$ as the recurrent parameters of a gradient-based dynamic system, i.e., $y_k = T_k(x, y_{k-1})$ with $T_k(x, y_{k-1}) = y_k = 1 - \eta \nabla_y f(x, y_{k-1}(x))$, these methods first calculate gradient representations of the LL objective and then perform either reverse/forward or automatic differentiation to obtain Eq. (5). However, the dynamic system generated by these GBMs can only reveal gradients of the LL subproblem, but completely miss descent information from the UL objective.

Theoretically, these existing convergence results all require that the LL dynamics $(y_K(x))$ is uniformly bounded on $x$ and $y_K(x)$ uniformly converges to $\gamma^*(x)$ as $K \to \infty$. We should also point out that the LLS condition actually plays the key role for most of existing GBMs (e.g., [12], [35]). These approaches often require restrictive assumptions (e.g., strong convexity) to meet this assumption. Besides, some researches [3], [27], [36] prove that UL value-function converges to a first-order stationary point, i.e., $\lim_{T \to \infty} \nabla \varphi(x_T) = 0$. To achieve the stationarity, they require the first-order Lipschitz assumption for the UL and LL objectives, the twice continuously differentiable property for the LL objective and some additional restrictive assumptions, such as nonsingular Hessian assumption in [3].

### 3 Bi-level Descent Aggregation

In this section, we establish a general algorithmic framework to solve BLOs formulated in Eq. (2). In particular, by incorporating the numerical BR mapping $y_K(x)$ into Eq. (2), we actually aim to solve the following approximated single-level optimization model:

$$\min_{x \in \mathcal{X}} \varphi_K(x).$$

(6)

It should be emphasized that different from these existing GBMs stated above, which only use the information of the LL subproblem to generate $\varphi_K(x)$ (i.e., obtain $y_K(x)$), here we formulate it as the value function of the following inner simple bi-level model:

$$\min_{y \in Y} F(x, y), \ s.t. \ y \in S(x).$$

(7)

Let $T_k(x, \cdot)^4$ stand for a schematic iterative module originated from a certain simple bi-level solution strategy on Eq. (7) (with a fixed UL variable x). Then we can write the general updating rule of y as follows:

$$y_{k+1}(x) = T_{k+1}(x, y_k(x)), \ k = 0, \cdots, K - 1,$$

(8)

where $y_0(x) = y_0$ is the initialization based on $y_0 \in Y$. For the particular form of $T_k(x, \cdot)$, here we would like to aggregate both the UL and LL subproblems to define it. Specifically, for a given x, we write the descent directions of the UL and LL objectives as

$$d_k^F(x) = s_u \nabla_y F(x, y_k(x)), \ d_k^L(x) = s_l \nabla_y f(x, y_k(x)),$$

where $s_u$, $s_l$ denote the corresponding step size parameters. Then we consider the following aggregated updating scheme as $T_k(x, \cdot)$, i.e.,

$$T_{k+1}(x, y_k(x)) = \text{Proj}_{y} \left( y_k(x) - \left( \mu_{k} d_k^F(x) + (1 - \mu) \beta_k d_k^L(x) \right) \right),$$

(9)

where $\text{Proj}_{y}$ denotes the projection on $Y$, $\mu \in (0, 1)$ and $\alpha_k, \beta_k \in (0, 1]$ are the aggregation parameters and $k = 0, \cdots, K - 1$. Here we should point out that the iteration scheme in [1, Eq. (10)] is just a specific case of Eq. (9) with $\beta_{k} = (1 - \mu_{k})/(1 - \mu)$.

As for solving the single-level problem in Eq. (6), we state that this UL optimization step straightforwardly follows the standard (stochastic) gradient scheme, which has been widely investigated in literature; see, e.g., [26], [14], [12], [27]. To close this section, we summarize the overall BDA scheme in the following Algorithm 1.

#### Remark 1

First of all, we emphasize that it will be demonstrated in the following sections that the main scope of introducing set constraints (i.e., $x \in X$ and $y \in Y$) in Eq. (2) is to guarantee the completeness of our theoretical analysis. Thus in most optimization scenarios, we can straightforwardly define large enough $X$ and $Y$ (e.g., the whole space) to make the projection operation $\text{Proj}$ inactive during our iterations. Besides, even if it requires to explicitly consider the set constraints for some specific applications, we actually simply introduce Clarke subdifferential (see [37] for detailed definition) for the projection operation during iterations.

### 4 A General Convergence Analysis Recipe

This part aims to provide a general convergence analysis recipe for GBMs (not only BDA, but also these existing

4. In fact, our theoretical analysis in Section 4 will introduce two essential properties, which can be used as guidance for designing $T_k$. In other words, any $T_k$ satisfying these two properties all can be used as our fundamental modules.

5. When the identity of UL iteration step $t$ is clear from the context, we omit the superscript $t$ and write $x$ instead of $x^t$. To present the algorithm steps in a more explicit form, we provide detailed iterations with the superscript $t$ of $x$ in Algorithm 1; see the experiment in Sections 8.1.
Algorithm 1 Bi-level Descent Aggregation Framework

**Input:** The necessary parameters and initialization.

**Output:** The optimized \( x, y \).

1. \( t = 0 \).
2. **while** Not Converge do
3.     for \( k = 0 \) to \( K - 1 \) do
4.         \% LL updating (line 5-8)
5.         \( d_k^l(x^t) = s_k \nabla_y F(x^t, y_k(x^t)) \),
6.         \( d_k^t(x^t) = s_l \nabla_\lambda f(x_k^t, y_k(x^t)) \),
7.         \( \hat{y}_{k+1}(x^t) = y_k(x^t) - (\mu \alpha_k d_k^t(x^t) + (1 - \mu) \beta d_k^t(x^t)) \),
8.         \( y_{k+1}(x^t) = \text{proj}_Y(\hat{y}_{k+1}(x^t)) \).
9.     end for
10. \% UL updating (line 11)
11.     \( x^{t+1} = \text{proj}_X(x^t - \lambda \nabla \varphi_K(x^t)) \).
12. \( t = t + 1 \).
13. **end while**

 approaches. That is, we first introduce two essential convergence properties for the UL and LL subproblems and then establish a general convergence analysis template to investigate the theoretical properties of gradient-based bi-level iterations.  

To conduct the convergence analysis, we first make the following standing assumption.

**Assumption 1.** \( F(x, y), \nabla_y F(x, y), f(x, y) \) and \( \nabla_f f(x, y) \) are continuous on \( X \times \mathbb{R}^m \). For any \( x \in X \), \( F(x, \cdot) : \mathbb{R}^m \to \mathbb{R} \) is \( L_P \)-smooth, convex and bounded below by \( M_0, f(x, \cdot) : \mathbb{R}^m \to \mathbb{R} \) is \( L_f \)-smooth and convex.

4.1 Two Essential Convergence Properties

Now we are ready to establish the new convergence analysis template, which describes the main steps to achieve the converge guarantees for our bi-level updating scheme (stated in Eqs. (6)-(8), with a schematic \( T_k \)). Basically, our proof recipe is based on the following two essential properties:

1. **UL objective convergence property:** For each \( x \in X \),
   \[ \lim_{K \to \infty} \varphi_K(x) = \varphi(x). \]
2. **LL objective convergence property:** \( \{y_K(x)\} \) is uniformly bounded on \( X \), and for any \( \epsilon > 0 \), there exists \( k(\epsilon) > 0 \) such that whenever \( K > k(\epsilon) \),
   \[ \sup_{x \in X} \{ f(x, y_K(x)) - f^*(x) \} \leq \epsilon. \]

Indeed, the general recipe provides us a criterion to design different stable algorithms. Under these two essential properties, we thoroughly analyze the bi-level optimization problem and provide comprehensive theoretical results. Based on the developed BDA algorithm scheme, we first provide convergence results towards global and local minimum in Section 4.2 and Section 4.3 respectively. Specifically, if a series of global/local solutions of approximation subproblems are found, then a global solution of the original bi-level problem can be approximately achieved.

4.2 Towards Global Minimum

Thanks to the continuity of \( f(x, y) \), we have the same semi-continuity over partial minimization as in [1]. In other words, with the continuity of \( f(x, y) \), we have that \( f^*(x) := \min_x f(x, y) \) is Upper Semi-Continuous (USC for short) on \( X \). Equipped with the above two properties (i.e., UL objective convergence property and LL objective convergence property), we establish our general convergence results in the following theorem for the schematic bi-level scheme in Eqs. (6)-(8).

**Theorem 1.** (Convergence towards Global Minimum) Suppose both the above UL and LL objective convergence properties hold and \( f(x, y) \) is continuous on \( X \times \mathbb{R}^m \). Let \( x_K \) be a \( \epsilon_K \)-minimum of \( \varphi_K(x) \), i.e.,
   \[ \varphi_K(x_K) \leq \varphi_K(x) + \epsilon_K, \quad \forall x \in X. \]

Then if \( \epsilon_K \to 0 \), we have

1. Any limit point \( \bar{x} \) of the sequence \( \{x_K\} \) satisfies that \( \bar{x} \in \text{arg min}_{x \in X} \varphi(x) \).
2. \( \inf_{x \in X} \varphi_K(x) \to \inf_{x \in X} \varphi(x) \) as \( K \to \infty \).

**Proof.** For any limit point \( \bar{x} \) of the sequence \( \{x_K\} \), let \( \{x_t\} \) be a subsequence of \( \{x_K\} \) such that \( x_t \to \bar{x} \in X \). As \( \{y_K(x)\} \) is uniformly bounded on \( X \), we can have a subsequence \( \{x_m\} \) of \( \{x_t\} \) satisfying \( y_m(x_m) \to \bar{y} \) for some \( \bar{y} \). It follows from the LL objective convergence property that for any \( \epsilon > 0 \), there exists \( M(\epsilon) > 0 \) such that for any \( m > M(\epsilon) \), we have
   \[ f(x_m, y_m(x_m)) \leq f^*(x_m) \leq \epsilon. \]

By letting \( m \to \infty \), and since \( f(\cdot, \bar{y}) \) is USC on \( X \), we have \( f(\bar{x}, \bar{y}) = f^*(\bar{x}) \leq \epsilon \). As \( \epsilon \) is arbitrarily chosen, we have \( f(\bar{x}, \bar{y}) \leq f^*(\bar{x}) \leq 0 \) and thus \( \bar{y} \in S(\bar{x}) \). Next, as \( F(\bar{x}, \bar{y}) \) is continuous at \( (\bar{x}, \bar{y}) \), for any \( \epsilon > 0 \), there exists \( M(\epsilon) > 0 \) such that for any \( m > M(\epsilon) \), it holds
   \[ F(\bar{x}, \bar{y}) \leq F(x_m, y_m(x_m)) + \epsilon. \]

Then, we have, for any \( m > M(\epsilon) \) and \( x \in X \),
   \[ \varphi(\bar{x}) = \inf_{y \in S(\bar{x})} F(\bar{x}, y) \leq F(x_m, y_m(x_m)) + \epsilon \leq \varphi_m(\bar{x}) + \epsilon + \epsilon_m. \]

Taking \( m \to \infty \) and by the UL objective convergence property and \( \epsilon_m \to 0 \), we have
   \[ \varphi(\bar{x}) \leq \lim_{m \to \infty} \varphi_m(\bar{x}) + \epsilon + \epsilon_m = \varphi(\bar{x}) + \epsilon, \quad \forall \bar{x} \in X. \]

By taking \( \epsilon \to 0 \), we have \( \varphi(\bar{x}) \leq \varphi(x) \), \( \forall x \in X \) which implies \( \bar{x} \in \text{arg min}_{x \in X} \varphi(x) \).

We next show that \( \inf_{x \in X} \varphi_K(x) \to \inf_{x \in X} \varphi(x) \) as \( K \to \infty \). For any \( x \in X \), \( \inf_{x \in X} \varphi_K(x) \leq \varphi_K(x) \), by taking \( K \to \infty \) and with the UL objective convergence property, we have
   \[ \limsup_{K \to \infty} \left\{ \inf_{x \in X} \varphi_K(x) \right\} \leq \varphi(x), \quad \forall x \in X, \]

and thus
   \[ \limsup_{K \to \infty} \left\{ \inf_{x \in X} \varphi_K(x) \right\} \leq \inf_{x \in X} \varphi(x). \]

7. This subscript \( K \) just corresponds to the subscript of \( \varphi_K \).
Then we have that any limit point \( \bar{x} \) which implies a contradiction to Eq. (11). Thus we have \( \epsilon > 0 \) and subsequence \( \{ x_l \} \) of \( \{ x_K \} \) such that

\[
\inf_{x \in X} \phi(x) < \inf_{x \in X} \phi(x) - \delta, \quad \forall \delta > 0.
\]  
(11)

Since \( X \) is compact, we can assume without loss of generality that \( x_l \to \bar{x} \in X \) by considering a subsequence. Then, as shown in above, we have \( \bar{x} \in \arg\min_{x \in X} \phi(x) \). And, by the same arguments for deriving Eq. (10), we can show that \( \forall \epsilon > 0, \exists \delta > 0 \) such that \( \forall l > k(\epsilon), \) it holds

\[
\phi(\bar{x}) \leq \phi(x_l) + \epsilon.
\]

By letting \( l \to \infty, \epsilon \to 0 \) and the definition of \( x_l \), we have

\[
\inf_{x \in X} \phi(x) = \phi(\bar{x}) \leq \lim_{l \to \infty} \left\{ \inf_{x \in X} \phi(x) \right\},
\]

which implies a contradiction to Eq. (11). Thus we have \( \inf_{x \in X} \phi(x) \to \inf_{x \in X} \phi(x) \) as \( K \to \infty \). \( \square \)

### 4.3 Towards Local Minimum

**Theorem 2.** (Convergence towards Local Minimum) Suppose both the LL and UL objective convergence properties hold and let \( x_K \) be a local \( \epsilon \)-minimum of \( \phi_K(x) \) with uniform neighborhood modulus \( \delta > 0 \), i.e.,

\[
\phi_K(x_K) \leq \phi_K(x) + \epsilon, \quad \forall x \in B_{\delta}(x_K) \cap X.
\]

Then we have that any limit point \( \bar{x} \) of the sequence \( \{ x_K \} \) is a local minimum of \( \phi \), i.e., there exists \( \delta > 0 \) such that

\[
\phi(\bar{x}) \leq \phi(x), \quad \forall x \in B_{\delta}(\bar{x}) \cap X.
\]

**Proof.** For any limit point \( \bar{x} \) of the sequence \( \{ x_K \} \), let \( \{ x_l \} \) be a subsequence of \( \{ x_K \} \) such that \( x_l \to \bar{x} \in X \) and \( x_l \in B_{\delta/2}(x) \). As \( \{ y_K(x) \} \) is uniformly bounded on \( X \), we can have a subsequence \( \{ y_m \} \) of \( \{ x_l \} \) satisfying \( y_m(x_m) \to y \) for some \( y \). It follows from the LL objective convergence property that for any \( \epsilon > 0 \), there exists \( M(\epsilon) > 0 \) such that for any \( m > M(\epsilon) \),

\[
f(x_m, y_m(x_m)) - f^*(x_m) \leq \epsilon.
\]

By letting \( m \to \infty \), and since \( f \) is continuous and \( f^*(x) \) is USC on \( X \), we have

\[
f(\bar{x}, y) - f^*(\bar{x}) \leq \epsilon.
\]

As \( \epsilon \) is arbitrarily chosen, we have \( f(\bar{x}, y) - f^*(\bar{x}) \leq 0 \) and thus \( y \in S(x) \). Next, as \( F \) is continuous at \( (\bar{x}, y) \), for any \( \epsilon > 0 \), there exists \( M(\epsilon) > 0 \) such that for any \( m > M(\epsilon) \), it holds

\[
F(\bar{x}, y) \leq F(x_m, y_m(x_m)) + \epsilon.
\]

Then, we have, for any \( m > M(\epsilon) \) and \( x \in X \),

\[
\phi(\bar{x}) = \inf_{y \in S(x)} F(\bar{x}, y) \leq F(x_m, y_m(x_m)) + \epsilon = \phi_m(x_m) + \epsilon.
\]

Next, as \( x_m \) is a local \( \epsilon \)-minimum of \( \phi_m(x) \) with uniform neighborhood modulus \( \delta \), it follows

\[
\phi_m(x_m) \leq \phi_m(x) + \epsilon, \quad \forall x \in B_{\delta}(x) \cap X.
\]

Since \( B_{\delta/2}(x) \subseteq B_{\delta/2}(x_m) \), we have that for any \( \epsilon > 0, \forall x \in B_{\delta/2}(x) \cap X \), there exists \( M(\epsilon) > 0 \) such that whenever \( m > M(\epsilon) \),

\[
\phi_m(x_m) + \epsilon \leq \phi_m(x) + \epsilon + \epsilon.
\]

Taking \( m \to \infty \) and by the UL objective convergence property and \( \epsilon_m \to 0, \forall x \in B_{\delta/2}(x) \cap X \) we have

\[
\phi(x) \leq \lim_{m \to \infty} \phi_m(x) + \epsilon_m + \epsilon = \phi(x) + \epsilon.
\]

By taking \( \epsilon \to 0 \), we have

\[
\phi(x) \leq \phi(x), \quad \forall x \in B_{\delta/2}(x) \cap X,
\]

which implies \( \bar{x} \in \arg\min_{x \in B_{\delta/2}(x) \cap X} \phi(x) \), i.e., \( \bar{x} \) is a local minimum of \( \phi \). \( \square \)

Note that this work provides a series of approximate optimization problems to the bi-level problem, and we establish the convergence of such approximation problems to the original bi-level problem (i.e., Eq. (1)). Such kind of result is commonly used for characterizing the convergence of approximation type optimization method on nonconvex problems, see, for examples, Theorem 17.1 in book [38] for the convergence of the quadratic penalty function method and Theorem 7 in paper [39] for convergence of the interior point method.

### 5 Convergence Properties of BDA

With the above discussions in Section 3, the BLO is reduced to optimize a simple bi-level problem in Eq. (7) w.r.t. the LL variable \( y \), and subsequently solve a single-level problem in Eq. (6) w.r.t. the LL variable \( x \). This part analyzes the convergence behavior of the developed iterative algorithm. In other words, this part is devoted to show that our proposed BDA meets two convergence properties stated in Section 4 (i.e., UL objective convergence property and LL objective convergence property).

Following the above roadmap, convergence behaviors of gradient-based bi-level methods can be systematically investigated. The desired convergence results can be successfully achieved once the embedded task-tailored iterative gradient-aggregation modules \( T_k \) meet the UL objective convergence property and the LL objective convergence property.

#### 5.1 UL Convergence Properties

To investigate the convergence behavior of the proposed simple bi-level iterations \( T_k \) in Eq. (9), with fixed \( x \), we first introduce the following two auxiliary variables

\[
z_{k+1}(x) = y_k(x) - s_{\delta}\nabla F(x, y_k(x)),
\]

\[
z_{k+1}(x) = y_k(x) - s_{\delta}\nabla f(x, y_k(x)).
\]

We further denote the optimal value and the optimal solution set of simple bi-level problem (i.e., Eq. (7)) by \( \phi(x) \) and \( S(x) \), respectively.

As the identity of \( x \) is clear from the context, in Section 5.1 and 5.2, for succinctness we will write \( \Psi(y) \) instead of \( F(x, y) \), \( \Psi^\star \) instead of \( \phi(x) \), \( \psi(y) \) instead of \( f(x, y) \), \( S \) instead of \( S(x) \), and \( \delta \) instead of \( \delta(x) \). Moreover, we will omit the notation \( x \) and use the notations \( y_k, z_{k+1}^l \) and \( z_{k+1}^l \) instead of the \( y_k(x), z_{k+1}^l(x) \) and \( z_{k+1}^l(x) \), respectively.

With inner iterative module, this part demonstrate the convergence behavior of simple bi-level. We first provide a descent inequality of function value in the following lemma.
**Table 1** Comparing the convergence results between our method and existing GBMs in different scenarios (i.e., BLO w/ and w/o LLS condition).

| Alg.       | W/ LLS w/ UL strong convexity | W/ LLS w/o UL strong convexity | W/ UL strong convexity | W/o UL strong convexity |
|------------|-------------------------------|-------------------------------|-----------------------|------------------------|
| **Existing GBMs** |                               |                               |                       |                        |
| UL         | $F(x, \cdot)$ is Lipschitz continuous. | Not available                 |                       |                        |
| LL         | $y_N(x) \xrightarrow{\text{u}} y$ uniformly bounded on $X$, $f(y_N(x)) \xrightarrow{\text{u}} f(x)$. | Not available                 |                       |                        |
|            | Main results: $x_N \xrightarrow{\text{u}} x^*$, $\lim_{n \to \infty} f(x_N)$ is Lipschitz continuous. | Not available                 |                       |                        |
| **Ours**   |                               |                               |                       |                        |
| [1]        | $F(x, \cdot)$ is Lipschitz continuous. | $F(x, \cdot)$ is $L_F$-smooth and $\sigma$-strongly convex. | Not available         |                        |
| LL         | $y_N(x) \xrightarrow{\text{u}} y$ uniformly bounded on $X$, $f(y_N(x)) \xrightarrow{\text{u}} f(x)$. | $f(x, \cdot)$ is $L_F$-smooth and convex, $S(x)$ is continuous. | Not available         |                        |
|            | Main results: $x_N \xrightarrow{\text{u}} x^*$, $\lim_{n \to \infty} f(x_N)$ is Lipschitz continuous. |                                       |                       |                        |

Here $\xrightarrow{\text{u}}$ and $\to$ represent the subsequential and uniform convergence, respectively. The superscript * denotes that it is the true optimal variables/values. “conti.” and “diff.” denote continuously and differentiable respectively.

**Lemma 1.** Let $\{y_k\}$ be the sequence generated by Eq. (9) with $\alpha_k, \beta_k \in (0, 1)$, $s_k \in (0, \frac{1}{L'} + s_{1-k})$ and $\mu \in (0, 1)$, then for any $y \in Y$, we have

$$
(1 - \mu)\beta_k f(x, y) + \frac{s_k}{s_{1-k}} F(x, y) \geq (1 - \mu)\beta_k f(x, z_{k+1}) + \frac{s_k}{s_{1-k}} F(x, z_{k+1}) + \frac{\beta_k}{2L} (1 - \alpha_k s_k L_F) ||y_k - z_{k+1}\|^2 + \frac{\beta_k}{2L} ||y - y_{k+1}\|^2 + \frac{1}{2s_k} (1 - \beta_k s_k L_F) ||y_k - z_{k+1}||^2 - \frac{1}{2s_k} ||y - y_k||^2.
$$

(12)

**Proof.** It follows from the definitions of $z_{k+1}^u$ and $z_{k+1}^l$ that

$$
0 = \alpha_k \nabla \Psi(y_k) + \frac{x_{k+1}^u - y}{s_k} \text{ and } 0 = \beta_k \nabla \psi(y_k) + \frac{x_{k+1}^l - y}{s_l}.
$$

(13)

Thus, for any $y$, we have

$$
0 = \alpha_k \langle \nabla \Psi(y_k), y - z_{k+1}^u \rangle + \frac{z_{k+1}^u - y}{s_k}, y - z_{k+1}^u \rangle.
$$

(14)

$$
0 = \beta_k \langle \nabla \psi(y_k), y - z_{k+1}^l \rangle + \frac{z_{k+1}^l - y}{s_l}, y - z_{k+1}^l \rangle.
$$

(15)

As $\psi$ is convex and $\nabla \psi$ is Lipschitz continuous with constant $L_F$, we have

$$
\langle \nabla \psi(y_k), y - z_{k+1}^l \rangle \\
\quad = \langle \nabla \psi(y_k), y - y_k \rangle + \langle \nabla \psi(y_k), y_k - z_{k+1}^l \rangle \\
\quad \leq \psi(y) - \psi(y_k) + \psi(y_k) - \psi(z_{k+1}^l) + \frac{L}{2} ||y_k - z_{k+1}^l||^2 \\
\quad = \psi(y) - \psi(z_{k+1}^l) + \frac{L}{2} ||y_k - z_{k+1}^l||^2.
$$

(16)

Combining with $\langle x_{k+1} - y, y - z_{k+1}^u \rangle = \frac{1}{2} (||y - y_k||^2 - ||y - z_{k+1}^l||^2 - ||y_k - z_{k+1}^u||^2)$ and Eq. (15) yields

$$
\beta_k \psi(y) \geq \beta_k \psi(z_{k+1}^u) - \frac{\beta_k}{2s_l} ||y - y_k||^2 + \frac{1}{2s_l} ||y - z_{k+1}^u||^2.
$$

(17)

As $\Psi$ is convex and $\nabla \Psi$ is Lipschitz continuous with constant $L_F$, by similar arguments, we can have

$$
\alpha_k \Psi(y) \geq \alpha_k \Psi(z_{k+1}^u) - \frac{\alpha_k}{2s} ||y - y_k||^2 + \frac{1}{2s} ||y - z_{k+1}^u||^2.
$$

(18)

Multiplying Eq. (17) and Eq. (18) by $1 - \mu$ and $s_{1-k} s_k$, respectively, and then summing them up implies that

$$
(1 - \mu)\beta_k \psi(y) + \frac{\beta_k}{2L} (1 - \alpha_k s_k L_F) ||y_k - z_{k+1}^u||^2 + \frac{\beta_k}{2L} ||y - y_k||^2 + \frac{1}{2s_k} (1 - \beta_k s_k L_F) ||y_k - z_{k+1}^u||^2.
$$

(19)

By the convexity of $\|\cdot\|^2$, we have

$$
(1 - \mu) ||y - z_{k+1}^u||^2 + \mu ||y - z_{k+1}^u||^2 + \mu ||y - z_{k+1}^u||^2 \\
\geq ||y - ((1 - \mu)z_{k+1}^u + \mu z_{k+1}^u)||^2.
$$

(20)
Then, since $\alpha_k, \beta_k \leq 1$, we obtain form Eq. (19) that for any $y \in \mathcal{Y}$,
\[
(1 - \mu)\beta_k \psi(y) + \frac{\mu \alpha_k \alpha_s \Psi(y)}{s_t} \geq (1 - \mu)\beta_k \psi(y(z_{k+1}^1)) + \frac{\mu \alpha_k \alpha_s \Psi(y(z_{k+1}^1)) - 1}{s_t} \|y - y_k\|^2 + \frac{1}{2s_t} \|y - y_{k+1}\|^2 + \frac{1 - \mu}{2} (1 - \beta_k s_t L_f) \|y_k - z_{k+1}^1\|^2 + \frac{\mu}{2s_t} (1 - \alpha_k s_t L_F) \|y_k - z_{k+1}^1\|^2 + \frac{1}{2s_t} \|y_k - z_{k+1}^1\|^2 + \mu \alpha_k ^2 \alpha_s ^2 \|y - y_{k+1}\|^2.
\]

This completes the proof.

**Lemma 2.** Let $\{a_k\}$ and $\{b_k\}$ be sequences of non-negative real numbers. Assume that there exists $n_0 \in \mathbb{N}$ such that
\[
a_{k+1} + b_k - a_k \leq 0, \quad \forall k \geq n_0.
\]
Then $\lim_{k \to \infty} a_k$ exists and $\sum_{k=0}^{\infty} b_k < \infty$.

**Proof.** Adding the inequality $a_{k+1} + b_k - a_k \leq 0$, from $k = n_0$, to $k = n - 1$, we get $a_n + \sum_{k=n_0}^{n-1} b_k \leq a_{n_0}$. By letting $n \to \infty$, we get $\sum_{k=n_0}^{\infty} b_k < \infty$. As $\{a_k\}_{k \geq n_0}$ is a non-negative decreasing sequence, $\lim_{k \to \infty} a_k$ exists.

The above Lemma 2 aims to analyze sequence inequality that will be applied in the following Theorem. We explore the boundness of inner iterative sequence in the following Lemma 3.

**Lemma 3.** Let $\{y_k\}$ be the sequence generated by Eq. (9) with $\alpha_k \in (0, 1)$, $\beta_k \in (0, 1)$, $s_u \in (0, \frac{1}{L_f})$, $s_t \in (0, \frac{1}{L_f})$ and $\mu \in (0, 1)$, then for any $\bar{y} \in \mathcal{S}(x)$, we have
\[
\|z_{k+1}^1 - \bar{y}\| \leq \|y_k - \bar{y}\|.
\]

Furthermore, when $\mathcal{Y}$ is compact, sequences $\{y_k\}, \{z_{k}^1\}, \{z_{k}^u\}$ are all bounded.

**Proof.** According to [28, Proposition 4.8, Proposition 4.33, Corollary 18.16], we know that if $0 \leq \beta_s \leq \frac{1}{2s_t}, 0 \leq \alpha_s \leq \frac{1}{2s_t}$, operators $I - \beta_s \partial\psi$ and $I - \alpha_s \partial\psi$ are both nonexpansive (i.e., 1-Lipschitz continuous). Then, since $z_{k+1} = y_k - \beta_s \partial\psi(y_k)$ and $\bar{y} = \bar{y} - \beta_s \partial\psi(\bar{y})$ for any $\bar{y} \in \mathcal{S}$, we have
\[
\|z_{k+1} - \bar{y}\| = \|y_k - \beta_s \partial\psi(y_k) - \bar{y} + \beta_s \partial\psi(\bar{y})\| \leq \|y_k - \bar{y}\|.
\]

If $\mathcal{Y}$ is compact, then the desired boundedness of $\{y_k\}$ follows directly from the iteration scheme in Eq. (9). And it follows from $\|z_{k+1} - \bar{y}\| \leq \|y_k - \bar{y}\|$ that $\{z_{k}^1\}$ is bounded. Next, because
\[
\|z_{k+1}^u - (\bar{y} - \alpha_s s_u \partial\psi(\bar{y}))\| \leq \|y_k - \bar{y}\|,
\]
and $\alpha_k \in (0, 1)$, we have $\{z_{k}^u\}$ is bounded.

With the above lemmas, we are now ready to obtain the convergence result of our proposed algorithm in the following theorem.

**Theorem 3.** Let $\{y_k(x)\}$ be the sequence generated by Eq. (9) with $\alpha_k \in (0, 1)$, $\alpha_s \geq 0$, $\sum \alpha_s = +\infty$, $\beta_k \in [\beta, 1]$ with some $\beta > 0$, $s_u \in (0, \frac{1}{L_f})$, $s_t \in (0, \frac{1}{L_f})$ and $\mu \in (0, 1)$, suppose that $\mathcal{Y}$ is compact, for any given $x$, if $\mathcal{S}(x)$ is nonempty, we have
\[
\lim_{k \to \infty} \text{dist}(y_k(x), \mathcal{S}(x)) = 0,
\]
and then
\[
\lim_{k \to \infty} F(x, y_k(x)) = \varphi(x).
\]

**Proof.** Let $\delta > 0$ be a constant satisfying $\delta < \frac{1}{s_t} \min \{ (1 - \mu)(1 - s_u L_f), \mu(1 - s_u L_F) \}$. We consider a sequence of $\{\tau_n\}$ defined by
\[
\tau_n := \max \{ k \in \mathbb{N} | k \leq n \} \text{ and } \tau | \|y_{k-1} - z_{k}^1\|^2 + \|y_{k-1} - z_{k}^u\|^2 + \frac{\mu \alpha_k \alpha_s \alpha_u}{s_t} \leq \delta | \|y_{k-1} - z_{k}^1\|^2 + \|y_{k-1} - z_{k}^u\|^2 + \frac{\mu \alpha_k \alpha_s \alpha_u}{s_t} | \Psi(z_{k}^1) - \Psi^* | < \delta \}.
\]

Inspired by [41], we consider the following two cases: (a) $\{\tau_n\}$ is finite, i.e., there exists $k_0 \in \mathbb{N}$ such that
\[
\delta | \|y_{k-1} - z_{k}^1\|^2 + \frac{1}{4s_t} \|y_{k-1} - z_{k}^1\|^2 + \frac{\mu \alpha_k \alpha_s \alpha_u}{s_t} | \Psi(z_{k}^1) - \Psi^* | < \delta
\]
for all $k \geq k_0$; (b) $\{\tau_n\}$ is not finite, i.e., for all $k_0 \in \mathbb{N}$, there exists $k \geq k_0$, such that
\[
\delta | \|y_{k-1} - z_{k}^1\|^2 + \|y_{k-1} - z_{k}^u\|^2 + \frac{\mu \alpha_k \alpha_s \alpha_u}{s_t} | \Psi(z_{k}^1) - \Psi^* | < \delta
\]
for all $k \geq k_0$. Let $\bar{y}$ be any point in $\mathcal{S}$, setting $y = \bar{y}$ in Eq. (12), as $\psi(y) = \min_{y \in \mathcal{Y}} \psi(y)$, we have $\psi(y_{k+1}) = \psi(y_{k-1})$ and $\alpha_k, \beta_k \leq 1$, we have
\[
\|z_{k+1} - \bar{y}\|^2 \geq \frac{\mu \alpha_k \alpha_s \alpha_u}{s_t} | \Psi(z_{k}^1) - \Psi^* | < \delta \quad \text{for all } k \geq k_0.<n>
bounded. Since $\Psi$ is continuous and $\lim_{k \to \infty} \|y_k - z_{k+1}^\tau\| = 0$, there exists $k_0 \geq k_1$ such that $\|\Psi'(y_k) - \Psi'(z_{k+1}^\tau)\| < \epsilon$ for all $k \geq k_2$ and thus $\Psi'(z_{k+1}^\tau) - \Psi' \geq \epsilon$ for all $k \geq k_2$. Then we have

$$\epsilon \sum_{k=k_2}^\infty \alpha_k \leq \sum_{k=k_2}^\infty \alpha_k \left( \Psi'(z_{k+1}^\tau) - \Psi' \right) < \infty,$$

where the last inequality follows from $\sum_{k=0}^\infty \alpha_k \left( \Psi'(z_{k}^\tau) - \Psi' \right) < \infty$. This result contradicts to the assumption $\sum_{k=0}^\infty \alpha_k = +\infty$. As $\{y_k\}$ is bounded, we can assume without loss of generality that $\lim_{k \to \infty} y_k = \bar{y}$ by taking a subsequence. By the continuity of $\Psi$, we have $\Psi'(y) = \lim_{k \to \infty} \Psi'(y_k) \leq \Psi'$. Next, let $k = \ell$ and $\ell \to \infty$ in Eq. (13), by the continuity of $\nabla \psi$, $\beta_k \geq \beta > 0$, and $\lim_{k \to \infty} \|y_k - z_{k+1}^\tau\| = 0$, we have

$$0 \in \nabla \psi(\bar{y}),$$

and thus $\bar{y} \in \mathcal{S}$. Combining with $\Psi'(\bar{y}) \leq \Psi'$, we show that $\bar{y} \in \mathcal{S}$. Then by taking $\bar{y} = y$ and since $\lim_{k \to \infty} \|y_k - y_k\|^2 = 0$ and thus $\lim_{k \to \infty} \text{dist}(y_k, \mathcal{S}) = 0$.

Case (b): We assume that $\{\tau_n\}$ is not finite and for any $k_0 \in \mathbb{N}$, there exists $k \geq k_0$ such that $\delta \|\bar{y} - z_{k+1}^\tau\|^2 + \delta \|y_k - z_{k+1}^\tau\|^2 + \frac{\mu + \alpha}{\tau_k} \left( (1 - \mu)z_{k+1}^\tau + \mu z_{k+1}^\tau\right) - \|y_k - y\| < 0$. It follows from the assumption that $\tau_n$ is well defined for $n$ large enough and $\lim_{n \to \infty} \tau_n = +\infty$.

By setting $y = \text{Proj}_\delta(y_k)$ in Eq. (12), we have

$$\frac{1}{2\tau_k} \text{dist}^2(y_k, \bar{S}) \geq \frac{1}{2\tau_k} \text{dist}^2(y_k, \bar{S}) + \left(\frac{(1-\mu)(1-s\beta_L)}{1-\mu} - \delta\right) \|y_k - z_{k+1}^\tau\|^2 + \left(\frac{\mu(1-s\beta_L)}{1-\mu} - \delta\right) \|y_k - z_{k+1}^\tau\|^2 + \delta \|y_k - z_{k+1}^\tau\|^2 + \frac{\mu}{\tau_k} \left( (1 - \mu)z_{k+1}^\tau + \mu z_{k+1}^\tau\right) - \|y_k - y\|^2 \geq 0,$$

and thus $\bar{y} \in \mathcal{S}$. As $\Psi' > \Psi'(z_{k}^\tau)$ for all $k \in \{\tau_n\}$ and hence $\Psi' > \Psi'(z_{k}^\tau)$ for all $k$. Then it follows from the continuity of $\Psi$ and $\lim_{k \to \infty} \|z_{k+1}^\tau - y_k\| = 0$ that $\Psi' \geq \Psi'(y)$, which implies $\bar{y} \in \mathcal{S}$ and $\lim_{k \to \infty} h_k = 0$. Now, as we have shown above that $\bar{y} \in \mathcal{S}$ for any limit point $\bar{y}$ of $\{y_k\}$, we can obtain from the boundness of $\{y_k\}$ and $\{h_n\}$ that $\lim_{n \to \infty} h_n = 0$. Thus $\lim_{n \to \infty} \tau_n = n$, and $\lim_{k \to \infty} \text{dist}(y_k, \mathcal{S}) = 0$.

5.2 LL Convergence Properties

Specially, when we take $\alpha_k = 1/(k+1)$, we have the following uniformly complexity estimation. We first denote $D = \sup \|y - y'\|$, $M_{\rho} := \sup_{y, y' \in \mathcal{Y}} \|\nabla \psi F(x, y)\|$ and $M_{\rho} := \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} \|\nabla \psi F(x, y)\|$. And it should be notice that $D, M_{\rho}$ and $M_{\rho}$ are all finite when $\mathcal{X}$ and $\mathcal{Y}$ are compact.

Lemma 4. Let $\{y_k\}$ be the sequence generated by Eq. (9) with $\alpha_k = 1/(k+1)$, $\beta_k \in [\beta, 1]$ with some $\beta > 0$, $|\beta_k - \beta| \leq \frac{c_3}{(k+1)^2}$ with some $c_3 > 0$, $s_u \in (0, 1/\mathcal{L})$, $s_l \in (0, 1/\mathcal{L})$ and $\mu \in (0, 1)$, then for any $y \in \mathcal{S}(x)$, we have

$$\frac{(1-\mu)c_3}{(k+1)^2} \|y_k - y\|^2 \leq \frac{\mu}{h_k} \|y_k - y\|^2 + \frac{\mu}{h_k} \|y_k - y\|^2 \leq \frac{\mu}{h_k} \|y_k - y\|^2.$$

Proof. According to [40], Proposition 4.8, Proposition 4.33, Corollary 18.16, we know that when $0 \leq \beta_k s_l \leq \frac{1}{\mathcal{L}}$, $0 \leq \alpha_k s_u \leq \frac{1}{\mathcal{L}}$, operators $I - \beta_k s_l \nabla \psi$, $I - \alpha_k s_u \nabla \psi$ and $\text{Proj}_{y_k}$ are all nonexpansive (i.e., 1-Lipschitz continuous). Next, as

$$y_{k+1} = \text{Proj}_{y_k} \left( (1 - \mu)z_{k+1}^\tau + \mu z_{k+1}^\tau\right) - \beta_k s_l \nabla \psi(y_k)), \}

According to the definition of $\tau_n$, we have for all $k \in \{\tau_n\}$,

$$\frac{\mu}{h_k} \|y_k - y\|^2 + \frac{\mu}{h_k} \|y_k - y\|^2 \leq \frac{\mu}{h_k} \|y_k - y\|^2 \leq \frac{\mu}{h_k} \|y_k - y\|^2.$$
by denoting $\Delta^k_i := \alpha_k - \kappa_{k-1}$ and $\Delta^k_j := \beta_k - \beta_{k-1}$, we have the following inequality

$$\|y_{k+1} - y_k\|^2 \leq \|y_{k+1} - z_{k+1}^1\|^2 + (1 - \mu)\|z_{k+1}^1 - z_k^1\|^2,$$

$$\leq \|I - \alpha_k s_k \nabla \psi(y_k - y_{k-1})\|^2 + \mu s_k \|\Delta_k^1\|^2 \|\nabla \psi(y_{k-1})\|^2 + 2\mu s_k \|\Delta_k^1\|^2 \|y_k - y_{k-1}\||\nabla \psi(y_{k-1})|| + (1 - \mu) (I - \beta_k s_k \nabla \psi)(y_k - y_{k-1})\|^2$$

$$+ 2(1 - \mu) s_k \|\Delta_k^1\|^2 \|I - \beta_k s_k \nabla \psi(y_k - y_{k-1})\||\nabla \psi(y_{k-1})|| + (1 - \mu) s_k \|\Delta_k^1\|^2 \|y_{k-1} - z_k^1\|^2$$

$$\leq \|y_k - y_{k-1}\|^2 + 2\mu s_k \|\Delta_k^1\|^2 \|y_k - y_{k-1}\||\nabla \psi(y_{k-1})|| + \frac{\mu \|\Delta_k^1\|^2 s_k}{\epsilon_k} \|y_{k-1} - z_k^1\|^2$$

$$+ 2(1 - \mu) s_k \|\Delta_k^1\|^2 \|y_k - y_{k-1}\||\nabla \psi(y_{k-1})||,$$

where the first inequality follows from the nonexpansiveness of $\text{Proj}_Y$ and the convexity of $\|\cdot\|^2$, the second inequality comes from the definitions of $z_{k+1}^1, z_k^1$ and the last inequality follows from the nonexpansiveness of $I - \beta_k s_k \nabla \psi$ and $I - \alpha_k s_k \nabla \psi$ and the definitions of $z_{k+1}^1, z_k^1$. Then, since $\alpha_k \geq \beta_k \geq \beta > 0, |\beta_k - \beta_{k-1}| \leq \frac{1}{\epsilon_k}$, with some $\epsilon_k > 0$, $s_k \in (0, \frac{1}{\epsilon_k})$, $\beta_k \in (0, 1]$ and $\mu \in (0, 1)$. Suppose $\hat{S}(\chi)$ is nonempty, $\chi$ is compact, $F(x, \cdot)$ is bounded below by $M$, we have for $k \geq 0$,

$$\|y_k(x) - z_k^1(x)\|^2 \leq \frac{2(\gamma_{k+1} + C) + \|\nabla \psi(y_k - y_{k-1})\|^2}{\epsilon_k^2} \|z_{k+1}^1(x) - z_k^1(x)\|^2,$$

$$f(z_{k+1}^1(x) - f) \leq \frac{D^2}{\epsilon_k^2} \sum_{k=1}^{\infty} \frac{1}{\epsilon_k},$$

where $C_3 := \frac{D^2 + 2s_k(\chi(x) - M)}{2} + \frac{D^2}{\epsilon_k^2}$, $C_4 := \frac{C_2}{\epsilon_k^2}$, $C_k := \frac{C_2}{\epsilon_k^2}$, $C_0 := \max\{2 + c_3^2, 3\}$.

Proof. Let $y$ be any point in $S$, and set $y = \hat{y}$ as in Eq. (12), since $\psi(y) = \min_{y \in \mathbb{R}^d} \psi(y) \leq \psi(z_{k+1}^1)$, we have

$$\frac{1}{2} \|y - y_k\|^2 + \frac{\mu}{\epsilon_k} (\Psi(y) - \Psi(z_{k+1}^1))$$

$$\geq \frac{1}{2} \|y - y_{k+1}\|^2 + \frac{1}{2} (1 - \mu)(1 - \beta_k s_k L_f) \|y_k - z_{k+1}^1\|^2$$

$$+ \frac{\mu}{\epsilon_k} (1 - \alpha_k s_k L_f) \|y_k - z_k^1\|^2$$

$$+ \frac{1}{2} \left\|\left(1 - \beta_k s_k \nabla \psi \right) y_k - \left(1 - \beta_k s_k \nabla \psi \right) z_k^1\right\|^2$$

$$\leq \frac{1}{2} \|y - y_0\|^2 + \frac{1}{2} \left\|\left(1 - \beta_k s_k \nabla \psi \right) y_k - \left(1 - \beta_k s_k \nabla \psi \right) z_k^1\right\|^2$$

$$\leq \frac{1}{2} \|y - y_0\|^2 + \frac{\mu}{\epsilon_k} \left\|\Psi(y) - \Psi(z_{k+1}^1)\right\|^2$$

Adding the Eq. (28) from $k = 0$ to $k = n - 1$, and since $\alpha_k, \beta_k \in (0, 1]$, we have

$$\frac{1}{2} \|y - y_n\|^2 + \frac{1}{2} (1 - s_k L_f) \sum_{k=0}^{n-1} \|y_k - z_{k+1}^1\|^2$$

$$+ \frac{1}{2} (1 - s_k L_f) \sum_{k=0}^{n-1} \|y_k - z_k^1\|^2$$

$$+ \frac{1}{2} \sum_{k=0}^{n-1} \left\|\left(1 - \beta_k s_k \nabla \psi \right) y_k - \left(1 - \beta_k s_k \nabla \psi \right) z_k^1\right\|^2$$

$$\leq \frac{1}{2} \|y - y_0\|^2 + \frac{1}{2} \left\|\left(1 - \beta_k s_k \nabla \psi \right) y_k - \left(1 - \beta_k s_k \nabla \psi \right) z_k^1\right\|^2$$

$$\leq \frac{1}{2} \|y - y_0\|^2 + s_u(1 + n) \left\|\Psi(y) - M_0\right\|^2$$

(29)

Then it follows from Eq. (29) and Eq. (30) that

$$\|y_{n+1} - y_n\|^2 \leq \min\{1 - s_k L_f, 1 - s_k u L_f\} \|y_k - y_{k-1}\|^2$$

$$\leq \min\{1 - s_k L_f, 1 - s_k u L_f\} \|y_k - y_{k-1}\|^2$$

$$+ c_3^2 (1 - \mu)(1 - s_k L_f) \|y_k - z_{k+1}^1\|^2 + 2\mu s_k D_M F$$

$$+ \frac{\mu}{\epsilon_k} \left\|\Psi(y) - \Psi(z_{k+1}^1)\right\|^2 + 2\mu s_k D_M F$$

$$\leq \max\{2 + c_3^2, 3\} \|y - y_0\|^2 + 2\mu s_k D_M F$$

$$+ 2\mu s_k D_M F + 2\mu s_k D_M F$$

$$\leq \max\{2 + c_3^2, 3\} \|y - y_0\|^2 + 2\mu s_k D_M F$$

$$+ 2\mu s_k D_M F$$

where the second inequality comes from $\|y_k - y_{k+1}\| = (1 - \mu)(1 - \beta_k s_k L_f) \|y_k - z_{k+1}^1\| + (1 - \beta_k s_k L_f) \|y_k - z_k^1\| + \mu \|\beta_k s_k L_f \| \leq 2\|y - y_0\| \leq 2\|y - y_0\| \leq D.$

Then, we have

$$\frac{1}{\epsilon_k} \|z_{k+1}^1 - y_k\|^2 \leq \frac{1}{\epsilon_k} \|z_{k+1}^1 - y_k\|^2 + \frac{1}{\epsilon_k} \|z_{k+1}^1 - y_k\|^2$$

$$\leq \frac{1}{\epsilon_k} \|z_{k+1}^1 - y_k\|^2 + \|z_{k+1}^1 - y_k\|^2$$

$$\leq \frac{1}{\epsilon_k} \|z_{k+1}^1 - y_k\|^2 + \|z_{k+1}^1 - y_k\|^2$$

$$\leq \frac{1}{\epsilon_k} \|z_{k+1}^1 - y_k\|^2 + \frac{1}{\epsilon_k} \|z_{k+1}^1 - y_k\|^2$$

$$\leq \frac{1}{\epsilon_k} \|z_{k+1}^1 - y_k\|^2 + \frac{1}{\epsilon_k} \|z_{k+1}^1 - y_k\|^2$$

$$\leq \frac{1}{\epsilon_k} \|z_{k+1}^1 - y_k\|^2 + \|z_{k+1}^1 - y_k\|^2$$

(33)

where the second inequality follows from the definition of $z_k^1$ and the last inequality comes from $\|y_k - y_{k+1}\| \leq D$ and $\beta_k \geq \beta$. This implies that for any $n > n_0 > 0$,

$$\frac{1}{\epsilon_k} \|z_{k+1}^1 - y_n\|^2 \leq \|z_{k+1}^1 - y_n\|^2$$

$$\leq \frac{1}{\epsilon_k} \|z_{k+1}^1 - y_n\|^2$$

$$\leq \|z_{k+1}^1 - y_n\|^2$$

$$\leq \|z_{k+1}^1 - y_n\|^2$$

(34)
where the last inequality follows from Eq. (32) that \( \|y_k - y_{k-1}\|^2 \leq C \|z_k\|^2 \leq C \|z_{k+1}\|^2 \) for all \( k \geq n_0 \), and it can be easily verified that the above inequality holds when \( m = 1 \). By Eq. (29), we have

\[
\frac{1}{2} (1 - \mu) (1 - s_i L_f) \sum_{k=0}^{n-1} \|y_k - z_{k+1}\|^2 \leq \frac{1}{2} \|y - y_0\|^2 + s_u (1 + \ln n) (\Psi^* - M_0).
\]

Then, for any \( n \), let \( m \) be the smallest integer such that \( m \geq n \), and let \( n_0 = n - m + 1 \), combining the above inequality with Eq. (34), we have

\[
\|y - y_0\|^2 + 2s_u (1 + \ln n) (\Psi^* - M_0) \geq \sum_{k=0}^{n} \|y_k - z_{k+1}\|^2 \geq m \beta^2 \|y_n - z_{n+1}\|^2 - C_2 \frac{(m-\alpha_n)(1+\ln n)}{\sqrt{n}}.
\]

where \( C_2 := (s^2 L^2 D + \frac{1}{2} (\mu - \beta)^2) \frac{\sqrt{T}}{\sqrt{n}} \).

Next, as \( n^4 + 1 \geq m \geq n^4 \), and hence \( n_0 \geq (m - 1)^4 - m + 1 \). Then \( 16n_0 - 2mn - (m - 1)(m - 1)(3m - 4)(5m - 4) - 1 > 0 \) when \( m \geq 2 \). Thus, when \( n \geq 2 \), we have

\[
\|y_n - z_{n+1}\|^2 \leq \frac{m \beta^2}{2} \|y_n - z_{n+1}\|^2 \leq 2(1 + \ln n) \frac{C_3}{\sqrt{n}}.
\]

where the last inequality follows from \( \sqrt{1 + \ln n} \leq 1 + \ln n \) and \( m \geq n^4 \). By the convexity of \( \psi \), and \( y_n - z_{n+1} = \beta_n s_i \nabla \psi (y_n) \), we have

\[
\psi (y_n) \leq \psi (\hat{y}) + \langle \nabla \psi (y_n), y_n - \hat{y} \rangle \leq \min_{y} \psi + \frac{1}{\beta_n s_i} \|y_n - z_{n+1}\|^2 \|y_n - \hat{y}\| \leq \min_{y} \psi + \frac{D}{\beta_n s_i} \sqrt{2C_2 + C_3 \frac{1}{\sqrt{n}}},
\]

This complete the proof. \( \square \)

5.3 Approximation Quality and Convergence of BDA

This part is devoted to the justification of the approximation quality and hence the convergence of our bi-level updating scheme (stated in Eqs. (6)-(8), with embedded \( T_k \) in Eq. (9)). Following the general proof recipe, we only need to verify that the convergence of \( T_k \) in Eq. (9) meets the LL objective convergence property and the LL objective convergence property.

Theorem 5. Suppose Assumptions 1 is satisfied, \( \mathcal{X} \) and \( \mathcal{Y} \) are compact, and \( \mathcal{S}(x) \) is nonempty for all \( x \in \mathcal{X} \). Let \( \{ y_k \} \) be the output generated by (9) with \( s_i \in (0, 1/L_f), \mu \in (0, 1), \alpha_k = \frac{1}{k+1}, \beta_k \in [\beta, 1] \) with some \( \beta > 0, \| \beta_k - \beta_k \| \leq \frac{c_0}{(k+1)^2} \) with some \( c_0 > 0 \), then we have that both the LL and UL objective convergence properties hold.

Proof. Since \( \mathcal{X} \) and \( \mathcal{Y} \) are both compact, and \( F(x, y) \) is continuous on \( \mathcal{X} \times \mathcal{Y} \), we have that \( F(x, y) \) is uniformly bounded above on \( \mathcal{X} \times \mathcal{Y} \) and thus \( \min_{y \in \mathcal{Y}} \mathcal{S}(x) \) \( F(x, y) \) is uniformly bounded above on \( \mathcal{X} \). And combining with the assumption that \( F(x, y) \) is uniformly bounded below with respect to \( y \) by \( M_0 \) for any \( x \in \mathcal{X}, \mathcal{Y} \) is compact, we can obtain from the Theorem 4 that there exists \( C > 0 \) such that for any \( x \in \mathcal{X} \), we have

\[
f(x, y_k(x)) - f^*(x) \leq C \sqrt{\frac{1 + \ln K}{K^2}}.
\]

As \( \sqrt{1 + \ln K} \rightarrow 0 \) as \( K \rightarrow \infty \), \( \{ y_k(x) \} \in \mathcal{Y} \), and \( \mathcal{Y} \) is compact, LL objective convergence property holds. Next, it follows from Theorem 3 that \( \varphi_k(x) \rightarrow \varphi(x) \) as \( K \rightarrow \infty \) for any \( x \in \mathcal{X} \) and thus UL objective convergence property holds. \( \square \)

Further, in the following, we will show that when \( f(x, y) \) is level-bounded in \( y \) uniformly in \( x \in \mathcal{X} \), compactness assumption on \( \mathcal{Y} \) in Theorem 5 can be safely removed and \( \mathcal{Y} \) can be taken as \( \mathbb{R}^m \).

Theorem 6. Suppose Assumptions 1 is satisfied, \( f(x, y) \) is level-bounded in \( y \) uniformly in \( x \in \mathcal{X} \), and \( \mathcal{S}(x) \) is nonempty for all \( x \in \mathcal{X} \). Let \( \{ y_k(x) \} \) be the output generated by (9) with \( s_l = s_u = \mu \in (0, 1), \) \( \alpha_k = \frac{1}{k+1}, \beta_k \in [\beta, 1] \) with some \( \beta > 0, \beta_k \leq \beta_k \), \( \beta_k - \beta_k \leq \frac{c_0}{(k+1)^2} \) with some \( c_0 > 0 \), then we have that both the LL and UL objective convergence properties hold.

Proof. According to the update scheme of \( y_{k+1} \) given in Eq. (9), \( y_{k+1} \) can be equivalently regarded as

\[
y_{k+1} = \arg \min_{y \in \mathcal{Y}} \langle \nabla_y \phi_k(x, y_k), y - y_k \rangle + \frac{1}{2\beta} \|y - y_k\|^2,
\]

where \( \phi_k(x, y) = \alpha_k M(x, y) + \beta_k (1 - \mu) f(x, y) \). Since \( s_l \in (0, 1/\max(L_f, l_f)), \alpha_k, \beta_k, \mu \in (0, 1), \) we have \( s_l \leq 1/L \phi_k \), \( \phi_k \) denotes the Lipschitz continuity constant of \( \nabla_y \phi_k(x, \cdot) \). Then, \( [42, \text{Lemma 10.4}] \) yields that \( \phi_k(x, y_k) \leq \phi_k(x, y_k) \).

Since \( f(x, y) \) is assumed to be level-bounded in \( y \) uniformly in \( x \in \mathcal{X} \), there exists \( m_0 \) such that \( f(x, y) \) is bounded below by \( m_0 \) on \( \mathbb{R}^n \times \mathcal{Y} \). By Assumption 1, \( F \) is bounded below by \( M_0 \). And as \( \alpha_k \) and \( \beta_k \) are positive and nonincreasing, it follows from the above inequality that

\[
\alpha_k \mu (F(x, y_{k+1}) - M_0) + \beta_k (1 - \mu) f(x, y_{k+1}) - m_0 \leq \alpha_k \mu (F(x, y_k) - M_0) + \beta_k (1 - \mu) f(x, y_k) - m_0.
\]

Thus \( \beta_k \in [\beta, 1] \) implies that \( \forall k \) holds the following

\[
\mathcal{Y} \in \mathcal{C}, \forall k \in \mathcal{X}.
\]

Then by the continuity of \( \nabla_y F(x, y) \) and \( \nabla_y f(x, y) \), there exists a compact set \( \mathcal{C} \subset \mathbb{R}^m \) such that

\[
y_{k+1}(x) \in \mathcal{C}, \forall k \in \mathcal{X},
\]

where \( y_{k+1}(x) \) is defined in Eq. (9). This implies that the sequence \( \{ y_k \} \) coincides with the one generated by the update scheme in Eq. (9) with \( \mathcal{Y} = \mathcal{C} \). Then since \( \mathcal{C} \) is compact, the conclusion follows from Theorem 5 immediately. \( \square \)

Remark 2. Following the analysis recipe, the entire Section 5 is devoted to show that the constructed algorithm (i.e., BDA) meet two convergence properties. In particular, the UL and LL convergence verifications are presented in Theorem 3 and
Theorem 4, respectively. The proof of Theorem 3 mainly relies on the sufficiently decreasing inequality given in Lemma 1. Theorem 6 discussed the convergence behavior of BDA without the compactness assumption on $\mathcal{Y}$. 

6 Stationarity Analysis of BDA

This part provides the convergence behavior of the problem that $\min_x \varphi(x)$ is solved to (approximate) stationarity. We consider the special case where $\mathcal{Y} = \mathbb{R}^m$ and the LL objective function $f(x, \cdot) : \mathbb{R}^m \to \mathbb{R}$ is $\sigma$-strongly convex for any $x \in \mathcal{X}$. In this case, the solution set of LL problem $\mathcal{S}(x)$ is a singleton, and we denote its unique solution by $y^*(x)$. In the following, we are going to show the convergence of BDA with respect to stationary points in this special case. Our analysis is partly inspired by [31]. We first make the following assumptions.

Assumption 2. $F$ and $f$ are both twice continuously differentiable. For any $x \in \mathcal{X}$, $f(x, \cdot) : \mathbb{R}^m \to \mathbb{R}$ is $\sigma$-strongly convex.

Before providing the convergence results, we begin with a lemma.

Lemma 5. Let $\{a_k\}$ and $\{b_k\}$ be sequences of non-negative real numbers. Assume that $b_k \to 0$ and there exist $\rho \in (0, 1)$, $n_0 \in \mathbb{N}$ such that $a_k + 1 \leq b_k + b_k, \forall k \geq n_0$. Then $\lim_{k \to \infty} a_k = 0$.

Proof. As $b_k \to 0$, there exists $B > 0$ such that $b_k \leq B$ for all $k$. And we have for any $k \geq n_0$,

$$a_{k+1} \leq a_k + 1 \leq b_k + b_k \leq \rho a_k + B,$$

which implies the boundedness of the sequence $\{a_k\}$ and thus there exists $A$ such that $a_k \leq A$ for all $k$.

For any $\epsilon > 0$, since $b_k \to 0$, there exists $k_1 > n_0$ such that $b_k \leq \frac{1-\rho^2}{2}$ for all $k \geq k_1$. Since $\lim_{k \to \infty} b_k = 0$, there exists $k_2 > k_1$ such that for any $k \geq k_2$, $\rho^k \leq \frac{\epsilon}{2}$ and hence $a_{k+1} \leq \epsilon$. As $\epsilon$ is arbitrarily chosen, we obtain that $\lim_{k \to \infty} a_k = 0$. \hfill \Box

By applying implicit function theorem on the optimality condition of the LL problem, we obtain that $y^*(x)$ is differentiable on $\mathcal{X}$ and its derivative is given by

$$\frac{\partial y^*(x)}{\partial x} = - \left( \nabla_{yy} f(\mathbb{x}, y^*(\mathbb{x})) \right)^{-1} \nabla_{yx} f(\mathbb{x}, y^*(\mathbb{x})).$$

Hence, the function $\varphi(x) = F(\mathbb{x}, y^*(\mathbb{x}))$ is also differentiable and its derivative is given by Eq. (4). With the above Lemma 5, we have the following proposition.

Proposition 1. Suppose Assumptions 1 and 2 are satisfied, $f(x, y)$ is level-bounded in $y$ uniformly in $x \in \mathcal{X}$, $\mathcal{X}$ is compact, $\mathcal{Y} = \mathbb{R}^m$, and $\mathcal{S}(x)$ is nonempty for all $x \in \mathcal{X}$. Let $\{y_k(x)\}$ be the output generated by (9) with $s_1 = 1, s = s \in (0, 1)$, $\alpha_k = \alpha_k, \lim_{k \to \infty} \alpha_k = 0, \beta_k \leq \alpha_k, \lim_{k \to \infty} \beta_k = 0, \beta_k \leq \beta_k, |\beta_k - \beta_k| \leq \epsilon^\theta, \beta_k - \beta_k \geq \epsilon^\theta, \text{with some } \epsilon > 0$, then we have

$$\sup_{x \in \mathcal{X}} \left\| \nabla \varphi_k(x) - \nabla \varphi(x) \right\| \to 0, \text{ as } k \to \infty.$$
compact, $\nabla_x f$ and $\nabla_y f$ are both uniformly continuous on $X \times C$, then, the fact that $\sup_{x \in X} ||y_k(x) - y(x)|| \rightarrow 0$ as $k \rightarrow \infty$ implies that $\sup_{x \in X} ||\nabla_x f(x, y_k)|| \rightarrow 0$ and $\sup_{x \in X} ||\nabla_y f(x, y_k)|| \rightarrow 0$ as $k \rightarrow \infty$. $\sigma$-strong convexity of $f(x, \cdot)$ yields the continuity of $\frac{\partial y^*(x)}{\partial x}$ on $X$ and thus $\sup_{x \in X} \left| \frac{\partial y^*(x)}{\partial x} \right| < +\infty$. Then, we have $\sup_{x \in X} \left| \frac{\partial y^*(x)}{\partial x} \right| \rightarrow 0$ as $k \rightarrow \infty$. Next, as $\nabla_x f$ and $\nabla_y f$ are both continuous on $X \times \mathbb{R}^m$, $X, C$ are compact, and $y_k(x) \in C$ for any $k$ and $x \in X$, then

$$\sup_{x \in X} \left( ||\nabla_x f(x, y_k(x))|| + ||\nabla_y f(x, y_k(x))|| \right) < +\infty.$$ Combining with the fact that $\sup_{x \in X} \left| \frac{\partial y^*(x)}{\partial x} \right| < +\infty$, we have $\sup_{x \in X} \Gamma_k(x) < +\infty$. Because $\alpha_k \rightarrow 0$ as $k \rightarrow +\infty$, we have $\alpha_k \mu \sup_{x \in X, \Gamma_k(x) > \epsilon} \Gamma_k(x) \rightarrow 0$ as $k \rightarrow \infty$. According to Lemma 5, we have

$$\sup_{x \in X} \left| \frac{\partial y_k(x)}{\partial x} - \frac{\partial y^*(x)}{\partial x} \right| \rightarrow 0, \text{ as } k \rightarrow \infty. \quad (38)$$

Recalling Eq. (4) and

$$\nabla \varphi_k(x) = \nabla_x f(x, y_k(x)) + \left( \frac{\partial y_k(x)}{\partial x} \right)^\top \nabla_y f(x, y_k(x)),$$ we have the following estimate

$$\sup_{x \in X} ||\nabla \varphi_k(x) - \nabla \varphi(x)|| \leq \sup_{x \in X} \left| \frac{\partial y_k(x)}{\partial x} - \frac{\partial y^*(x)}{\partial x} \right| \sup_{x \in X} ||\nabla_y f(x, y_k(x))|| + \sup_{x \in X} ||\nabla_y f(x, y_k(x))||$$

where $\sup_{x \in X} ||\nabla_y f(x, y_k(x))|| = \sup_{x \in X} \left| \frac{\partial y_k(x)}{\partial x} - \frac{\partial y^*(x)}{\partial x} \right| \sup_{x \in X} ||\nabla_y f(x, y_k(x))|| + \sup_{x \in X} ||\nabla_y f(x, y_k(x))||$ and $\sup_{x \in X} ||\nabla_y f(x, y_k(x))|| = \sup_{x \in X} \left| \frac{\partial y_k(x)}{\partial x} - \frac{\partial y^*(x)}{\partial x} \right| \sup_{x \in X} ||\nabla_y f(x, y_k(x))|| + \sup_{x \in X} ||\nabla_y f(x, y_k(x))||$. Since $\nabla_x f$ and $\nabla_y f$ are continuous on $X \times \mathbb{R}^m$ and thus uniformly continuous on compact set $X \times C$, then the fact that $\sup_{x \in X} \left| \frac{\partial y_k(x)}{\partial x} - \frac{\partial y^*(x)}{\partial x} \right| \sup_{x \in X} ||\nabla_y f(x, y_k(x))|| < +\infty$ as $k \rightarrow \infty$ implies that $\sup_{x \in X} ||\nabla_y f(x, y_k(x))|| \rightarrow 0$ and $\sup_{x \in X} ||\nabla_y f(x, y_k(x))|| \rightarrow 0$ as $k \rightarrow \infty$. The continuity of $\frac{\partial \varphi(x)}{\partial x}$ on $X$ yields $\sup_{x \in X} \left| \frac{\partial \varphi(x)}{\partial x} \right| < +\infty$ and thus $\sup_{x \in X} ||\nabla_y f(x, y_k(x))|| \rightarrow 0$ as $k \rightarrow \infty$. Next, as $\nabla_y f$ is continuous on $X \times \mathbb{R}^m$ and $y_k(x)$ belongs to a compact set for any $k$ and $x \in X$, it holds that $\sup_{x \in X} ||\nabla_y f(x, y_k(x))|| < +\infty$ for any $k$. Then Eq. (38) implies that $\sup_{x \in X} \left| \frac{\partial \varphi(x)}{\partial x} \right| < +\infty$ as $k \rightarrow \infty$. Then we get the conclusion directly from Lemma 5.

**Theorem 7.** Suppose Assumptions 1 and 2 are satisfied, $f(x, y)$ is level-bounded in $y$ uniformly in $x \in X$, $X$ is compact, $\mathcal{Y}$ is nonempty for all $x \in X$. Let $\{y_k(x)\}$ be the output generated by (9) with $s_1 = s_2 = \epsilon_s (0.1/\max(L, f_1))$, $\mu \in (0, 1), \alpha_k \leq \alpha_{k-1}, \lim_{k \rightarrow \infty} \alpha_k = 0, \beta_k \in [\beta, 1]$ with some $\beta > 0$, $\beta_k \leq \beta_{k-1}, |\beta_k - \beta_{k-1}| \leq \frac{c_{\beta}}{k} \epsilon_{\beta}$ with some $c_{\beta} > 0$, and let $x_k$ be a $\epsilon_k$-stationary point of $\varphi(x)$, i.e., $\epsilon_k = \nabla \varphi_k(x_k)$.

Then if $\epsilon_k \rightarrow 0$, we have that any limit point $\bar{x}$ of the sequence $\{x_k\}$ is a stationary point of $\varphi$, i.e.,

$$0 = \nabla \varphi(\bar{x}).$$
Consider sequence $x^y_{\mathbf{K}}$, the existing bi-level based methods still may fail to reach an optimal solution. For example, with $x \in [-100, 100]$ and $y \in \mathbb{R}^2$, we consider the following BLO problem:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} ||x - z||^2 + ||y - e||^2,$$

subject to ($y, z \in \mathbb{R}^n$),

$$\min_{y \in \mathbb{R}^2} \frac{1}{2} ||y||^2 - x^T y,$$

(40)

where $X = [-100, 100] \times \cdots \times [-100, 100] \subset \mathbb{R}^n$, $e$ denotes the vector whose elements are all equal to 1. By simple calculation, we know that the unique optimal solution of Eq. (40) is $x^* = y^* = z^* = e$. However, if adopting the existing gradient-based scheme in Eq. (39) with initialization $y_0, z_0 = (0, 0)$ and varying step size $s_k^y \in (0, 1)$, we have that $y_k = (1 - \prod_{k=0}^{K-1} (1 - s_k^y)) x$ and $z_k = 0$. Then the approximated problem of Eq. (40) amounts to

$$\min_{x \in \mathbb{R}^n} F(x, y_k, z_k) = ||x||^2 + (1 - \prod_{k=0}^{K-1} (1 - s_k^y)) ||x - e||^2.$$

Consider sequence $x^*_{\mathbf{K}} = \arg \min_{x \in \mathbb{R}^n} F(x, y_k, z_k)$, it follows from the first-order optimality condition that

$$0 = 4 ||x^*_{\mathbf{K}}||^2 x^*_{\mathbf{K}} + 4a_K ||a_K x^*_{\mathbf{K}} - e||^2 (a_K x^*_{\mathbf{K}} - e),$$

(41)

where $a_K = (1 - \prod_{k=0}^{K-1} (1 - s_k^y))$. Then, if sequence $\{x^*_{\mathbf{K}}\}$ converge to a limit point $e$, and since $\{a_K\}$ is bounded, there exist subsequences $\{x^*_{\mathbf{K}_i}\} \subset \{x^*_{\mathbf{K}}\}$ and $\{a_{K_i}\} \subset \{a_K\}$ such that $\{x^*_{\mathbf{K}_i}\} \rightarrow e$ and $\{a_{K_i}\} \rightarrow a$. By considering subsequences $\{x^*_{\mathbf{K}_i}\}$ and $\{a_{K_i}\}$ in Eq. (41) and taking $K_i \rightarrow \infty$, we should have

$$0 = ||e||^2 e + a \bar{a} \bar{e} - e||^2 (\bar{a}e - e)$$

and thus $0 = 1 + (\bar{a} - 1)^3 \bar{a}$. However, since $a_K = (1 - \prod_{k=0}^{K-1} (1 - s_k^y)) \in [0, 1]$, then $\bar{a} \in [0, 1]$ and

$$1 + (\bar{a} - 1)^3 \bar{a} \geq 1 - ||(\bar{a} - 1)\bar{a}|| \geq \frac{3}{4} > 0,$$

which is a contradiction to $0 = 1 + (\bar{a} - 1)^3 \bar{a}$. Therefore, any subsequence of $\{x^*_{\mathbf{K}}\}$ cannot converge to the true solution (i.e., $x^* = e$).

Remark 3. Actually, even with strongly convex UL objective w.r.t. LL variable $y$, the existing bi-level based methods still may fail to converge to the true solution.
Algorithmically, this work establishes a more general framework, in which we introduce the projection-based operations to handle set constrains in BLOs and design more flexible strategy to set the aggregation parameters during iterations. We also established a one-stage fast approximation to BDA for solving large-scale real-world problems.

We design a high-dimensional counter-example (see Eq. (40)) and conduct various experiments to verify our new theoretical findings, i.e., the efficiency of BDA for BLOs without the LLS condition and both the UL and LL objectives are convex but not strongly convex. We further do new experiments to more clearly analyze the components of BDA and report more results on real applications (e.g., with new evaluation metrics, on more challenging benchmarks and compared with more state-of-the-art approaches).

### 7.3 One-stage BDA: A Fast Implementation

Multi-step of the LL iteration modules $T_k$ will cause a lot of memory consumption that may be an obstacle in modern massive-scale deep learning applications. Thus it would be useful to simplify iteration steps. This part provides an extension scheme leveraging a one-stage simplification to reduce complicated gradient-based calculation steps [17]. By setting $K = 1$ in Eq. (9), the algorithm reads as

$$y_1(x) = T_1(x, y_0) = \text{Proj}_Y (y_0 - s\partial_y \phi(x, y_0)),$$

where \(\phi(x, y_0) = \alpha F(x, y_0) + \beta f(x, y_0)\) and \(\alpha, \beta \in (0, 1]\) denote the aggregation parameters. Indeed, if \((y_0 - s\partial_y \phi(x, y_0)) \notin \mathcal{Y}\), with this one-stage simplification, we can simplify the back-propagation calculation with the following finite difference approximation

$$\frac{d\phi_1(x)}{dx} \approx \frac{\partial F(x, y_1) - \partial F(x, y_0)}{s},$$

where \(h^\pm_0 = y_0 \pm \epsilon \partial_f(x, y_1) / \partial y_1\) and \(\partial_f(x, y) = \alpha \partial_x F(x, y) + \beta \partial_x f(x, y)\). Since \(y_1\) can be a big interval, this case (i.e., \((y_0 - s\partial_y \phi(x, y_0)) \notin \mathcal{Y}\)) is often satisfied in general. If \((y_0 - s\partial_y \phi(x, y_0)) \notin \mathcal{Y}\), the above back-propagation can be calculated by the following form

$$\frac{d\phi_1(x)}{dx} \approx \frac{\partial F(x, y_1) - \partial F(x, y_0)}{s} \pm 2 \epsilon \partial_x \phi(x, h^\pm_0),$$

where \(h^\pm_0 = y_0 \pm \epsilon \text{Proj}_Y (z_0 + \epsilon/2 \partial F(x, y_1) / \partial y_1)\) and \(h^{\pm}_0 = y_0 \pm \epsilon \text{Proj}_Y (z_0 - \epsilon/2 \partial F(x, y_1) / \partial y_1)\) with \(z_0 = y_0 - s\partial_y \phi(x, y_0)\).

### 8 Experimental Results

This section first verify the numerical results and then evaluate the performance of our proposed method on different problems.

#### 8.1 Numerical Evaluations

Our numerical results are investigated based on the synthetic BLO described in Section 7.1, i.e., Counter Example in Eq. (40). As stated in Section 7.1, this deterministic bi-level formulation satisfies all the assumptions required in Section 4,
We considered different numerical metrics, such as the convergence speed of our iterative sequences. To show the influence of the LL iterations (i.e., \( t \geq 5 \)), BDA is close to the optimal solution \( y^* \) while RHG cannot. In the above figures, we set \( \alpha_k = 0.5/k \), \( k = 1, \cdots, K \), \( s_u = s_l = 0.1 \), and \( \beta_k = 1, \mu = 0.1 \).

Figure 4 plotted numerical results of the proposed BDA and RHG [14, 12] with ten different initialization points. We considered different numerical metrics, such as the relationship of \((x, y)\) with optimal solution \((x^*, y^*)\) and the distance between \(F(x, y_K)\) and \(F^*\) (i.e., \( |F(x, y_K) - F^*|\)), for evaluations. It needs to be noted that we select five different initial points to show the performance of \(|F(x, y_K) - F^*|\). As can be observed that RHG is always hard to obtain the correct solution, even start from different initialization points. It is mainly because that the solution set of the LL subproblem in Eq. (40) is not a singleton, which does not satisfy the fundamental assumption of RHG. In contrast, the proposed method can obtain a truly optimal solution in all these scenarios. The initialization only slightly affects the convergence speed of our iterative sequences.

To explore the performance under projection operator denoted in Eq. (9), we report in Figure 5 the results (i.e., \(|x - x^*|\) and \(|y - y^*|\)) of comparing the performance with and without projection (i.e., \( w / \) Projection, \( w/o \) Projection). In this experiment, we set the initial value far away from the optimal point with relatively close projection interval \( \gamma \). As can be seen, with the projection operator, the iteration sequences reach convergence with fewer steps.

Figure 6 evaluated the convergence behaviors of BDA with different choices of \( \alpha_k \). We set \( \beta_k = 1, \mu = 0.5, s_u = 0.1 \) and \( s_l = 0.1 \). By setting \( \alpha_k = 0 \), we were unable to use the UL information guiding the LL updating. Thus it is hard to obtain proper feasible solutions for the UL approximation subproblem. When choosing a fixed \( \alpha_k \) 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.

### 8.2 Hyper-parameter Optimization

For the hyper-parameter optimization problem, the key idea is to choose a set of optimal hyper-parameters for a given machine learning task. In this experiment, we consider a specific hyper-parameter optimization example (i.e., data hyper-cleaning [14, 27]) to evaluate the developed bi-level algorithm. This task aims to train a linear classifier on a machine learning task. In this experiment, we consider a specific hyper-parameter optimization example (i.e., data hyper-cleaning [14, 27]) to evaluate the developed bi-level algorithm. This task aims to train a linear classifier on a given image set, but part of the training labels are corrupted. Here we consider soft-max regression (with parameters \( y \)) as our classifier and introduce hyper-parameters \( x \) to weight examples for training. We define the LL objective as the following weighted training loss:

\[
f(x, y) = \sum_{(u_i, v_i) \in D_{tr}} [\sigma(x)] \ell(y; u_i, v_i),
\]

where \( x \) is the hyper-parameter vector to penalize the objective for different training samples, \( \ell(y; u_i, v_i) \) means the cross-entropy function with the classification parameter \( y \), and data pairs \((u_i, v_i)\) and denote \( D_{tr} \) and \( D_{val} \) as the training and validation sets, respectively. Here \( \sigma(x) \) denotes the element-wise sigmoid function on \( x \) 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 \( \ell_2 \) regularization on the validation set, i.e.,

\[
F(x, y) = \sum_{(u_i, v_i) \in D_{val}} \ell(y(x); u_i, v_i).
\]
In particular, the UL and LL objective $F$ and $f$ w.r.t. $y$ is required to be convex. To satisfy this requirement, we design the classifier with a fully connected layer.

We applied our BDA and One-stage BDA (O-BDA) together with the bi-level based methods, i.e., Implicit HG (IHG) [35], RHG and Truncated RHG (T-RHG) [27]. We first conduct the experiment on two datasets (MNIST dataset [44] and Fashion MNIST dataset [45]) that each with 5000 training examples (i.e., $D_{tr}$), 5000 validation examples (i.e., $D_{val}$) and a test set with the remaining 60000 samples. We randomly chose 2500 training samples from $D_{tr}$ and pollute the labels.

We use validation accuracy (i.e., Val. Acc.), test accuracy (i.e., Test Acc.), F1-score and running times as the metrics of our developed algorithm. As shown in Table 2, the developed method perform the best both on MNIST and Fashion MNIST dataset. The LL iterations are $K = 200$ and $K = 50$ on MNIST and Fashion MNIST, respectively. For T-RHG, we chose 100-step and 25-step truncated back-propagation respectively from $K = 200$ and $K = 50$ to guarantee its convergence. Besides, the developed O-BDA still perform better when comparing with the existing bi-level based methods.

### 8.3 Meta-Learning

Meta-learning aims to leverage a large number of similar few-shot tasks to learn an algorithm that should work well on novel tasks in which only a few labeled samples are available. In particular, we consider the few-shot learning problem [46], [47], where each task is to discriminate $N$ separate classes and it is to learn the hyper-parameter $x$ 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 that the network architecture is with four-layer CNNs followed by fully connected layer, we separate the network architecture into two parts: the cross-task intermediate representation layers (parameterized by $x$) outputs the meta features and the multinomial logistic regression layer (parameterized by $y^j$) as our ground classifier for the $j$-th task. We also collect a meta training data set $D = \{D_j\}$, where $D_j = D_{tr}^j \cup D_{val}^j$ is linked to the $j$-th task. Then for the $j$-th task, we consider the cross-entropy function $\ell(x, y^j; D_{tr}^j)$ as the task-specific loss and thus the LL objective can be defined as

$$f(x, \{y^j\}) = \sum_j \ell(x, y^j; D_{tr}^j).$$

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

$$F(x, \{y^j\}) = \sum_j \ell(x, y^j; D_{val}^j).$$

Our experiments are conducted on Omniglot [48] and MiniImageNet [46] benchmarks. We compared our BDA to several state-of-the-art approaches, such as MAML [28], Meta-SGD [49], Reptile [29], iMAML [2], RHG, and T-RHG. As shown in Table 3, BDA compared well to these methods and achieved the highest classification accuracy except in the 5-way task. Further, with more complex problems (such as 20-way, 30-way and 40-way), BDA shows significant advantages.
1-shot are shown in Figure 7.

This work established a flexible descent aggregation framework to handle the LLS issue. Then this work strictly proved the convergence of the developed framework without the LLS assumption and the strong convexity in the UL objective. Focusing on different solution qualities (namely, global, local, and stationarity), this work elaborated the convergence results respectively. Finally, extensive experiments justified our theoretical results and demonstrated the superiority of the proposed algorithm for hyper-parameter optimization and meta-learning.

9 Conclusions

This work established a flexible descent aggregation framework with task-tailored iteration dynamics modules to solve bi-level tasks by formulating BLO in Eq. (2) from the viewpoint of optimistic bi-level. We provided a new algorithmic framework to handle the LLS issue. Then, this work strictly proved the convergence of the developed framework without the LLS assumption and the strong convexity in the UL objective. Focusing on different solution qualities (namely, global, local, and stationarity), this work elaborated the convergence results respectively. Finally, extensive experiments justified our theoretical results and demonstrated the superiority of the proposed algorithm for hyper-parameter optimization and meta-learning.

ACKNOWLEDGEMENTS

This work is partially supported by the National Key R&D Program of China (2020YFB1313503), the National Natural Science Foundation of China (Nos. 61922019 and 11971220), the Fundamental Research Funds for the Central Universities, the Shenzhen Science and Technology Program (No. RCX2020071414700072), the Stable Support Plan Program of Shenzhen Natural Science Fund (No. 20200925152128002), the Guangdong Basic and Applied Basic Research Foundation and the Pacific Institute for the Mathematical Sciences (PIMS).

REFERENCES

[1] R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang, “A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton,” in ICML, 2020, pp. 6305–6315.
[2] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine, “Meta-learning with implicit gradients,” in NeurIPS, 2019, pp. 113–124.
[3] M. MacKay, P. Vicil, J. Lorrain, D. Duvenaud, and R. Grosse, “Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions,” ICLR, 2019.
[4] K. Kunisch and T. Pock, “A bilevel optimization approach for parameter learning in variational models,” SIAM Journal on Imaging Sciences, vol. 6, no. 2, pp. 938–983, 2013.
[5] R. Liu, J. Gao, J. Zhang, D. Meng, and Z. Lin, “Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond,” CoRR, abs/2101.11517, 2021.
[6] S. Dempe, Bilevel optimization: theory, algorithms and applications. TU Bergakademie Freiberg Mining Academy and Technical University, 2018.
[7] S. Dempe, B. S. Mordukhovich, and A. B. Zemkoho, “Two-level value function approach to non-smooth optimistic and pessimistic bilevel programs,” Optimization, vol. 68, no. 2-3, pp. 433–455, 2019.
[8] S. Dempe, J. Dutta, and B. Mordukhovich, “New necessary optimality conditions in optimistic bilevel programming,” Optimization, vol. 56, no. 5-6, pp. 577–604, 2007.
[9] B. Kohli, “Optimality conditions for optimistic bilevel programming problem using convexifiers,” Journal of Optimization Theory and Applications, vol. 152, no. 3, pp. 632–651, 2012.
[10] L. Lampariello, S. Sagratella et al., “Numerically tractable optimistic bilevel problems.” Computational Optimization and Applications, vol. 76, no. 2, pp. 277–303, 2020.
[11] R. Liu, Y. Liu, S. Zeng, and J. Zhang, “Towards gradient-based bilevel optimization with non-convex followers and beyond,” CoRR, abs/2110.00455, 2021.
[12] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, “Bilevel programming for hyperparameter optimization and meta-learning,” in ICLR, 2018, pp. 1563–1572.
[13] D. Zügner and S. Günnemann, “Adversarial attacks on graph neural networks via meta learning,” ICLR, 2019.
[14] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil, “Forward and reverse gradient-based hyperparameter optimization,” in ICML, 2017, pp. 1165–1173.
[15] T. Okuno, A. Takeda, and A. Kawana, “Hyperparameter learning via bilevel nonsmooth optimization,” CoRR, abs/1806.01520, 2018.
[16] Z. Yang, Y. Chen, M. Hong, and Z. Wang, “Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost,” in NeurIPS, 2019, pp. 8351–8363.
[17] H. Liu, K. Simonyan, and Y. Yang, “Darts: Differentiable architecture search,” in ICLR, 2019.
[18] B. Wu, X. Dai, P. Zhang, Y. Wang, F. Sun, Y. Wu, Y. Tian, P. Vajda, Y. Jia, and K. Keutzer, “Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search,” in IEEE CVPR, 2019, pp. 10734–10742.
[19] N. Nakai, T. Matsubara, and K. Uehara, “Att-darts: Differentiable neural architecture search for attention,” in IEEE IJCNN, 2020, pp. 1–8.
[20] Y. Hu, X. Wu, and R. He, “Tf-nas: Rethinking three search freedoms of latency-constrained differentiable neural architecture search,” in ECCV, 2020, pp. 123–139.
[21] J. C. de los Reyes, C.-B. Schönlieb, and T. Valkonen, “Bilevel parameter learning for higher-order total variation regularization models,” *Journal of Mathematical Imaging and Vision*, vol. 57, no. 1, pp. 1–25, 2017.

[22] J. Chen, P. Mu, R. Liu, X. Fan, and Z. Luo, “Flexible bilevel image layer modeling for robust denoising,” in *IEEE ICME*, 2020, pp. 1–6.

[23] R. Liu, P. Mu, J. Chen, X. Fan, and Z. Luo, “Investigating task-driven latent feasibility for nonconvex image modeling,” *IEEE Transactions on Image Processing*, vol. 29, pp. 7629–7640, 2020.

[24] R. G. Jeroslow, “The polynomial hierarchy and a simple model for competitive analysis,” *Mathematical Programming*, vol. 32, no. 2, pp. 146–164, 1985.

[25] E. Weinan, “A proposal on machine learning via dynamical systems,” *Communications in Mathematics and Statistics*, vol. 5, no. 1, pp. 1–11, 2017.

[26] D. Maclaurin, D. Duvenaud, and R. Adams, “Gradient-based hyperparameter optimization through reversible learning,” in *ICML*, 2015, pp. 2113–2122.

[27] A. Shaban, C.-A. Cheng, N. Hatch, and B. Boots, “Truncated backpropagation for bilevel optimization,” in *AISTATS*, 2019, pp. 1723–1732.

[28] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *ICML*, 2017, pp. 1126–1135.

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

[30] J. Lorraine and D. Duvenaud, “Stochastic hyperparameter optimization through hypernetworks,” *CoRR*, abs/1802.09419, 2018.

[31] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo, “On the iteration complexity of hypergradient computation,” in *ICML*, 2020, pp. 3748–3758.

[32] J. Lorraine, P. Vicol, and D. Duvenaud, “Optimizing millions of hyperparameters by implicit differentiation,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1540–1552.

[33] Q. Bertrand, Q. Klopfenstein, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon, “Implicit differentiation of lasso-type models for hyperparameter optimization,” in *ICML*, 2020, pp. 810–821.

[34] H. Xiao, K. Rasul, and S. Zaremba, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” in *CoRR*, abs/1708.07747, 2017.

[35] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra et al., “Matching networks for one shot learning,” in *NeurIPS*, 2016, pp. 3630–3638.

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

[37] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, “Human-level concept learning through probabilistic program induction,” *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[38] Z. Li, F. Zhou, F. Chen, and H. Li, “Meta-sgd: Learning to learn quickly for few shot learning,” *CoRR*, abs/1707.08835, 2017.

[39] A. Prakash, J. Roster, D. Florencio, and C. Zhang, “Repr: Improved training of convolutional filters,” in *IEEE CVPR*, 2019, pp. 10666–10675.