A Generic 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, gradient-based methods for solving bi-level optimization tasks have drawn a great deal of interest from the machine learning community. However, to calculate the gradient of the best response, existing research always relies on the singleton of the lower-level solution set (a.k.a., Lower-Level Singleton, LLS). In this work, by formulating bi-level models from an optimistic bi-level viewpoint, we first establish a novel Bi-level Descent Aggregation (BDA) framework, which aggregates hierarchical objectives of both upper level and lower level. The flexibility of our framework benefits from the embedded replaceable task-tailored iteration dynamics modules, thereby capturing a wide range of bi-level learning tasks. Theoretically, we derive a new methodology to prove the convergence of BDA framework without the LLS restriction. Besides, the new proof recipe we propose is also engaged to improve the convergence results of conventional gradient-based bi-level methods under the LLS simplification. Furthermore, we employ a one-stage technique to accelerate the back-propagation calculation in a numerical manner. Extensive experiments justify our theoretical results and demonstrate the superiority of the proposed algorithm for hyper-parameter optimization and meta-learning tasks.

Index Terms—Bi-level programming, 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 applications. Mathematically, BLO can be formulated as the following optimization problem:

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

where the Upper-Level (UL) objective F is a jointly continuous function, the UL constraint $\mathcal{X}$ is a compact set, the set-valued mapping $\mathcal{S}(\mathbf{x})$ indicates the solution set of the Lower-Level (LL) subproblem parameterized by $\mathbf{x}$, and $\mathcal{Y}\subseteq \mathrm{dom}F$ is a compact convex set. Without loss of generality, the (parameterized) LL subproblem can be stated as:

$$\min_{\mathbf{y}} f(\mathbf{x},\mathbf{y}), \tag{2}$$

where the Lower-Level (LL) objective f is jointly continuous. Indeed, the BLO model 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.,

$$\mathbf{y}\in\mathcal{S}(\mathbf{x})=\arg\min_{\mathbf{y}} f(\mathbf{x},\mathbf{y}), \tag{3}$$

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 $\mathcal{S}(\mathbf{x})$ in Eq. (3) is not a singleton [1], [2]. In this work, we always call the condition that $\mathcal{S}(\mathbf{x})$ is a singleton as Lower-Level Singleton (or LLS for short).

Although early works on BLO can date back to the nineteen seventies [2], it was not until the last decade that a large amount of bi-level optimization models were established to capture machine learning applications, including meta learning [3], [4], [5], hyper-parameter optimization [6], [7], [8], reinforcement learning [9], generative adversarial learning [10], [11], neural architecture search [12], [13], [14], [15] and image processing [16], [17], [18], [19], [20], and etc.

Early studies focused on numerical methods to solve BLO in Eqs. (1)-(2). Classical solution schemes for BLO which had gained popularity from the optimization community can only manage BLO models with simple structures. For example, by using first-order optimality conditions to replace the LL subproblem, the original BLO in Eqs. (1)-(2) is reformulated into a single-level optimization problem with equilibrium constraints; see, e.g., [21], [22], [7]. However, these approaches involve a large number of auxiliary variables and constraints, thus cannot address complex machine learning tasks with high-dimensional data set.

Recently, for solving BLO, variety of gradient-based methods associated with neural networks have attracted great attentions. Basically, the key idea behind these gradient-based methods is to optimize BLO with approximate best response Jacobin (i.e., the gradient of the best response

- *R. Liu is with the DUT-RU International School of Information Science & Engineering, Dalian University of Technology, and also with the Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian 116024, China. E-mail: rsliu@dlut.edu.cn.*
- *P. Mu is with the School of Mathematical Sciences, Dalian University of Technology, and also with the Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian 116024, China. E-mail: muyifan11@mail.dlut.edu.cn.*
- *X. Yuan and S. Zeng are with the Department of Mathematics, The University of Hong Kong, Hong Kang, China. E-mail: xmyuan@hku.hk, zengsz@connect.hku.hk.*
- *J. Zhang is with the Department of Mathematics, SUSTech International Center for Mathematics, Southern University of Science and Technology, National Center for Applied Mathematics Shenzhen, Shenzhen, Guangdong, China. (Corresponding author, E-mail: zhangj9@sustech.edu.cn.)*

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

function about the UL variable $\mathbf{x}$). According to different features of approximations, these gradient-based methods for BLO can be classified into two main categories: implicit best response and explicit best response.

The first category, i.e., implicit best response, which people also refer to as implicit differentiation ([4], [23], [24], [25] and [26]), relies on the observation that it is possible to replace the LL subproblem by an implicit equation. Specifically, the gradient of implicit best response takes advantage of the Implicit Function Theorem and computes the Jacobian by implicit differentiation equations, which depends only on the solution to the LL optimization and effectively decouples the UL gradient computation from the choice of LL optimizer. These implicit equation methods derive exact best response gradients but involve computing a Hessian matrix and its inverse, which could be computationally expensive for large-scale problems. To address this issue, the Hessian inverse is always approximated by calculating Neumann series [24] or solving a linear system with Conjugate Gradient (CG) method [26].

For the other category, i.e., explicit best response or automatic differentiation, the best response gradients are obtained by automatic differentiation through iterations of the LL gradient descent. This explicit structure mainly includes three schemes: recurrence-based [27], [6], [3], [28], initialization-based [29], [30] and proxy-based scheme [31], [8]. Specifically, recurrence-based best response 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 [29], [30], 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 [31], [8], a so-called hyper-network is trained to map LL gradients for their hierarchical optimization.

Practical experiments have demonstrated that the aforementioned gradient-based bi-level methods are very powerful for solving machine learning applications. The experimental performance and numerical efficiency have been witnessed in diversified learning tasks, but research on the theoretical convergence is still in its infancy (as summarized in Table 1). Indeed, all the mentioned gradient-based methods require the LLS condition in Eq. (2) to simplify their optimization processes and gain theoretical guarantees. For example, [3], [28] enforce the strong convexity assumption to the LL subproblem, which is even stronger than the LLS and very restrictive for real-world complex tasks.

In response to this limitation, in this work we propose a novel framewrok termed Bi-level Descent Aggregation (BDA); see an early version in [32]. Note that, for existing methods within the explicit best response category, the explicit best response 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., [3, Section 3]. The BDA scheme, which roughly falls into the explicit best response category, as the approximation is also constructed in terms of optimization dynamics, differs substantially from other gradient-based methods. In particular, the BDA is a new framework from the optimistic bi-level viewpoint; see Eq. (4). It establishes suitable optimization dynamics which suffice to ensure the desired good approximations. Indeed, in order to achieve such a good approximation, instead of replacing the LL subproblem with dynamics, the BDA actually replaces the inner simple bi-level subproblem with dynamics; see Eq. (5) for description of the inner simple bi-level. While the inner simple bi-level subproblem contains both the LL and the UL objectives, the BDA optimization iteration dynamics characterize an aggregation of both the LL and the UL descent informations. Consequently, the inner simple bi-level dynamics converge to some minimizers of both the LL and UL 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 *LL objective convergence property* and *UL objective convergence property*; see Section 2 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 *LL objective convergence property* and *UL objective convergence property* without imposing any strong convexity assumptions in either UL or LL subproblems. 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 formulating BLO in Eqs. (1)-(2) from the viewpoint of optimistic bi-level, this work provides a new generic bi-level algorithmic framework. Embedded with a task-tailored iterative gradient-aggregation dynamics for solving the inner simple bi-level, our framework owns the ability to flexibly capture different types of learning tasks.
- We establish a general convergence analysis template and an associated proof recipe for our proposed algorithmic framework. In particular, the convergence of our developed algorithm can be strictly justified without the singleton assumption on the LL subproblem, which is always required by existing approaches. Please see Table 1 for more details.
- Our analysis further demonstrates that the iterative gradient-aggregation dynamics (i.e., Eq. (8)) can also be interpreted as a new scheme for solving the simple bi-level problem without the UL strong convexity.

TABLE 1
Comparing the convergence results between our method and existing gradient-based bi-level methods in different scenarios (i.e., BLO with and without LLS condition).

Alg.			LLS	w/o LLS	
				with UL Strongly Convex	w/o Strongly Convex
Existing Gradient-based Bi-level Methods		UL	$F(\mathbf{x},\cdot)$ is Lipschitz continuous.	Not available	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: $\mathbf{x}_K \xrightarrow{s} \mathbf{x}^*$, $\inf_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) \to \inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x})$.		
Ours	[32]	UL	$F(\mathbf{x},\cdot)$ is Lipschitz continuous.	$F(\mathbf{x},\cdot)$ is L_F-smooth, and σ-strongly convex.	Not available
		LL	$\{\mathbf{y}_K(\mathbf{x})\}$ is uniformly bounded on $\mathcal{X}$, $f(\mathbf{x},\mathbf{y}_K(\mathbf{x})) \xrightarrow{u} f^*(\mathbf{x})$.	$f(\mathbf{x},\cdot)$ is L_f-smooth and convex, $\mathcal{S}(\mathbf{x})$ is continuous.	
			$f(\mathbf{x},\mathbf{y})$ is level-bounded in $\mathbf{y}$ locally uniformly in $\mathbf{x}\in\mathcal{X}$.		
		Main results: $\mathbf{x}_K \xrightarrow{s} \mathbf{x}^*$, $\inf_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) \to \inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x})$.		
	This Work	UL	$F(\mathbf{x},\cdot)$ is Lipschitz continuous.	$F(\mathbf{x},\cdot)$ is L_F-smooth, convex and bounded below.	
		LL	$\{\mathbf{y}_K(\mathbf{x})\}$ is uniformly bounded on $\mathcal{X}$, $f(\mathbf{x},\mathbf{y}_K(\mathbf{x})) \xrightarrow{u} f^*(\mathbf{x})$.	$f(\mathbf{x},\cdot)$ is L_f-smooth and convex.	
		Main results: $\mathbf{x}_K \xrightarrow{s} \mathbf{x}^*$, $\inf_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) \to \inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x})$.		

Here $\xrightarrow{s}$ and $\xrightarrow{u}$ represent the subsequential and uniform convergence, respectively. The superscript * denotes that it is the true optimal variables/values.

Consequently, the assumptions required by our framework have been significantly weakened.

- Last in order of occurrence but just as important, we design a new high-dimensional counter-example (i.e., Example 1), which explicitly illustrates the limitations of existing gradient-based bi-level approaches and verifies our above theoretical investigations. A variety of numerical experiments have also been conducted to demonstrate the superiority of the proposed algorithmic framework.

2 The Proposed Algorithm

Form an optimistic bi-level perspective, this section first initializes a generic framework for designing an explicit best response type algorithm for solving BLO (i.e., Eqs. (1)-(2)). Besdies, this section also introduces a developed template for analyzing the convergence algorithms within the framework.

2.1 Our Algorithmic Framework

From an optimistic bi-level viewpoint[1], we can reformulate Eqs. (1)-(2) as

$$\min_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x}), \text{ with } \varphi(\mathbf{x}) = \inf_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{S}(\mathbf{x})} F(\mathbf{x},\mathbf{y}). \tag{4}$$

Such reformulation reduces BLO 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.,

$$\min_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x},\mathbf{y}), \ s.t. \ \mathbf{y}\in\mathcal{S}(\mathbf{x}), \text{ (with fixed } \mathbf{x}). \tag{5}$$

1. For more theoretical details of optimistic BLO, we refer to [2] and the references therein.

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

$$\mathbf{y}_{k+1}(\mathbf{x}) = \mathcal{T}_{k+1}(\mathbf{x},\mathbf{y}_k(\mathbf{x})), \ k=0,\cdots,K-1, \tag{6}$$

where $\mathbf{y}_0(\mathbf{x}) = \mathbf{y}_0$ with some vector $\mathbf{y}_0 \in \mathcal{Y}$ and $\mathcal{T}_k(\mathbf{x},\cdot)$ stands for a schematic iterative module originated from a certain simple bi-level solution strategy on Eq. (5) with a fixed UL variable $\mathbf{x}$. Then we can replace $\varphi(\mathbf{x})$ by $F(\mathbf{x},\mathbf{y}_K(\mathbf{x}))$ and approximate Eq. (4) as:

$$\min_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) = F(\mathbf{x},\mathbf{y}_K(\mathbf{x})). \tag{7}$$

With the above procedure, the BLO in Eqs. (1)-(2) is approximated by a sequence of standard single-level optimization problems.

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)-(7), with a schematic $\mathcal{T}_k$). Basically, our proof methodology consists of two main steps:

(1) **LL objective convergence property:** $\{\mathbf{y}_K(\mathbf{x})\}$ is uniformly bounded on $\mathcal{X}$, and for any $\epsilon>0$, there exists $k(\epsilon)>0$ such that whenever $K>k(\epsilon)$,

$$\sup_{\mathbf{x}\in\mathcal{X}} \{f(\mathbf{x},\mathbf{y}_K(\mathbf{x})) - f^*(\mathbf{x})\} \le \epsilon.$$

(2) **UL objective convergence property:** For each $\mathbf{x}\in\mathcal{X}$,

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

With the above discussions, solving the BLO reduces to solve a simple bi-level problem in Eq. (5) w.r.t. the LL variable $\mathbf{y}$, and subsequently solve a single-level problem in Eq. (7) w.r.t. the UL variable $\mathbf{x}$.

2.2 Improved Iteration Modules

To capture more general bi-level applications without UL strong convexity, we shall construct an implementable solution strategy for solving the inner simple bi-level subproblem (i.e., Eq. (5)). Thus, in this part, we propose a gradient type method for solving simple bi-level problems (i.e., Eq. (5)) with merely convex UL and LL objectives. In particular, the descent informations of both the UL and LL objectives are aggregated to design $\mathcal{T}_k$. For a given $\mathbf{x}$, the descent directions of the UL and LL objectives can be defined separately as

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

where s_u, s_l are their step size parameters. The new simple bi-level algorithm then reads as

$$\begin{aligned}&\hat{\mathbf{y}}_{k+1}(\mathbf{x}) = \mathbf{y}_k(\mathbf{x}) - \left(\mu\alpha_k \mathbf{d}_k^F(\mathbf{x}) + (1-\mu)\beta_k \mathbf{d}_k^f(\mathbf{x})\right)\\ &\mathcal{T}_{k+1}\left(\mathbf{x}, \mathbf{y}_k(\mathbf{x})\right) = \mathbf{y}_{k+1}(\mathbf{x}) = \mathtt{Proj}_{\mathcal{Y}}\left(\hat{\mathbf{y}}_{k+1}(\mathbf{x})\right),\end{aligned} \tag{8}$$

where $k = 0, \ldots K-1$, $\mu \in (0,1)$ and $\alpha_k, \beta_k \in (0,1]$ denote the aggregation parameters. Proj means the projection operator. Actually, if we set $\beta_k = (1-\mu\alpha_k)/(1-\mu)$, the above iterative module $\tilde{\mathbf{y}}_{k+1}$ reduces to the iteration scheme in [32, Eq. (10)].

3 Theoretical Investigations

In this section, we first derive a general convergence proof recipe in Section 3.1 according to the developed analysis template (stated in Section 2.1). Following this roadmap, convergence behaviors of gradient-based bi-level methods can be systematically investigated (Section 3.2).

3.1 General Convergence Recipe

We brief some definitions, which are necessary for our analysis. One may also refer to [33] for more details on these variational analysis properties.

Definition 1. *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 defined as*

$$\begin{aligned}\limsup_{\mathbf{x}\to\bar{\mathbf{x}}} \varphi(\mathbf{x}) &= \lim_{\delta\to 0}\left[\sup\nolimits_{\mathbf{x}\in\mathbb{B}_\delta(\bar{\mathbf{x}})} \varphi(\mathbf{x})\right]\\ &= \inf_{\delta>0}\left[\sup\nolimits_{\mathbf{x}\in\mathbb{B}_\delta(\bar{\mathbf{x}})} \varphi(\mathbf{x})\right],\\ \liminf_{\mathbf{x}\to\bar{\mathbf{x}}} \varphi(\mathbf{x}) &= \lim_{\delta\to 0}\left[\inf\nolimits_{\mathbf{x}\in\mathbb{B}_\delta(\bar{\mathbf{x}})} \varphi(\mathbf{x})\right]\\ &= \sup_{\delta>0}\left[\inf\nolimits_{\mathbf{x}\in\mathbb{B}_\delta(\bar{\mathbf{x}})} \varphi(\mathbf{x})\right],\end{aligned}$$

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

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

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

Thanks to the continuity of $f(\mathbf{x}, \mathbf{y})$, we have the following semi-continuity over partial minimization.

Lemma 1. *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 $\mathcal{X}$.*

Proof. For any sequence $\{\mathbf{x}_t\} \subseteq \mathcal{X}$ satisfying $\mathbf{x}_t \to \bar{\mathbf{x}} \in \mathcal{X}$, and given any $\epsilon > 0$, let $\bar{\mathbf{y}} \in \mathbb{R}^m$ satisfy $f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \le f^*(\bar{\mathbf{x}}) + \epsilon$. As f is continuous at $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, there exists $T > 0$ such that

$$f^*(\mathbf{x}_t) \le f(\mathbf{x}_t, \bar{\mathbf{y}}) \le f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) + \epsilon \le f^*(\bar{\mathbf{x}}) + 2\epsilon, \quad \forall t > T,$$

and thus

$$\limsup_{t\to\infty} f^*(\mathbf{x}_t) \le f^*(\bar{\mathbf{x}}) + 2\epsilon.$$

By taking $\epsilon \to 0$, we get $\limsup_{k\to\infty} f^*(\mathbf{x}_t) \le f^*(\bar{\mathbf{x}})$. □

Equipped with the above two properties (i.e., *LL objective convergence property* and *UL objective convergence property*), we can establish our general convergence results in the following theorems for the schematic bi-level scheme in Eqs. (6)-(7).

Theorem 1. *(Convergence towards Global Minimum) Suppose both the above LL and UL objective convergence properties hold and $f(\mathbf{x}, \mathbf{y})$ is continuous on $\mathcal{X} \times \mathbb{R}^m$. 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 $K \to \infty$.*

Proof. For any limit point $\bar{\mathbf{x}}$ of the sequence $\{\mathbf{x}_K\}$, let $\{\mathbf{x}_l\}$ be a subsequence of $\{\mathbf{x}_K\}$ such that $\mathbf{x}_l \to \bar{\mathbf{x}} \in \mathcal{X}$. As $\{\mathbf{y}_K(\mathbf{x})\}$ is uniformly bounded on $\mathcal{X}$, we can have a subsequence $\{\mathbf{x}_m\}$ of $\{\mathbf{x}_l\}$ satisfying $\mathbf{y}_m(\mathbf{x}_m) \to \bar{\mathbf{y}}$ for some $\bar{\mathbf{y}}$. 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

$$f(\mathbf{x}_m, \mathbf{y}_m(\mathbf{x}_m)) - f^*(\mathbf{x}_m) \le \epsilon.$$

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

$$f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) - f^*(\bar{\mathbf{x}}) \le \epsilon.$$

As ϵ is arbitrarily chosen, we have $f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) - f^*(\bar{\mathbf{x}}) \le 0$ and thus $\bar{\mathbf{y}} \in \mathcal{S}(\bar{\mathbf{x}})$.

Next, as F is continuous at $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, for any $\epsilon > 0$, there exists $M(\epsilon) > 0$ such that for any $m > M(\epsilon)$, it holds

$$F(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \le F(\mathbf{x}_m, \mathbf{y}_m(\mathbf{x}_m)) + \epsilon.$$

Then, we have, for any $m > M(\epsilon)$ and $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}\varphi(\bar{\mathbf{x}}) &= \inf_{\mathbf{y}\in\mathcal{S}(\bar{\mathbf{x}})} F(\bar{\mathbf{x}}, \mathbf{y})\\ &\le F(\bar{\mathbf{x}}, \bar{\mathbf{y}})\\ &\le F(\mathbf{x}_m, \mathbf{y}_m(\mathbf{x}_m)) + \epsilon\\ &= \varphi_m(\mathbf{x}_m) + \epsilon\\ &\le \varphi_m(\mathbf{x}) + \epsilon.\end{aligned} \tag{9}$$

Taking $m \to \infty$ and by the *UL objective convergence property*, we have

$$\varphi(\bar{\mathbf{x}}) \le \lim_{m\to\infty} \varphi_m(\mathbf{x}) + \epsilon = \varphi(\mathbf{x}) + \epsilon, \quad \forall \mathbf{x} \in \mathcal{X}.$$

By taking $\epsilon \to 0$, we have

$$\varphi(\bar{\mathbf{x}}) \le \varphi(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X},$$

which implies $\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x})$.

We next show that $\inf_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) \to \inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x})$ as $K \to \infty$. Since for any $\mathbf{x} \in \mathcal{X}$,

$$\inf_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) \le \varphi_K(\mathbf{x}),$$

by taking $K \to \infty$ and with the *UL objective convergence property*, we have

$$\limsup_{K\to\infty} \left\{ \inf_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) \right\} \le \varphi(\mathbf{x}), \qquad \forall \mathbf{x} \in \mathcal{X},$$

and thus

$$\limsup_{K\to\infty} \left\{ \inf_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) \right\} \le \inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x}).$$

So, if $\inf_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) \to \inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x})$ does not hold, then there exist $\delta > 0$ and subsequence $\{\mathbf{x}_l\}$ of $\{\mathbf{x}_K\}$ such that

$$\inf_{\mathbf{x}\in\mathcal{X}} \varphi_l(\mathbf{x}) < \inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x}) - \delta, \quad \forall l. \tag{10}$$

Since $\mathcal{X}$ is compact, we can assume without loss of generality that $\mathbf{x}_l \to \bar{\mathbf{x}} \in \mathcal{X}$ by considering a subsequence. Then, as shown in above, we have $\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x})$. And, by the same arguments for deriving Eq. (9), we can show that for any $\epsilon > 0$, there exists $k(\epsilon) > 0$ such that for any $l > k(\epsilon)$, it holds

$$\varphi(\bar{\mathbf{x}}) \le \varphi_l(\mathbf{x}_l) + \epsilon.$$

By letting $l \to \infty$, $\epsilon \to 0$ and the definition of $\mathbf{x}_l$, we have

$$\inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x}) = \varphi(\bar{\mathbf{x}}) \le \liminf_{l\to\infty} \left\{ \inf_{\mathbf{x}\in\mathcal{X}} \varphi_l(\mathbf{x}) \right\},$$

which implies a contradiction to Eq. (10). Thus we have $\inf_{\mathbf{x}\in\mathcal{X}} \varphi_K(\mathbf{x}) \to \inf_{\mathbf{x}\in\mathcal{X}} \varphi(\mathbf{x})$ as $K \to \infty$.

□

Theorem 2. *(Convergence towards Local Minimum) Suppose both the LL and UL objective convergence properties hold and let $\mathbf{x}_K$ be a local minimum of $\varphi_K(\mathbf{x})$ with uniform neighborhood modulus $\delta > 0$. Then we have that any limit point $\bar{\mathbf{x}}$ of the sequence $\{\mathbf{x}_K\}$ is a local minimum of φ, i.e., there exists $\tilde{\delta} > 0$ such that*

$$\varphi(\bar{\mathbf{x}}) \le \varphi(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{B}_{\tilde{\delta}}(\bar{\mathbf{x}}) \cap \mathcal{X}.$$

Proof. For any limit point $\bar{\mathbf{x}}$ of the sequence $\{\mathbf{x}_K\}$, let $\{\mathbf{x}_l\}$ be a subsequence of $\{\mathbf{x}_K\}$ such that $\mathbf{x}_l \to \bar{\mathbf{x}} \in \mathcal{X}$ and $\mathbf{x}_l \in \mathbb{B}_{\delta/2}(\bar{\mathbf{x}})$. As $\{\mathbf{y}_K(\mathbf{x})\}$ is uniformly bounded on $\mathcal{X}$, we can have a subsequence $\{\mathbf{x}_m\}$ of $\{\mathbf{x}_l\}$ satisfying $\mathbf{y}_m(\mathbf{x}_m) \to \bar{\mathbf{y}}$ for some $\bar{\mathbf{y}}$. 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

$$f(\mathbf{x}_m, \mathbf{y}_m(\mathbf{x}_m)) - f^*(\mathbf{x}_m) \le \epsilon.$$

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

$$f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) - f^*(\bar{\mathbf{x}}) \le \epsilon.$$

As ϵ is arbitrarily chosen, we have $f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) - f^*(\bar{\mathbf{x}}) \le 0$ and thus $\bar{\mathbf{y}} \in \mathcal{S}(\bar{\mathbf{x}})$.

Next, as F is continuous at $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, for any $\epsilon > 0$, there exists $M(\epsilon) > 0$ such that for any $m > M(\epsilon)$, it holds

$$F(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \le F(\mathbf{x}_m, \mathbf{y}_m(\mathbf{x}_m)) + \epsilon.$$

Then, we have, for any $m > M(\epsilon)$ and $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned} \varphi(\bar{\mathbf{x}}) &= \inf_{\mathbf{y}\in\mathcal{S}(\bar{\mathbf{x}})} F(\bar{\mathbf{x}}, \mathbf{y}) \\ &\le F(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \\ &\le F(\mathbf{x}_m, \mathbf{y}_m(\mathbf{x}_m)) + \epsilon \\ &= \varphi_m(\mathbf{x}_m) + \epsilon. \end{aligned}$$

Next, as $\mathbf{x}_m$ is a local minimum of $\varphi_m(\mathbf{x})$ with uniform neighborhood modulus δ, it follows

$$\varphi_m(\mathbf{x}_m) \le \varphi_m(\mathbf{x}), \ \forall \mathbf{x} \in \mathbb{B}_\delta(\mathbf{x}_m) \cap \mathcal{X}.$$

Since $\mathbb{B}_{\delta/2}(\bar{\mathbf{x}}) \subseteq \mathbb{B}_{\delta/2+\|\mathbf{x}_m-\bar{\mathbf{x}}\|}(\mathbf{x}_m) \subseteq \mathbb{B}_\delta(\mathbf{x}_m)$, we have that for any $\epsilon > 0$, $\forall \mathbf{x} \in \mathbb{B}_{\delta/2}(\bar{\mathbf{x}}) \cap \mathcal{X}$, there exists $M(\epsilon) > 0$ such that whenever $m > M(\epsilon)$,

$$\varphi_m(\mathbf{x}_m) + \epsilon \le \varphi_m(\mathbf{x}) + \epsilon \le \varphi(\mathbf{x}) + \epsilon.$$

Taking $m \to \infty$ and by the UL objective convergence property, we have

$$\varphi(\bar{\mathbf{x}}) \le \lim_{m\to\infty} \varphi_m(\mathbf{x}) + \epsilon = \varphi(\mathbf{x}) + \epsilon, \quad \forall \mathbf{x} \in \mathbb{B}_{\delta/2}(\bar{\mathbf{x}}) \cap \mathcal{X}.$$

By taking $\epsilon \to 0$, we have

$$\varphi(\bar{\mathbf{x}}) \le \varphi(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{B}_{\delta/2}(\bar{\mathbf{x}}) \cap \mathcal{X},$$

which implies $\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{B}_{\delta/2}(\bar{\mathbf{x}})\cap\mathcal{X}} \varphi(\mathbf{x})$, i.e, $\bar{x}$ is a local minimum of φ. □

3.2 Convergence Analysis

The desired convergence can be successfully achieved once the embedded task-tailored iterative gradient-aggregation modules $\mathcal{T}_k$ meet the *LL objective convergence property* and the *UL objective convergence property*. As the *LL objective convergence property* is weak and hence easy to meet, it provides us more flexibility to design algorithms. In particular, the iteration module in Eq. (8) is constructed for solving simple bi-level inner subproblem in Eq. (5) with merely convex UL objective.

This section is devoted to the justification of the approximation quality and hence the convergence of our bi-level updating scheme (stated in Eqs. (6)-(7), with embedded $\mathcal{T}_k$ in Eq. (8)). Following the general proof recipe, we only need to verify that the convergence of $\mathcal{T}_k$ in Eq. (8) meets the *LL objective convergence property* and the *UL objective convergence property*. To investigate the convergence behavior of the proposed simple bi-level iterations $\mathcal{T}_k$ in Eq. (8), with fixed $\mathbf{x}$, we first introduce two auxiliary variables

$$\mathbf{z}^u_{k+1}(\mathbf{x}) = \mathbf{y}_k(\mathbf{x}) - s_u\alpha_k\nabla F(\mathbf{x}, \mathbf{y}_k(\mathbf{x}))$$

and

$$\mathbf{z}^l_{k+1}(\mathbf{x}) = \mathbf{y}_k(\mathbf{x}) - s_l\beta_k\nabla f(\mathbf{x}, \mathbf{y}_k(\mathbf{x})).$$

We further denote the optimal value and the optimal solution set of simple bi-level problem (i.e., Eq. (5)) by $\varphi(\mathbf{x})$ and $\hat{\mathcal{S}}(\mathbf{x})$, respectively. First, we show the convergence result for our proposed algorithm, i.e., Eqs. (6)-(7), with embedded $\mathcal{T}_k$ in Eq. (8).

Theorem 3. *Let $\{\mathbf{y}_k(\mathbf{x})\}$ be the sequence generated by Eq. (8) with $\alpha_k \in (0, 1]$, $\alpha_k \searrow 0$, $\sum \alpha_k = +\infty$, $\beta_k \in [\underline{\beta}, 1]$ with some*

$\underline{\beta} > 0$, $s_u \in (0, \frac{1}{L_F})$, $s_l \in (0, \frac{1}{L_f})$ *and* $\mu \in (0,1)$, *suppose that* $\mathcal{Y}$ *is compact, for any given* $\mathbf{x}$, *if* $\hat{\mathcal{S}}(\mathbf{x})$ *is nonempty , we have*

$$\lim_{k\to\infty} \operatorname{dist}(\mathbf{y}_k(\mathbf{x}), \hat{\mathcal{S}}(\mathbf{x})) = 0,$$

and then

$$\lim_{k\to\infty} F(\mathbf{x}, \mathbf{y}_k(\mathbf{x})) = \varphi(\mathbf{x}).$$

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

Theorem 4. *Let* $\{\mathbf{y}_k(\mathbf{x})\}$ *be the sequence generated by Eq.* (8) *with* $\alpha_k = \frac{1}{k+1}$, $\beta_k \in [\underline{\beta}, 1]$ *with some* $\underline{\beta} > 0$, $|\beta_k - \beta_{k-1}| \leq \frac{c_\beta}{(k+1)^2}$ *with some* $c_\beta > 0$, $s_u \in (0, \frac{1}{L_F})$, $s_l \in (0, \frac{1}{L_f})$ *and* $\mu \in (0,1)$. *Suppose* $\hat{\mathcal{S}}(\mathbf{x})$ *is nonempty,* $\mathcal{Y}$ *is compact,* $F(\mathbf{x}, \cdot)$ *is bounded below by* M_0, *we have for* $k \geq 2$,

$$\begin{aligned}
&\|\mathbf{y}_k(\mathbf{x}) - \mathbf{z}^l_{k+1}(\mathbf{x})\|^2 \leq \tfrac{(2C_2+C_3)}{\underline{\beta}^2} \tfrac{1+\ln k}{k^{\frac{1}{4}}},\\
&f(\mathbf{z}^l_{k+1}(\mathbf{x})) - \min f \leq \tfrac{D}{\underline{\beta}^2 s_l}\sqrt{(2C_2+C_3)}\sqrt{\tfrac{1+\ln k}{k^{\frac{1}{4}}}},
\end{aligned}$$

where $C_3 := \frac{D^2+2s_u(\varphi(\mathbf{x})-M_0)}{(1-\mu)(1-s_l L_f)}$, $C_2 := (s_l^2 L_f^2 D + \frac{4DL_f}{\underline{\beta}})\sqrt{C_1}$, $C_1 := \frac{C_0(D^2+2s_u(\varphi(\mathbf{x})-M_0))+2\mu s_u D M_F + 2(1-\mu)s_l c_\beta D M_f}{\min\{(1-s_l L_f),(1-s_u L_F),1\}}$ *and* $C_0 = \max\{2 + c_\beta^2/\underline{\beta}^2, 3\}$.

Now we are ready to establish our fundamental LL and UL objective convergence properties required in Theorem 1.

Theorem 5. *Suppose Assumptions 1 is satisfied,* $\mathcal{X}$ *and* $\mathcal{Y}$ *are compact, and* $\hat{\mathcal{S}}(\mathbf{x})$ *is nonempty for all* $\mathbf{x} \in \mathcal{X}$. *Let* $\{\mathbf{y}_k(\mathbf{x})\}$ *be the output generated by* (8) *with* $s_l \in (0, 1/L_f)$, $s_u \in (0, 1/L_F)$, $\mu \in (0,1)$, $\alpha_k = \frac{1}{k+1}$, $\beta_k \in [\underline{\beta}, 1]$ *with some* $\underline{\beta} > 0$, $|\beta_k - \beta_{k-1}| \leq \frac{c_\beta}{(k+1)^2}$ *with some* $c_\beta > 0$, *then we have that both the LL and UL objective convergence properties hold.*

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

$$f(\mathbf{x}, \mathbf{y}_K(\mathbf{x})) - f^*(\mathbf{x}) \leq C\sqrt{\frac{1+\ln K}{K^{\frac{1}{4}}}}.$$

As $\sqrt{\frac{1+\ln K}{K^{\frac{1}{4}}}} \to 0$ as $K \to \infty$, $\{\mathbf{y}_K(\mathbf{x})\} \subset \mathcal{Y}$, and $\mathcal{Y}$ is compact, *LL objective convergence property* holds. Next, it follows from Theorem 3 that $\varphi_K(\mathbf{x}) \to \varphi(\mathbf{x})$ as $K \to \infty$ for any $\mathbf{x} \in \mathcal{X}$ and thus *UL objective convergence property* holds. □

3.3 Proof of Theorem 3

As the identity of $\mathbf{x}$ is clear from the context, in Section 3.3 and 3.4, for succinctness we will write $\Psi(\mathbf{y})$ instead of $F(\mathbf{x}, \mathbf{y})$, Ψ^* instead of $\varphi(\mathbf{x})$, $\psi(\mathbf{y})$ instead of $f(\mathbf{x}, \mathbf{y})$, $\mathcal{S}$ instead of $\mathcal{S}(\mathbf{x})$, and $\hat{\mathcal{S}}$ instead of $\hat{\mathcal{S}}(\mathbf{x})$. Moreover, we will omit the notation $\mathbf{x}$ and use the notations $\mathbf{y}_k$, $\mathbf{z}^u_{k+1}$ and $\mathbf{z}^l_{k+1}$ instead of $\mathbf{y}_k(\mathbf{x})$, $\mathbf{z}^u_{k+1}(\mathbf{x})$ and $\mathbf{z}^l_{k+1}(\mathbf{x})$, respectively.

TABLE 2
Summarize the updating schemes of UL variable $\mathbf{x}$ and LL variable $\mathbf{y}$ under different conditions (i.e., with and without LLS assumption). "Traditional" and "Ours" respectively denote the iteration schemes with and without LLS condition.

Types	Gradient-based Methods (Traditional)	BDA (Ours)
Dynamics	$\min_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$	$\min_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x}, \mathbf{y}),$ $s.t., \mathbf{y} \in \arg\min_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$
	$\mathbf{y}_K \to \mathcal{S}(\mathbf{x})$	$\mathbf{y}_K \to \hat{\mathcal{S}}(\mathbf{x})$
Approximation	$\min_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}_K(\mathbf{x}))$	

Lemma 2. *Let* $\{\mathbf{y}_k\}$ *be the sequence generated by Eq.* (8) *with* $\alpha_k, \beta_k \in (0, 1]$, $s_u \in (0, \frac{1}{L_F})$, $s_l \in (0, \frac{1}{L_f})$ *and* $\mu \in (0,1)$, *then for any* $\mathbf{y} \in \mathcal{Y}$, *we have*

$$\begin{aligned}
&(1-\mu)\beta_k f(\mathbf{x}, \mathbf{y}) + \tfrac{\mu s_u \alpha_k}{s_l} F(\mathbf{x}, \mathbf{y})\\
&\geq (1-\mu)\beta_k f(\mathbf{x}, \mathbf{z}^l_{k+1}) + \tfrac{\mu s_u \alpha_k}{s_l} F(\mathbf{x}, \mathbf{z}^u_{k+1}) - \tfrac{1}{2s_l}\|\mathbf{y} - \mathbf{y}_k\|^2\\
&+\tfrac{1}{2s_l}\|\mathbf{y} - \mathbf{y}_{k+1}\|^2 + \tfrac{(1-\mu)}{2s_l}(1 - \beta_k s_l L_f)\|\mathbf{y}_k - \mathbf{z}^l_{k+1}\|^2\\
&+\tfrac{\mu}{2s_l}(1 - \alpha_k s_u L_F)\|\mathbf{y}_k - \mathbf{z}^u_{k+1}\|^2\\
&+\tfrac{1}{2s_l}\left\|((1-\mu)\mathbf{z}^l_{k+1} + \mu\mathbf{z}^u_{k+1}) - \mathbf{y}_{k+1}\right\|^2.
\end{aligned} \tag{11}$$

Proof. It follows from the definitions of $\mathbf{z}^u_{k+1}$ and $\mathbf{z}^l_{k+1}$ that

$$\begin{aligned}
0 &= \alpha_k \nabla\Psi(\mathbf{y}_k) + \tfrac{\mathbf{z}^u_{k+1} - \mathbf{y}_k}{s_u},\\
0 &= \beta_k \nabla\psi(\mathbf{y}_k) + \tfrac{\mathbf{z}^l_{k+1} - \mathbf{y}_k}{s_l},
\end{aligned} \tag{12}$$

and thus for any $\mathbf{y}$, we have

$$0 = \alpha_k\langle\nabla\Psi(\mathbf{y}_k), \mathbf{y} - \mathbf{z}^u_{k+1}\rangle + \langle\tfrac{\mathbf{z}^u_{k+1} - \mathbf{y}_k}{s_u}, \mathbf{y} - \mathbf{z}^u_{k+1}\rangle, \tag{13}$$

and

$$0 = \beta_k\langle\nabla\psi(\mathbf{y}_k), \mathbf{y} - \mathbf{z}^l_{k+1}\rangle + \langle\tfrac{\mathbf{z}^l_{k+1} - \mathbf{y}_k}{s_l}, \mathbf{y} - \mathbf{z}^l_{k+1}\rangle. \tag{14}$$

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

$$\begin{aligned}
&\langle\nabla\psi(\mathbf{y}_k), \mathbf{y} - \mathbf{z}^l_{k+1}\rangle\\
&= \langle\nabla\psi(\mathbf{y}_k), \mathbf{y} - \mathbf{y}_k\rangle + \langle\nabla\psi(\mathbf{y}_k), \mathbf{y}_k - \mathbf{z}^l_{k+1}\rangle\\
&\leq \psi(\mathbf{y}) - \psi(\mathbf{y}_k) + \psi(\mathbf{y}_k) - \psi(\mathbf{z}^l_{k+1}) + \tfrac{L_f}{2}\|\mathbf{y}_k - \mathbf{z}^l_{k+1}\|^2\\
&= \psi(\mathbf{y}) - \psi(\mathbf{z}^l_{k+1}) + \tfrac{L_f}{2}\|\mathbf{y}_k - \mathbf{z}^l_{k+1}\|^2.
\end{aligned} \tag{15}$$

Combining with $\langle\mathbf{z}^l_{k+1} - \mathbf{y}_k, \mathbf{y} - \mathbf{z}^l_{k+1}\rangle = \frac{1}{2}(\|\mathbf{y} - \mathbf{y}_k\|^2 - \|\mathbf{y} - \mathbf{z}^l_{k+1}\|^2 - \|\mathbf{y}_k - \mathbf{z}^l_{k+1}\|^2)$ and Eq. (14) yields

$$\begin{aligned}
\beta_k\psi(\mathbf{y}) \geq \beta_k\psi(\mathbf{z}^l_{k+1}) &- \tfrac{1}{2s_l}\|\mathbf{y} - \mathbf{y}_k\|^2 + \tfrac{1}{2s_l}\|\mathbf{y} - \mathbf{z}^l_{k+1}\|^2\\
&+\tfrac{1}{2s_l}(1 - \beta_k s_l L_f)\|\mathbf{y}_k - \mathbf{z}^l_{k+1}\|^2.
\end{aligned} \tag{16}$$

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

$$\begin{aligned}
\alpha_k\Psi(\mathbf{y}) \geq \alpha_k\Psi(\mathbf{z}^u_{k+1}) &- \tfrac{1}{2s_u}\|\mathbf{y} - \mathbf{y}_k\|^2 + \tfrac{1}{2s_u}\|\mathbf{y} - \mathbf{z}^u_{k+1}\|^2\\
&+\tfrac{1}{2s_u}(1 - \alpha_k s_u L_F)\|\mathbf{y}_k - \mathbf{z}^u_{k+1}\|^2.
\end{aligned} \tag{17}$$

Multiplying Eq. (16) and Eq. (17) by $1-\mu$ and $\frac{s_u\mu}{s_l}$, respectively, and then summing them up implies that

$$\begin{aligned}&(1-\mu)\beta_k\psi(\mathbf{y})+\tfrac{\mu s_u\alpha_k}{s_l}\Psi(\mathbf{y})\\&\geq(1-\mu)\beta_k\psi(\mathbf{z}_{k+1}^l)+\tfrac{\mu s_u\alpha_k}{s_l}\Psi(\mathbf{z}_{k+1}^u)-\tfrac{1}{2s_l}\|\mathbf{y}-\mathbf{y}_k\|^2\\&+\tfrac{1}{2s_l}\left((1-\mu)\|\mathbf{y}-\mathbf{z}_{k+1}^l\|^2+\mu\|\mathbf{y}-\mathbf{z}_{k+1}^u\|^2\right)\\&+\tfrac{(1-\mu)}{2s_l}(1-\beta_ks_lL_f)\|\mathbf{y}_k-\mathbf{z}_{k+1}^l\|^2\\&+\tfrac{\mu}{2s_l}(1-\alpha_ks_uL_F)\|\mathbf{y}_k-\mathbf{z}_{k+1}^u\|^2.\end{aligned}\tag{18}$$

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

$$\begin{aligned}&(1-\mu)\|\mathbf{y}-\mathbf{z}_{k+1}^l\|^2+\mu\|\mathbf{y}-\mathbf{z}_{k+1}^u\|^2\\&\geq\|\mathbf{y}-\left((1-\mu)\mathbf{z}_{k+1}^l+\mu\mathbf{z}_{k+1}^u\right)\|^2.\end{aligned}$$

Next, as $\mathtt{Proj}_{\mathcal{Y}}$ is firmly nonexpansive (see, e.g.,[34, Proposition 4.8]), for any $\mathbf{y}\in\mathcal{Y}$, we have

$$\begin{aligned}&\left\|\mathbf{y}-\left((1-\mu)\mathbf{z}_{k+1}^l+\mu\mathbf{z}_{k+1}^u\right)\right\|^2\\&\geq\|\mathbf{y}-\mathbf{y}_{k+1}\|^2+\left\|\left((1-\mu)\mathbf{z}_{k+1}^l+\mu\mathbf{z}_{k+1}^u\right)-\mathbf{y}_{k+1}\right\|^2.\end{aligned}\tag{19}$$

Then, since $\alpha_k,\beta_k\leq1$, we obtain form Eq. (18) that for any $\mathbf{y}\in\mathcal{Y}$,

$$\begin{aligned}&(1-\mu)\beta_k\psi(\mathbf{y})+\tfrac{\mu s_u\alpha_k}{s_l}\Psi(\mathbf{y})\\&\geq(1-\mu)\beta_k\psi(\mathbf{z}_{k+1}^l)+\tfrac{\mu s_u\alpha_k}{s_l}\Psi(\mathbf{z}_{k+1}^u)-\tfrac{1}{2s_l}\|\mathbf{y}-\mathbf{y}_k\|^2\\&+\tfrac{1}{2s_l}\|\mathbf{y}-\mathbf{y}_{k+1}\|^2+\tfrac{(1-\mu)}{2s_l}(1-\beta_ks_lL_f)\|\mathbf{y}_k-\mathbf{z}_{k+1}^l\|^2\\&+\tfrac{\mu}{2s_l}(1-\alpha_ks_uL_F)\|\mathbf{y}_k-\mathbf{z}_{k+1}^u\|^2\\&+\tfrac{1}{2s_l}\left\|\left((1-\mu)\mathbf{z}_{k+1}^l+\mu\mathbf{z}_{k+1}^u\right)-\mathbf{y}_{k+1}\right\|^2.\end{aligned}\tag{20}$$

This completes the proof. □

Lemma 3. *Let $\{a_k\}$ and $\{b_k\}$ be sequences of non-negative real numbers. Assume that there exists $n_0\in\mathbb{N}$ such that*

$$a_{k+1}+b_k-a_k\leq0,\quad\forall k\geq n_0.$$

Then $\lim_{k\to\infty}a_k$ exists and $\sum_{k=1}^{\infty}b_k<\infty$.

Proof. Adding the inequality

$$a_{k+1}+b_k-a_k\leq0,$$

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. □

Lemma 4. *Let $\{\mathbf{y}_k\}$ be the sequence generated by Eq. (8) with $\alpha_k\in(0,1]$, $\beta_k\in(0,1]$, $s_u\in(0,\frac{1}{L_F})$, $s_l\in(0,\frac{1}{L_f})$ and $\mu\in(0,1)$, then for any $\bar{\mathbf{y}}\in\mathcal{S}(\mathbf{x})$, we have*

$$\|\mathbf{z}_{k+1}^l-\bar{\mathbf{y}}\|\leq\|\mathbf{y}_k-\bar{\mathbf{y}}\|.\tag{21}$$

Furthermore, when $\mathcal{Y}$ is compact, sequences $\{\mathbf{y}_k\}$, $\{\mathbf{z}_k^l\}$, $\{\mathbf{z}_k^u\}$ are all bounded.

Proof. According to [34, Proposition 4.8, Proposition 4.33, Corollary 18.16], we know that when $0\leq\beta_ks_l\leq\frac{1}{L_f}$, $0\leq\alpha_ks_u\leq\frac{1}{L_F}$, operators $Id-\beta_ks_l\nabla\psi$ and $Id-\alpha_ks_u\nabla\Psi$ are both nonexpansive (i.e., 1-Lipschitz continuous). Then, since $\mathbf{z}_{k+1}^l=\mathbf{y}_k-\beta_ks_l\nabla\psi(\mathbf{y}_k)$ and $\bar{\mathbf{y}}=\bar{\mathbf{y}}-\beta_ks_l\nabla\psi(\bar{\mathbf{y}})$ for any $\bar{\mathbf{y}}\in\mathcal{S}$, we have

$$\begin{aligned}\|\mathbf{z}_{k+1}^l-\bar{\mathbf{y}}\|&=\|\mathbf{y}_k-\beta_ks_l\nabla\psi(\mathbf{y}_k)-\bar{\mathbf{y}}+\beta_ks_l\nabla\psi(\bar{\mathbf{y}})\|\\&\leq\|\mathbf{y}_k-\bar{\mathbf{y}}\|.\end{aligned}$$

If $\mathcal{Y}$ is compact, then the desired boundedness of $\{\mathbf{y}_k\}$ follows directly from the iteration scheme in Eq. (8). And it follows from $\|\mathbf{z}_{k+1}^l-\bar{\mathbf{y}}\|\leq\|\mathbf{y}_k-\bar{\mathbf{y}}\|$ that $\{\mathbf{z}_k^l\}$ is bounded. Next, because

$$\|\mathbf{z}_{k+1}^u-(\bar{\mathbf{y}}-\alpha_ks_u\nabla\Psi(\bar{\mathbf{y}}))\|\leq\|\mathbf{y}_k-\bar{\mathbf{y}}\|,$$

and $\alpha_k\in(0,1]$, we have $\{\mathbf{z}_k^u\}$ is bounded. □

Now, we are ready to give the proof of Theorem 3.

Proof of Theorem 3. Let $\delta>0$ be a constant satisfying $\delta<\frac{1}{2s_l}\min\{(1-\mu)(1-s_lL_f),\mu(1-s_uL_F)\}$. We consider a sequence of $\{\tau_n\}$ defined by

$$\begin{aligned}\tau_n:=\max\Big\{k\in\mathbb{N}\mid k\leq n\text{ and }&\delta\|\mathbf{y}_{k-1}-\mathbf{z}_k^l\|^2\\&+\delta\|\mathbf{y}_{k-1}-\mathbf{z}_k^u\|^2+\tfrac{1}{4s_l}\left\|\left((1-\mu)\mathbf{z}_k^l+\mu\mathbf{z}_k^u\right)-\mathbf{y}_k\right\|^2\\&+\tfrac{\mu s_u\alpha_{k-1}}{s_l}\left(\Psi(\mathbf{z}_k^u)-\Psi^*\right)<0\Big\}.\end{aligned}$$

Inspired by [35], we consider the following two cases: (a) $\{\tau_n\}$ is finite, i.e., there exists $k_0\in\mathbb{N}$ such that

$$\begin{aligned}&\delta\|\mathbf{y}_{k-1}-\mathbf{z}_k^l\|^2+\tfrac{1}{4s_l}\left\|\left((1-\mu)\mathbf{z}_k^l+\mu\mathbf{z}_k^u\right)-\mathbf{y}_k\right\|^2\\&+\delta\|\mathbf{y}_{k-1}-\mathbf{z}_k^u\|^2+\tfrac{\mu s_u\alpha_{k-1}}{s_l}\left(\Psi(\mathbf{z}_k^u)-\Psi^*\right)\geq0,\end{aligned}$$

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

$$\begin{aligned}&\delta\|\mathbf{y}_{k-1}-\mathbf{z}_k^u\|^2+\tfrac{1}{4s_l}\left\|\left((1-\mu)\mathbf{z}_k^l+\mu\mathbf{z}_k^u\right)-\mathbf{y}_k\right\|^2+\\&\delta\|\mathbf{y}_{k-1}-\mathbf{z}_k^l\|^2+\tfrac{\mu s_u\alpha_{k-1}}{s_l}\left(\Psi(\mathbf{z}_k^u)-\Psi^*\right)<0.\end{aligned}$$

Case (a): We assume that $\{\tau_n\}$ is finite and there exists $k_0\in\mathbb{N}$ such that

$$\begin{aligned}&\delta\|\mathbf{y}_{k-1}-\mathbf{z}_k^l\|^2+\tfrac{1}{4s_l}\left\|\left((1-\mu)\mathbf{z}_k^l+\mu\mathbf{z}_k^u\right)-\mathbf{y}_k\right\|^2\\&+\delta\|\mathbf{y}_{k-1}-\mathbf{z}_k^u\|^2+\tfrac{\mu s_u\alpha_{k-1}}{s_l}\left(\Psi(\mathbf{z}_k^u)-\Psi^*\right)\geq0,\end{aligned}\tag{22}$$

for all $k\geq k_0$. Let $\bar{\mathbf{y}}$ be any point in $\hat{\mathcal{S}}$, setting $\mathbf{y}=\bar{\mathbf{y}}$ in Eq. (11), as $\psi(\bar{\mathbf{y}})=\min_{\mathbf{y}\in\mathbb{R}^n}\psi(\mathbf{y})\leq\psi(\mathbf{z}_{k+1}^l)$, $\mu\in(0,1)$ and $\alpha_k,\beta_k\leq1$, we have

$$\begin{aligned}&\tfrac{1}{2s_l}\|\bar{\mathbf{y}}-\mathbf{y}_k\|^2\\&\geq\tfrac{1}{2s_l}\|\bar{\mathbf{y}}-\mathbf{y}_{k+1}\|^2+\left(\tfrac{(1-\mu)(1-s_lL_f)}{2s_l}-\delta\right)\|\mathbf{y}_k-\mathbf{z}_{k+1}^l\|^2\\&+\left(\tfrac{\mu(1-s_uL_F)}{2s_l}-\delta\right)\|\mathbf{y}_k-\mathbf{z}_{k+1}^u\|^2\\&+\tfrac{1}{4s_l}\left\|\left((1-\mu)\mathbf{z}_{k+1}^l+\mu\mathbf{z}_{k+1}^u\right)-\mathbf{y}_{k+1}\right\|^2\\&+\delta\|\mathbf{y}_k-\mathbf{z}_{k+1}^l\|^2+\delta\|\mathbf{y}_k-\mathbf{z}_{k+1}^u\|^2\\&+\tfrac{1}{4s_l}\left\|\left((1-\mu)\mathbf{z}_{k+1}^l+\mu\mathbf{z}_{k+1}^u\right)-\mathbf{y}_{k+1}\right\|^2\\&+\tfrac{\mu s_u\alpha_k}{s_l}\left(\Psi(\mathbf{z}_{k+1}^u)-\Psi^*\right).\end{aligned}\tag{23}$$

For all $k\geq k_0$, $0<\delta<\frac{1}{2s_l}\min\{(1-\mu)(1-s_lL_f),\mu(1-s_uL_F)\}$ and Eq. (23) imply

$$\begin{aligned}&\tfrac{1}{2s_l}\|\bar{\mathbf{y}}-\mathbf{y}_k\|^2\\&\geq\tfrac{1}{2s_l}\|\bar{\mathbf{y}}-\mathbf{y}_{k+1}\|^2+\delta\|\mathbf{y}_k-\mathbf{z}_{k+1}^l\|^2+\delta\|\mathbf{y}_k-\mathbf{z}_{k+1}^u\|^2\\&+\tfrac{1}{4s_l}\left\|\left((1-\mu)\mathbf{z}_{k+1}^l+\mu\mathbf{z}_{k+1}^u\right)-\mathbf{y}_{k+1}\right\|^2\\&+\tfrac{\mu s_u\alpha_k}{s_l}\left(\Psi(\mathbf{z}_{k+1}^u)-\Psi^*\right).\end{aligned}$$

Combining with Eq. (22), it follows from Lemma 3 that

$$\begin{aligned}
&\sum_{k=0}^{\infty} \|\mathbf{y}_k - \mathbf{z}_{k+1}^l\|^2 < \infty,\\
&\sum_{k=0}^{\infty} \|\mathbf{y}_k - \mathbf{z}_{k+1}^u\|^2 < \infty,\\
&\sum_{k=0}^{\infty} \left\|\left((1-\mu)\mathbf{z}_{k+1}^l + \mu\mathbf{z}_{k+1}^u\right) - \mathbf{y}_{k+1}\right\|^2 < \infty,\\
&\sum_{k=0}^{\infty} \alpha_k \left(\Psi(\mathbf{z}_{k+1}^u) - \Psi^*\right) < \infty,
\end{aligned}$$

and $\lim_{k\to\infty} \|\bar{\mathbf{y}} - \mathbf{y}_k\|^2$ exists.

We now show that there exists subsequence $\{\mathbf{y}_\ell\} \subseteq \{\mathbf{y}_k\}$ such that $\lim_{\ell\to\infty} \Psi(\mathbf{y}_\ell) \le \Psi^*$. This is obviously true if for any $\hat{k} > 0$, there exists $k > \hat{k}$ such that $\Psi(\mathbf{y}_k) \le \Psi^*$. Thus, we just need to consider the case where there exists $\hat{k} > 0$ such that $\Psi(\mathbf{y}_k) > \Psi^*$ for all $k \ge \hat{k}$. If there does not exist subsequence $\{\mathbf{y}_\ell\} \subseteq \{\mathbf{y}_k\}$ such that $\lim_{\ell\to\infty} \Psi(\mathbf{y}_\ell) \le \Psi^*$, there must exist $\epsilon > 0$ and $k_1 \ge \max\{\hat{k}, k_0\}$ such that $\Psi(\mathbf{y}_k) - \Psi^* \ge 2\epsilon$ for all $k \ge k_1$. As $\mathcal{Y}$ is compact, it follows from Lemma 4 that sequences $\{\mathbf{y}_k\}$ and $\{\mathbf{z}_k^u\}$ are both bounded. Then since Ψ is continuous and $\lim_{k\to\infty} \|\mathbf{y}_k - \mathbf{z}_{k+1}^u\| = 0$, there exists $k_2 \ge k_1$ such that $|\Psi(\mathbf{y}^k) - \Psi(\mathbf{z}_{k+1}^u)| < \epsilon$ for all $k \ge k_2$ and thus $\Psi(\mathbf{z}_{k+1}^u) - \Psi^* \ge \epsilon$ for all $k \ge k_2$. Then we have

$$\epsilon \sum_{k=k_2}^{\infty} \alpha_k \le \sum_{k=k_2}^{\infty} \alpha_k \left(\Psi(\mathbf{z}_{k+1}^u) - \Psi^*\right) < \infty,$$

where the last inequality follows from $\sum_{k=0}^{\infty} \alpha_k \left(\Psi(\mathbf{z}_{k+1}^u) - \Psi^*\right) < \infty$. This result contradicts to the assumption $\sum_{k=0}^{\infty} \alpha_k = +\infty$. As $\{\mathbf{y}_\ell\}$ is bounded, we can assume without loss of generality that $\lim_{\ell\to\infty} \mathbf{y}_\ell = \tilde{\mathbf{y}}$ by taking a subsequence. By the continuity of Ψ, we have $\Psi(\tilde{\mathbf{y}}) = \lim_{\ell\to\infty} \Psi(\mathbf{y}_\ell) \le \Psi^*$. Next, let $k = \ell$ and $\ell \to \infty$ in Eq. (12) , by the continuity of $\nabla\psi$, $\beta_k \ge \underline{\beta} > 0$, and $\lim_{k\to\infty} \|\mathbf{y}_k - \mathbf{z}_{k+1}^l\| = 0$, we have

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

and thus $\tilde{\mathbf{y}} \in \mathcal{S}$. Combining with $\Psi(\tilde{\mathbf{y}}) \le \Psi^*$, we show that $\tilde{\mathbf{y}} \in \hat{\mathcal{S}}$. Then by taking $\bar{\mathbf{y}} = \tilde{\mathbf{y}}$ and since $\lim_{k\to\infty} \|\bar{\mathbf{y}} - \mathbf{y}_k\|^2$ exists, we have $\lim_{k\to\infty} \|\bar{\mathbf{y}} - \mathbf{y}_k\|^2 = 0$ and thus $\lim_{k\to\infty} \operatorname{dist}(\mathbf{y}_k, \hat{\mathcal{S}}) = 0$.

Case (b): We assume that $\{\tau_n\}$ is not finite and for any $k_0 \in \mathbb{N}$, there exists $k \ge k_0$ such that $\delta\|\mathbf{y}_{k-1} - \mathbf{z}_k^l\|^2 + \delta\|\mathbf{y}_{k-1} - \mathbf{z}_k^u\|^2 + \frac{1}{4s_l}\left\|\left((1-\mu)\mathbf{z}_k^l + \mu\mathbf{z}_k^u\right) - \mathbf{y}^k\right\|^2 + \frac{\mu s_u \alpha_{k-1}}{s_l}\left(\Psi(\mathbf{z}_k^u) - \Psi^*\right) < 0$. It follows from the assumption that τ_n is well defined for n large enough and $\lim_{n\to\infty} \tau_n = +\infty$. We assume without loss of generality that τ_n is well defined for all n.

By setting $\mathbf{y} = \texttt{Proj}_{\hat{\mathcal{S}}}(\mathbf{y}_k)$ in Eq. (11), we have

$$\begin{aligned}
&\tfrac{1}{2s_l}\operatorname{dist}^2(\mathbf{y}_k, \hat{\mathcal{S}})\\
&\ge \tfrac{1}{2s_l}\operatorname{dist}^2(\mathbf{y}_{k+1}, \hat{\mathcal{S}}) + \left(\tfrac{(1-\mu)(1-s_l L_f)}{2s_l} - \delta\right)\|\mathbf{y}_k - \mathbf{z}_{k+1}^l\|^2\\
&+ \left(\tfrac{\mu(1-s_u L_F)}{2s_l} - \delta\right)\|\mathbf{y}_k - \mathbf{z}_{k+1}^u\|^2 + \delta\|\mathbf{y}_k - \mathbf{z}_{k+1}^l\|^2\\
&+ \delta\|\mathbf{y}_k - \mathbf{z}_{k+1}^u\|^2 + \tfrac{1}{4s_l}\left\|\left((1-\mu)\mathbf{z}_{k+1}^l + \mu\mathbf{z}_{k+1}^u\right) - \mathbf{y}_{k+1}\right\|^2\\
&+ \tfrac{1}{4s_l}\left\|\left((1-\mu)\mathbf{z}_{k+1}^l + \mu\mathbf{z}_{k+1}^u\right) - \mathbf{y}_{k+1}\right\|^2\\
&+ \tfrac{\mu s_u \alpha_k}{s_l}\left(\Psi(\mathbf{z}_{k+1}^u) - \Psi^*\right) + \beta_k\left(\psi(\mathbf{z}_{k+1}^l) - \min\psi\right).
\end{aligned} \tag{24}$$

Suppose $\tau_n \le n-1$, and by the definition of τ_n, we have

$$\begin{aligned}
\delta\|\mathbf{y}_k - \mathbf{z}_{k+1}^l\|^2 + \delta\|\mathbf{y}_k - \mathbf{z}_{k+1}^u\|^2 + \tfrac{\mu s_u \alpha_k}{s_l}\left(\Psi(\mathbf{z}_{k+1}^u) - \Psi^*\right)\\
+ \tfrac{1}{4s_l}\left\|\left((1-\mu)\mathbf{z}_{k+1}^l + \mu\mathbf{z}_{k+1}^u\right) - \mathbf{y}_{k+1}\right\|^2 \ge 0,
\end{aligned}$$

for all $\tau_n \le k \le n-1$. Then

$$h_{k+1} - h_k \le 0, \quad \tau_n \le k \le n-1, \tag{25}$$

where $h_k := \frac{1}{2s_l}\operatorname{dist}^2(\mathbf{y}_k, \hat{\mathcal{S}})$. Adding these $n - \tau_n$ inequalities, we have

$$h_n \le h_{\tau_n}. \tag{26}$$

Eq. (26) is also true when $\tau_n = n$ because $h_{\tau_n} = h_n$. Once we are able to show that $\lim_{n\to\infty} h_{\tau_n} = 0$, we can obtain from Eq. (26) that $\lim_{n\to\infty} h_n = 0$.

By the definition of $\{\tau_n\}$, $\Psi^* > \Psi(\mathbf{z}_k^u)$ for all $k \in \{\tau_n\}$. Since $\mathcal{Y}$ is compact, according to Lemma 4, both $\{\mathbf{y}_{\tau_n}\}$ and $\{\mathbf{z}_{\tau_n}^u\}$ are bounded, and hence $\{h_{\tau_n}\}$ is bounded. As Ψ is assumed to be continuous, there exists M_0 such that

$$0 \le \Psi^* - \Psi(\mathbf{z}_k^u) \le \Psi^* - M_0.$$

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

$$\begin{aligned}
&\delta(\|\mathbf{y}_{k-1} - \mathbf{z}_k^l\|^2 + \|\mathbf{y}_{k-1} - \mathbf{z}_k^u\|^2)\\
&+ \tfrac{1}{4s_l}\left\|\left((1-\mu)\mathbf{z}_k^l + \mu\mathbf{z}_k^u\right) - \mathbf{y}_k\right\|^2\\
&< \tfrac{\mu s_u \alpha_{k-1}}{s_l}\left(\Psi^* - \Psi(\mathbf{z}_k^u)\right) \le \tfrac{\mu s_u \alpha_{k-1}}{s_l}\left(\Psi^* - M_0\right).
\end{aligned}$$

As $\lim_{n\to\infty} \tau_n = +\infty$, $\alpha_k \to 0$, we have

$$\begin{aligned}
\lim_{n\to\infty} \|\mathbf{y}_{\tau_n - 1} - \mathbf{z}_{\tau_n}^l\| = 0,\\
\lim_{n\to\infty} \|\mathbf{y}_{\tau_n - 1} - \mathbf{z}_{\tau_n}^u\| = 0,\\
\lim_{n\to\infty} \|\left((1-\mu)\mathbf{z}_{\tau_n}^l + \mu\mathbf{z}_{\tau_n}^u\right) - \mathbf{y}_{\tau_n}\| = 0.
\end{aligned}$$

Let $\tilde{\mathbf{y}}$ be any limit point of $\{\mathbf{y}_{\tau_n}\}$, and $\{\mathbf{y}_\ell\}$ be the subsequence of $\{\mathbf{y}_{\tau_n}\}$ such that

$$\lim_{\ell\to\infty} \mathbf{y}_\ell = \tilde{\mathbf{y}},$$

as $\lim_{n\to\infty} \|\mathbf{y}_{\tau_n - 1} - \mathbf{y}_{\tau_n}\| \le \lim_{n\to\infty}(\|\mathbf{y}_{\tau_n - 1} - \left((1-\mu)\mathbf{z}_{\tau_n}^l + \mu\mathbf{z}_{\tau_n}^u\right)\| + \|\left((1-\mu)\mathbf{z}_{\tau_n}^l + \mu\mathbf{z}_{\tau_n}^u\right) - \mathbf{y}_{\tau_n}\|) = 0$. We have $\lim_{\ell\to\infty} \mathbf{y}_{\ell-1} = \tilde{\mathbf{y}}$. Let $k = \ell - 1$ and $\ell \to \infty$ in Eq. (12), by the continuity of $\nabla\psi$, $\beta_k \ge \underline{\beta} > 0$ and $\lim_{\ell\to\infty} \|\mathbf{y}_{\ell-1} - \mathbf{z}_\ell^l\| = 0$. Then, we have

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

and thus $\tilde{\mathbf{y}} \in \mathcal{S}$. As $\Psi^* > \Psi(\mathbf{z}_k^u)$ for all $k \in \{\tau_n\}$ and hence $\Psi^* > \Psi(\mathbf{z}_\ell^u)$ for all ℓ. Then it follows from the continuity of Ψ and $\lim_{n\to\infty} \|\mathbf{z}_{\tau_n}^u - \mathbf{y}_{\tau_n}\| = 0$ that $\Psi^* \ge \Psi(\tilde{\mathbf{y}})$, which implies $\tilde{\mathbf{y}} \in \hat{\mathcal{S}}$ and $\lim_{\ell\to 0} h_\ell = 0$. Now, as we have shown above that $\tilde{\mathbf{y}} \in \hat{\mathcal{S}}$ for any limit point $\tilde{\mathbf{y}}$ of $\{\mathbf{y}_{\tau_n}\}$, we can obtain from the boundness of $\{\mathbf{y}_{\tau_n}\}$ and $\{h_{\tau_n}\}$ that $\lim_{n\to\infty} h_{\tau_n} = 0$. Thus $\lim_{n\to\infty} h_n = 0$, and $\lim_{k\to\infty} \operatorname{dist}(\mathbf{y}_k, \mathcal{S}) = 0$. □

3.4 Proof of Theorem 4

Lemma 5. *Let $\{\mathbf{y}_k\}$ be the sequence generated by Eq. (8) with $\alpha_k = \frac{1}{k+1}$, $\beta_k \in [\underline{\beta}, 1]$ with some $\underline{\beta} > 0$, $|\beta_k - \beta_{k-1}| \le \frac{c_\beta}{(k+1)^2}$ with some $c_\beta > 0$, $s_u \in (0, \frac{1}{L_F})$, $s_l \in (0, \frac{1}{L_f})$ and $\mu \in (0,1)$, then for any $\bar{\mathbf{y}} \in \mathcal{S}(\mathbf{x})$, we have*

$$\begin{aligned}
\|\mathbf{y}_{k+1} - \mathbf{y}_k\|^2 \le \|\mathbf{y}_k - \mathbf{y}_{k-1}\|^2 + \tfrac{\mu}{(k+1)^2}\|\mathbf{y}_{k-1} - \mathbf{z}_k^u\|^2\\
+ \tfrac{2(1-\mu)s_l c_\beta D M_f}{(k+1)^2} + \tfrac{2\mu s_u D M_F}{k(k+1)} + \tfrac{(1-\mu)c_\beta^2}{\underline{\beta}^2(k+1)^4}\|\mathbf{y}_{k-1} - \mathbf{z}_k^l\|^2.
\end{aligned}$$

Proof. According to [34, Proposition 4.8, Proposition 4.33, Corollary 18.16], we know that when $0 \leq \beta_k s_l \leq \frac{1}{L_f}$, $0 \leq \alpha_k s_u \leq \frac{1}{L_F}$, operators $Id - \beta_k s_l \nabla \psi$, $Id - \alpha_k s_u \nabla \Psi$ and $\texttt{Proj}_{\mathcal{Y}}$ are all nonexpansive (i.e., 1-Lipschitz continuous). Next, as

$$\begin{aligned}\mathbf{y}_{k+1} &= \texttt{Proj}_{\mathcal{Y}}\left(\mu \mathbf{z}_{k+1}^u + (1-\mu)\mathbf{z}_{k+1}^l\right) \\ &= \texttt{Proj}_{\mathcal{Y}}\left(\mathbf{y}_k - (\mu\alpha_k s_u \nabla\Psi(\mathbf{y}_k) + (1-\mu)\beta_k s_l \nabla\psi(\mathbf{y}_k))\right),\end{aligned}$$

by denoting $\Delta_\alpha^k := \alpha_k - \alpha_{k-1}$ and $\Delta_\beta^k := \beta_k - \beta_{k-1}$, we have the following inequality

$$\begin{aligned}&\|\mathbf{y}_{k+1} - \mathbf{y}_k\|^2 \\ &\leq \mu\|\mathbf{z}_{k+1}^u - \mathbf{z}_k^u\|^2 + (1-\mu)\|\mathbf{z}_{k+1}^l - \mathbf{z}_k^l\|^2, \\ &\leq \mu\|(Id - \alpha_k s_u \nabla\Psi)(\mathbf{y}_k - \mathbf{y}_{k-1})\|^2 + \mu s_u^2 |\Delta_\alpha^k|^2 \|\nabla\Psi(\mathbf{y}_{k-1})\|^2 \\ &+2\mu s_u |\delta_\alpha^k| \|(Id - \alpha_k s_u \nabla\Psi)(\mathbf{y}_k - \mathbf{y}_{k-1})\| \|\nabla\Psi(\mathbf{y}_{k-1})\| \\ &+(1-\mu)\|(Id - \beta_k s_l \nabla\psi)(\mathbf{y}_k - \mathbf{y}_{k-1})\|^2 \\ &+2(1-\mu)s_l|\Delta_\beta^k| \|(Id - \beta_k s_l \nabla\psi)(\mathbf{y}_k - \mathbf{y}_{k-1})\| \|\nabla\psi(\mathbf{y}_{k-1})\| \\ &+(1-\mu)s_l^2|\Delta_\beta^k|^2\|\nabla\psi(\mathbf{y}_{k-1})\|^2 \\ &\leq \|\mathbf{y}_k - \mathbf{y}_{k-1}\|^2 + 2\mu s_u |\Delta_\alpha^k| \|\mathbf{y}_k - \mathbf{y}_{k-1}\| \|\nabla\Psi(\mathbf{y}_{k-1})\| \\ &+\frac{\mu|\Delta_\alpha^k|^2}{\alpha_{k-1}^2}\|\mathbf{y}_{k-1} - \mathbf{z}_k^u\|^2 + \frac{(1-\mu)|\Delta_\beta^k|^2}{\beta_{k-1}^2}\|\mathbf{y}_{k-1} - \mathbf{z}_k^l\|^2 \\ &+2(1-\mu)s_l|\Delta_\beta^k| \|\mathbf{y}_k - \mathbf{y}_{k-1}\| \|\nabla\psi(\mathbf{y}_{k-1})\|,\end{aligned}$$

where the first inequality follows from the nonexpansiveness of $\texttt{Proj}_{\mathcal{Y}}$ and the convexity of $\|\cdot\|^2$, the second inequality comes from the definitions of $\mathbf{z}_k^u, \mathbf{z}_k^l$ and the last inequality follows from the nonexpansiveness of $Id - \beta_k s_l \nabla\psi$ and $Id - \alpha_k s_u \nabla\Psi$ and the definitions of $\mathbf{z}_k^u, \mathbf{z}_k^l$. Then, since $\alpha_k = \frac{1}{k+1}$, $\beta_k \geq \underline{\beta} > 0$, $|\beta_k - \beta_{k-1}| \leq \frac{c_\beta}{(k+1)^2}$, $D = \sup_{\mathbf{y},\mathbf{y}' \in \mathcal{Y}} \|\mathbf{y} - \mathbf{y}'\|$, $\sup_{\mathbf{y}\in\mathcal{Y}} \|\nabla\Psi(\mathbf{y})\| \leq M_\Psi$ and $\sup_{\mathbf{y}\in\mathcal{Y}} \|\nabla\psi(\mathbf{y})\| \leq M_\psi$, we have the following result

$$\begin{aligned}\|\mathbf{y}_{k+1} - \mathbf{y}_k\|^2 \leq\ &\|\mathbf{y}_k - \mathbf{y}_{k-1}\|^2 + \tfrac{\mu}{(k+1)^2}\|\mathbf{y}_{k-1} - \mathbf{z}_k^u\|^2 \\ &+\tfrac{2(1-\mu)s_l c_\beta D M_\psi}{(k+1)^2} + \tfrac{2\mu s_u D M_\Psi}{k(k+1)} + \tfrac{(1-\mu)c_\beta^2}{\underline{\beta}^2(k+1)^4}\|\mathbf{y}_{k-1} - \mathbf{z}_k^l\|^2.\end{aligned}$$

□

Proof of Theorem 4. Let $\bar{\mathbf{y}}$ be any point in $\mathcal{S}$, and set $\mathbf{y} = \bar{\mathbf{y}}$ in Eq. (11), since $\psi(\bar{\mathbf{y}}) = \min_{\mathbf{y}\in\mathbb{R}^n} \psi(\mathbf{y}) \leq \psi(\mathbf{z}_{k+1}^l)$, we have

$$\begin{aligned}&\tfrac{1}{2}\|\bar{\mathbf{y}} - \mathbf{y}_k\|^2 + \tfrac{\mu s_u}{k+1}\left(\Psi^* - \Psi(\mathbf{z}_u^{k+1})\right) \\ &\geq \tfrac{1}{2}\|\bar{\mathbf{y}} - \mathbf{y}_{k+1}\|^2 + \tfrac{1}{2}(1-\mu)(1-\beta_k s_l L_f)\|\mathbf{y}_k - \mathbf{z}_{k+1}^l\|^2 \\ &+\tfrac{1}{2}\mu(1-\alpha_k s_u L_F)\|\mathbf{y}_k - \mathbf{z}_{k+1}^u\|^2 \\ &+\tfrac{1}{2}\left\|\left((1-\mu)\mathbf{z}_{k+1}^l + \mu\mathbf{z}_{k+1}^u\right) - \mathbf{y}_{k+1}\right\|^2\end{aligned} \tag{27}$$

Adding the Eq. (27) from $k = 0$ to $k = n-1$, and since $\alpha_k, \beta_k \in (0, 1]$, we have

$$\begin{aligned}&\tfrac{1}{2}\|\bar{\mathbf{y}} - \mathbf{y}_n\|^2 + \tfrac{1}{2}(1-\mu)(1-s_l L_f)\sum_{k=0}^{n-1}\|\mathbf{y}_k - \mathbf{z}_{k+1}^l\|^2 \\ &+\tfrac{1}{2}\mu(1-s_u L_F)\sum_{k=0}^{n-1}\|\mathbf{y}_k - \mathbf{z}_{k+1}^u\|^2 \\ &+\tfrac{1}{2}\sum_{k=0}^{n-1}\left\|\left((1-\mu)\mathbf{z}_{k+1}^l + \mu\mathbf{z}_{k+1}^u\right) - \mathbf{y}_{k+1}\right\|^2 \\ &\leq \tfrac{1}{2}\|\bar{\mathbf{y}} - \mathbf{y}^0\|^2 + \sum_{k=0}^{n-1}\tfrac{s_u}{k+1}\left(\Psi^* - \Psi(\mathbf{z}_{k+1}^u)\right) \\ &\leq \tfrac{1}{2}\|\bar{\mathbf{y}} - \mathbf{y}^0\|^2 + s_u(1+\ln n)\left(\Psi^* - M_0\right),\end{aligned} \tag{28}$$

where the last inequality follows from the assumption that $\inf \Psi \geq M_0$. By Lemma 5, we have

$$\begin{aligned}\|\mathbf{y}_{k+1} - \mathbf{y}_k\|^2 \leq\ &\|\mathbf{y}_k - \mathbf{y}_{k-1}\|^2 + \tfrac{\mu}{(k+1)^2}\|\mathbf{y}_{k-1} - \mathbf{z}_k^u\|^2 \\ &+\tfrac{2(1-\mu)s_l c_\beta D M_f}{(k+1)^2} + \tfrac{2\mu s_u D M_F}{k(k+1)} + \tfrac{(1-\mu)c_\beta^2}{\underline{\beta}^2(k+1)^4}\|\mathbf{y}_{k-1} - \mathbf{z}_k^l\|^2.\end{aligned} \tag{29}$$

and thus

$$\begin{aligned}&n\|\mathbf{y}_n - \mathbf{y}_{n-1}\|^2 \leq \sum_{k=0}^{n-1}\|\mathbf{y}_{k+1} - \mathbf{y}_k\|^2 + \mu\sum_{k=0}^{n-1}\|\mathbf{y}_k - \mathbf{z}_{k+1}^u\|^2 \\ &+\tfrac{(1-\mu)c_\beta^2}{\underline{\beta}^2}\sum_{k=0}^{n-1}\|\mathbf{y}_k - \mathbf{z}_{k+1}^l\|^2 + 2\mu s_u D M_F + 2(1-\mu)s_l c_\beta D M_f.\end{aligned} \tag{30}$$

Then it follows from Eq. (28) and Eq. (30) that

$$\begin{aligned}&\min\left\{(1-s_l L_f), (1-s_u L_F), 1\right\} n\|\mathbf{y}_n - \mathbf{y}_{n-1}\|^2 \\ &\leq \min\left\{(1-s_l L_f), (1-s_u L_F), 1\right\}\sum_{k=0}^{n-1}\|\mathbf{y}_{k+1} - \mathbf{y}_k\|^2 \\ &+\tfrac{c_\beta^2}{\underline{\beta}^2}(1-\mu)(1-s_l L_f)\sum_{k=0}^{n-1}\|\mathbf{y}_k - \mathbf{z}_{k+1}^l\|^2 + 2\mu s_u D M_F \\ &+\mu(1-s_u L_F)\sum_{k=0}^{n-1}\|\mathbf{y}_k - \mathbf{z}_{k+1}^u\|^2 + 2(1-\mu)s_l c_\beta D M_f \\ &\leq (2 + \tfrac{c_\beta^2}{\underline{\beta}^2})(1-\mu)(1-s_l L_f)\sum_{k=0}^{n-1}\|\mathbf{y}_k - \mathbf{z}_{k+1}^l\|^2 \\ &+3\mu(1-s_u L_F)\sum_{k=0}^{n-1}\|\mathbf{y}_k - \mathbf{z}_{k+1}^u\|^2 + 2(1-\mu)s_l c_\beta D M_f \\ &+2\sum_{k=0}^{n-1}\left\|\left((1-\mu)\mathbf{z}_{k+1}^l + \mu\mathbf{z}_{k+1}^u\right) - \mathbf{y}_{k+1}\right\|^2 + 2\mu s_u D M_F \\ &\leq \max\{2 + \tfrac{c_\beta^2}{\underline{\beta}^2}, 3\}\left(\|\bar{\mathbf{y}} - \mathbf{y}_0\|^2 + 2s_u(1+\ln n)\left(\Psi^* - M_0\right)\right) \\ &+2\mu s_u D M_F + 2(1-\mu)s_l c_\beta D M_f,\end{aligned}$$

where the second inequality comes from $\mathbf{y}_k - \mathbf{y}_{k+1} = (1-\mu)(\mathbf{y}_k - \mathbf{z}_{k+1}^l) + \mu(\mathbf{y}_k - \mathbf{z}_{k+1}^u) + (1-\mu)\mathbf{z}_{k+1}^l + \mu\mathbf{z}_{k+1}^u - \mathbf{y}_{k+1}$ and the convexity of $\|\cdot\|^2$. Combining with $\|\bar{\mathbf{y}} - \mathbf{y}_0\| \leq D$, we have

$$\|\mathbf{y}_n - \mathbf{y}_{n-1}\|^2 \leq \frac{C_1(1+\ln n)}{n}, \tag{31}$$

where $C_1 := (\max\{2 + c_\beta^2/\underline{\beta}^2, 3\}(D^2 + 2s_u(\Psi^* - M_0)) + 2\mu s_u D M_F + 2(1-\mu)s_l c_\beta D M_f)/\min\{(1-s_l L_f), (1-s_u L_F), 1\}$. Next, by Lemma 4, we have for all k,

$$\|\mathbf{z}_{k+1}^l - \mathbf{y}_k\| \leq \|\mathbf{z}_{k+1}^l - \bar{\mathbf{y}}\| + \|\mathbf{y}_k - \bar{\mathbf{y}}\| \leq 2\|\mathbf{y}_k - \bar{\mathbf{y}}\| \leq 2D.$$

Then, we have

$$\begin{aligned}&\tfrac{1}{\beta_k^2}\|\mathbf{z}_{k+1}^l - \mathbf{y}_k\|^2 \\ &\leq \tfrac{2}{\beta_{k-1}}\|\mathbf{z}_k^l - \mathbf{y}_{k-1}\| \|\tfrac{\mathbf{z}_{k+1}^l - \mathbf{y}_k}{\beta_k} - \tfrac{\mathbf{z}_k^l - \mathbf{y}_{k-1}}{\beta_{k-1}}\| \\ &+\tfrac{1}{\beta_{k-1}^2}\|\mathbf{z}_k^l - \mathbf{y}_{k-1}\|^2 + \|\tfrac{\mathbf{z}_{k+1}^l - \mathbf{y}_k}{\beta_k} - \tfrac{\mathbf{z}_k^l - \mathbf{y}_{k-1}}{\beta_{k-1}}\|^2 \\ &\leq \tfrac{1}{\beta_{k-1}^2}\|\mathbf{z}_k^l - \mathbf{y}_{k-1}\|^2 + s_l^2\|\nabla\psi(\mathbf{y}_k) - \nabla\psi(\mathbf{y}_{k-1})\|^2 \\ &+\tfrac{4D}{\beta_{k-1}}\|\nabla\psi(\mathbf{y}_k) - \nabla\psi(\mathbf{y}_{k-1})\| \\ &\leq \tfrac{1}{\beta_{k-1}^2}\|\mathbf{z}_k^l - \mathbf{y}_{k-1}\|^2 + s_l^2 L_f^2\|\mathbf{y}_k - \mathbf{y}_{k-1}\|^2 + \tfrac{4DL_f}{\beta_{k-1}}\|\mathbf{y}_k - \mathbf{y}_{k-1}\| \\ &\leq \tfrac{1}{\beta_{k-1}^2}\|\mathbf{z}_k^l - \mathbf{y}_{k-1}\|^2 + (s_l^2 L_f^2 D + \tfrac{4DL_f}{\underline{\beta}})\|\mathbf{y}_k - \mathbf{y}_{k-1}\|,\end{aligned} \tag{32}$$

where the second inequality follows from the definition of $\mathbf{z}_k^l$ and the last inequality comes from $\|\mathbf{y}_k - \mathbf{y}_{k-1}\| \leq D$ and $\beta_k \geq \underline{\beta}$. This implies that for any $n > n_0 > 0$,

$$\begin{aligned}\tfrac{1}{\beta_n^2}\|\mathbf{z}_{n+1}^l - \mathbf{y}_n\|^2 \leq\ &(s_l^2 L_f^2 D + \tfrac{4DL_f}{\underline{\beta}})\sum_{k=n_0+1}^{n}\|\mathbf{y}_k - \mathbf{y}_{k-1}\| \\ &+\tfrac{1}{\beta_{n_0}^2}\|\mathbf{z}_{n_0+1}^l - \mathbf{y}_{n_0}\|^2.\end{aligned}$$

Thus, since $\beta_k \in [\underline{\beta}, 1]$, for any $m \geq 2$ and $n_0 = n - m + 1$, the following holds

$$\begin{aligned}
&m\underline{\beta}^2\|\mathbf{z}_{n+1}^l - \mathbf{y}_n\|^2 \\
&\leq (s_l^2 L_f^2 D + \tfrac{4DL_f}{\underline{\beta}}) \sum_{k=n_0+1}^{n} (k-n_0)\|\mathbf{y}_k - \mathbf{y}_{k-1}\| \\
&+ \sum_{k=n_0}^{n} \|\mathbf{z}_{k+1}^l - \mathbf{y}_k\|^2 \\
&\leq \sum_{k=n_0}^{n} \|\mathbf{z}_{k+1}^l - \mathbf{y}_k\|^2 + (s_l^2 L_f^2 D + \tfrac{4DL_f}{\underline{\beta}})\sqrt{C_1}\tfrac{m(m-1)}{2}\tfrac{\sqrt{(1+\ln n_0)}}{\sqrt{n_0}},
\end{aligned} \tag{33}$$

where the last inequality follows from Eq. (31) that $\|\mathbf{y}_k - \mathbf{y}_{k-1}\|^2 \leq \frac{C_1(1+\ln n_0)}{n_0}$ for all $k \geq n_0$, and it can be easily verified that the above inequality holds when $m = 1$. By Eq. (28), we have

$$\begin{aligned}
&\tfrac{1}{2}(1-\mu)(1-s_l L_f)\sum_{k=0}^{n-1}\|\mathbf{y}_k - \mathbf{z}_{k+1}^l\|^2 \\
&\leq \tfrac{1}{2}\|\bar{\mathbf{y}} - \mathbf{y}_0\|^2 + s_u(1+\ln n)\,(\Psi^* - M_0)\,.
\end{aligned}$$

Then, for any n, let m be the smallest integer such that $m \geq n^{\frac{1}{4}}$ and let $n_0 = n - m + 1$, combining the above inequality with Eq. (33), we have

$$\begin{aligned}
&\tfrac{\|\bar{\mathbf{y}}-\mathbf{y}_0\|^2 + 2s_u(1+\ln n)(\Psi^* - M_0)}{(1-\mu)(1-s_l L_f)} \\
&\geq \sum_{k=n_0}^{n}\|\mathbf{y}_k - \mathbf{z}_{k+1}^l\|^2 \\
&\geq m\underline{\beta}^2\|\mathbf{y}_n - \mathbf{z}_{n+1}^l\|^2 - C_2\tfrac{m(m-1)}{2}\tfrac{\sqrt{(1+\ln n_0)}}{\sqrt{n_0}},
\end{aligned}$$

where $C_2 := (s_l^2 L_\psi^2 D + \frac{4DL_f}{\underline{\beta}})\sqrt{C_1}$.

Next, as $n^{\frac{1}{4}} + 1 \geq m \geq n^{\frac{1}{4}}$, and hence $n_0 \geq (m-1)^4 - m + 1$. Then $16n_0 - m^2(m-1)^2 \geq (m-1)[(m-1)(3m-4)(5m-4) - 1] > 0$ when $m \geq 2$. Thus, when $n \geq 2$, we have $m \geq 2$ and $\frac{m(m-1)}{2}\frac{\sqrt{(1+\ln n_0)}}{\sqrt{n_0}} \leq 2\sqrt{(1+\ln n_0)}$. Then, let $C_3 := \frac{D^2 + 2s_u(\Psi^* - M_0)}{(1-\mu)(1-s_l L_f)}$, we have for any $n \geq 2$,

$$\begin{aligned}
\|\mathbf{y}_n - \mathbf{z}_{n+1}^l\|^2 &\leq \frac{1}{m\underline{\beta}^2}\left(C_3(1+\ln n) + 2C_2\sqrt{(1+\ln n_0)}\right) \\
&\leq \frac{(2C_2 + C_3)}{\underline{\beta}^2}\frac{1+\ln n}{n^{\frac{1}{4}}},
\end{aligned}$$

where the last inequality follows from $\sqrt{1+\ln n_0} \leq 1 + \ln n$ and $m \geq n^{\frac{1}{4}}$. By the convexity of ψ, and $\mathbf{y}_n - \mathbf{z}_{n+1}^l = \beta_n s_l \nabla\psi(\mathbf{y}_n)$, we have

$$\begin{aligned}
\psi(\mathbf{y}_n) &\leq \psi(\bar{\mathbf{y}}) + \langle\nabla\psi(\mathbf{y}_n), \mathbf{y}_n - \bar{\mathbf{y}}\rangle \\
&= \min\psi + \frac{1}{\beta_n s_l}\langle\mathbf{y}_n - \mathbf{z}_{n+1}^l, \mathbf{y}_n - \bar{\mathbf{y}}\rangle \\
&\leq \min\psi + \frac{1}{\beta_n s_l}\|\mathbf{y}_n - \mathbf{z}_{n+1}^l\|\|\mathbf{y}_n - \bar{\mathbf{y}}\| \\
&\leq \min\psi + \frac{D}{\underline{\beta}^2 s_l}\sqrt{(2C_2 + C_3)\frac{1+\ln n}{n^{\frac{1}{4}}}}.
\end{aligned}$$

□

4 Discussion and Extension

This section provides a comparison with the existing LLS scheme by a high dimension counter-example in Section 4.1 and develops an one-stage extension scheme in Section 4.2.

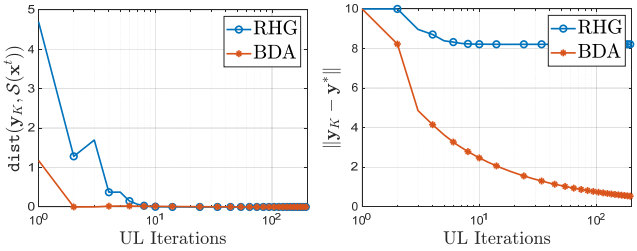

Fig. 1. An evaluation of the convergence behavior about the LL variable $\mathbf{y}$. We compare our BDA with gradient-based BLO algorithm (i.e., RHG). We set the initial points $(\mathbf{x}, \mathbf{y}) = (0, 0)$, $n = 50$ and $K = 20$. $\mathbf{x}^t$ denotes the UL variable at the t-th UL iterations.

4.1 Comparing with Existing LLS Theories

As aforementioned, a number of gradient-based methods have been proposed to solve BLO 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 BLO in Eqs. (1)-(2), they actually solve the following simplification:

$$\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x}, \mathbf{y}),\ s.t.\ \mathbf{y} = \arg\min_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}), \tag{34}$$

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 $\{\mathbf{y}_k(\mathbf{x})\}_{k=0}^K$ parameterized by $\mathbf{x}$ can be generated, e.g.,

$$\mathbf{y}_{k+1}(\mathbf{x}) = \mathbf{y}_k(\mathbf{x}) - s_l\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_k(\mathbf{x})),\ k = 0, \cdots, K-1, \tag{35}$$

where $s_l > 0$ is an appropriately chosen step size. Then by considering $\mathbf{y}_K(\mathbf{x})$ (i.e., the output of Eq. (35) 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. (35), we can calculate the derivative of $F(\mathbf{x}, \mathbf{y}_K(\mathbf{x}))$ (w.r.t. $\mathbf{x}$) to optimize Eq. (34) by automatic differentiation techniques [6], [36]. We summarize the updating schemes of UL and LL subproblem in Table 2.

The UL objective F is indeed a function of both the UL variable $\mathbf{x}$ and the LL variable $\mathbf{y}$. Conventional gradient-based bi-level methods (Eq. (35)) 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.

Example 1. *(Counter-Example) With $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^n$, we consider the following BLO problem:*

$$\begin{aligned}
&\min_{\mathbf{x}\in X}\|\mathbf{x} - \mathbf{z}\|^4 + \|\mathbf{y} - \mathbf{e}\|^4, \\
&s.t.\ (\mathbf{y}, \mathbf{z}) \in \arg\min_{\mathbf{y}\in\mathbb{R}^n, \mathbf{z}\in\mathbb{R}^n} \tfrac{1}{2}\|\mathbf{y}\|^2 - \mathbf{x}^\top\mathbf{y},
\end{aligned} \tag{36}$$

where $X = [-100, 100] \times \cdots [-100, 100] \subset \mathbb{R}^n$, $\mathbf{e}$ denotes the vector whose elements are all equal to 1. By simple calculation, we

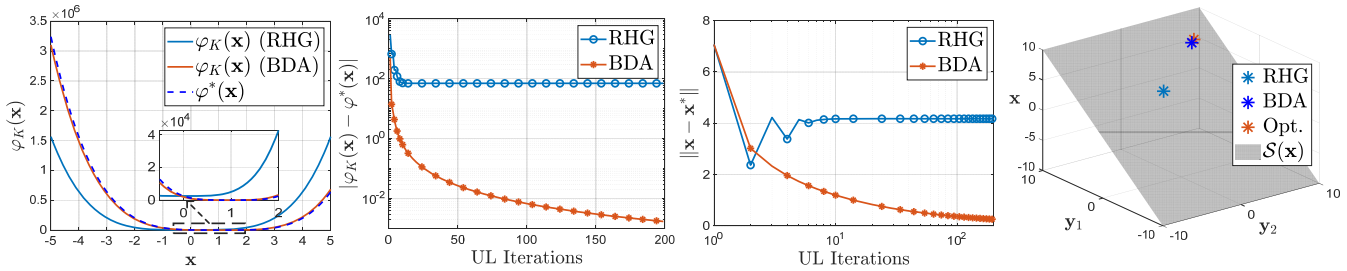

Fig. 2. Illustrating the convergence behavior of gradient-based BLO algorithms about the UL variable $\mathbf{x}$. We set the initial points $(\mathbf{x}, \mathbf{y}) = (0, 0)$, $n = 50$ and $K = 20$. In the first subfigure, $\varphi_K(\mathbf{x})$ and $\varphi^*(\mathbf{x})$ denote the UL objective with LL computational solution $\mathbf{y}_K$ and the optimal LL solution $\mathbf{y}^*$ respectively. The second and third subfigures respectively show the errors of UL objective (i.e., $|\varphi_K(\mathbf{x}) - \varphi^*(\mathbf{x})|$) and UL variable (i.e., $\|\mathbf{x} - \mathbf{x}^*\|$). The last subfigure illustrates the relationship among Optimal solution (short for "Opt.", the red star) and the iteration solutions of RHG and BDA.

know that the unique optimal solution of Eq. (36) *is* $\mathbf{x}^* = \mathbf{y}^* = \mathbf{z}^* = \mathbf{e}$. *However, if adopting the existing gradient-based scheme in Eq.* (35) *with initialization* $(\mathbf{y}_0, \mathbf{z}_0) = (0, 0)$ *and varying step size* $s_l^k \in (0, 1)$, *we have that* $\mathbf{y}_K = (1 - \prod_{k=0}^{K-1}(1 - s_l^k))\mathbf{x}$ *and* $\mathbf{z}_K = 0$. *Then the approximated problem of Eq.* (36) *amounts to*

$$\min_{\mathbf{x}\in X} F(\mathbf{x}, \mathbf{y}_K, \mathbf{z}_K) = \|\mathbf{x}\|^4 + \|(1 - \prod_{k=0}^{K-1}(1 - s_l^k))\mathbf{x} - \mathbf{e}\|^4.$$

Consider sequence

$$\mathbf{x}_K^* = \arg\min_{\mathbf{x}\in X} F(\mathbf{x}, \mathbf{y}_K, \mathbf{z}_K),$$

it follows from the first-order optimality condition that,

$$0 = 4\|\mathbf{x}_K^*\|^2\mathbf{x}_K^* + 4a_K\|a_K\mathbf{x}_K^* - \mathbf{e}\|^2(a_K\mathbf{x}_K^* - \mathbf{e}), \quad (37)$$

where $a_K = (1 - \prod_{k=0}^{K-1}(1 - s_l^k))$. *Then, if sequence* $\{\mathbf{x}_K^*\}$ *converge to a limit point* $\mathbf{e}$, *and since* $\{a_K\}$ *is bounded, there exist subsequences* $\{\mathbf{x}_{K_\ell}^*\} \subset \{\mathbf{x}_K^*\}$ *and* $\{a_{K_\ell}\} \subset \{a_K\}$ *such that* $\{\mathbf{x}_{K_\ell}^*\} \to \mathbf{e}$ *and* $\{a_{K_\ell}\} \to \bar{a}$. *By considering subsequences* $\{\mathbf{x}_{K_\ell}^*\}$ *and* $\{a_{K_\ell}\}$ *in* (37) *and taking* $K_\ell \to \infty$, *we should have*

$$\begin{aligned} 0 &= \|\mathbf{e}\|^2\mathbf{e} + \bar{a}\|\bar{a}\mathbf{e} - \mathbf{e}\|^2(\bar{a}\mathbf{e} - \mathbf{e}), \\ &= [1 + (\bar{a} - 1)^3\bar{a}]\|\mathbf{e}\|^2\mathbf{e} \end{aligned}$$

and thus

$$0 = 1 + (\bar{a} - 1)^3\bar{a}.$$

However, since $a_K = (1 - \prod_{k=0}^{K-1}(1 - s_l^k)) \in [0, 1]$, *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 subseuqnce of* $\{\mathbf{x}_K^*\}$ *cannot converge to the true solution (i.e.,* $\mathbf{x}^* = \mathbf{e}$*).*

Remark 1. *Actually, even with strongly convex UL objective w.r.t. LL variable* $\mathbf{y}$, *the existing bi-level based methods still may fail to reach an optimal solution. For example, with* $\mathbf{x} \in [-100, 100]$ *and* $\mathbf{y} \in \mathbb{R}^2$, *we consider the following BLO problem:*

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

By simple calculation, we know that the unique optimal solution of Eq. (1) *is* $\mathbf{x}^* = 1, \mathbf{y}^* = (1, 1)$. *However, if adopting the existing gradient-based scheme in Eq.* (35) *with initialization*

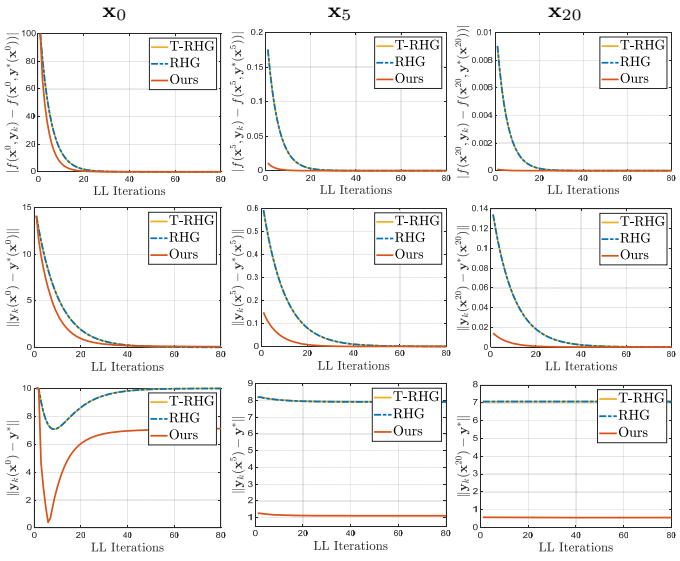

Fig. 3. LL iteration curves of gradient-based BLO algorithms (T-RHG, RHG and Ours) under three fixed $\mathbf{x}$ (i.e., $\mathbf{x}^0$, $\mathbf{x}^5$, $\mathbf{x}^{20}$). The objective errors (i.e., $|f(\mathbf{x}^0, \mathbf{y}_k) - f(\mathbf{x}^0, \mathbf{y}_*(\mathbf{x}^0))|$), variable errors with $\mathbf{y}_*(\mathbf{x})$ and $\mathbf{y}^*$ (i.e., $\|\mathbf{y}_k - \mathbf{y}_*(\mathbf{x}^0)\|$ and $\|\mathbf{y}_k - \mathbf{y}^*\|$) are respectively plotted from the top to the bottom row. $\mathbf{y}_*(\mathbf{x})$ and $\mathbf{y}^*$ denote the optimal solution with and without relationship about $\mathbf{x}$.

$\mathbf{y}_0 = (0, 0)$ *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* $[\mathbf{y}_K]_2 = 0$. *By defining* $\varphi_K(\mathbf{x}) = F(\mathbf{x}, \mathbf{y}_K)$, *we have* $\mathbf{x}_K^* = \frac{(1 - \prod_{k=0}^{K-1}(1 - s_l^k))}{1 + (1 - \prod_{k=0}^{K-1}(1 - s_l^k))^2}$. *It is easy to check that* $\mathbf{x}_K^* \leq \frac{1}{2}$. *So* $\mathbf{x}_K^*$ *cannot converge to the true solution (i.e.,* $\mathbf{x}^* = 1$*).*

Remark 2. *In applications, to achieve the LLS, people sometimes add a strongly convex regularization term to the LL subproblem. We must clarify that this strategy is only heuristic, which usually causes unpredictable large deviation from the true solution.*

Indeed, even the strongly convex regularization is set to be vanishing, such an appoximation procedure cannot guarantee any convergence to the true solution. We will take the counter-example in Remark 1 again for illustration. Specifically, we introduce a quadratic term $1/2\varepsilon\|\mathbf{y}_2\|^2$ *to the LL subproblem*

$$\min_{\mathbf{y}\in\mathbb{R}^2} \frac{1}{2}\|\mathbf{y}_1\|^2 + \frac{1}{2}\varepsilon\|\mathbf{y}_2\|^2 - \mathbf{x}^\top\mathbf{y}_1.$$

Apparently, the LL objective becomes strongly convex. But it can be checked that the optimal solution to such bilevel problem with

regularized LL

$$\min_{\mathbf{x}\in[-100,100]} \tfrac{1}{2}\|\mathbf{x}-\mathbf{y}_2\|^2+\tfrac{1}{2}\|\mathbf{y}_1-1\|^2,$$
$$s.t.\ \mathbf{y}\in\arg\min_{\mathbf{y}\in\mathbb{R}^2}\tfrac{1}{2}\|\mathbf{y}_1\|^2+\tfrac{1}{2}\varepsilon\|\mathbf{y}_2\|^2-\mathbf{x}^\top\mathbf{y}_1,$$

becomes $\mathbf{x}^*(\varepsilon)=\frac{1}{2}$, $\mathbf{y}_1^*(\varepsilon)=\frac{1}{2}$, $\mathbf{y}_2^*(\varepsilon)=0$ *which is obviously no longer the true solution to the original bilevel model. Moreover, even with* ε *tending* 0, *unfortunately,* $\mathbf{x}^*(\varepsilon)$, $\mathbf{y}^*(\epsilon)$ *and* $\mathbf{y}_2^*(\epsilon)$ *still fail to converge to the true solution* $(1,1,1)$.

To demonstrate the convergence behavior of our BDA and the most popular bi-level method (i.e., RHG [6], [3]), we first illustrate the optimization procedure of LL variable (i.e., $\mathbf{y}_K$) in Figure 1. As can be observed that the LL variable $\mathbf{y}_K$ can converge to the LL solution set $\mathcal{S}(\mathbf{x}^t)$ for both RHG and our BDA in the left subfigure. But, the LL variable of our method can find the optimal point, i.e., $\mathbf{y}^*$, while RHG cannot. Note that we set the dimension $n=50$.

In Figure 2, comparing with RHG, we then demonstrate the optimization procedure of UL variable (i.e., $\mathbf{x}$). In the first subfigure, under fixed LL iterative solution $\mathbf{y}_K$, the UL objective $\varphi_K(\mathbf{x})$ illustrates that our BDA can efficiently fit the optimal objective function (i.e., $\varphi^*(\mathbf{x})$) for any UL variable. To further demonstrate the convergence behavior, we plotted the errors of the UL objective (i.e., $|\varphi_K(\mathbf{x})-\varphi(\mathbf{x})|$) and variable (i.e., $\|\mathbf{x}-\mathbf{x}^*\|$) in the second and third subfigures. With the above illustration, we summarize the relationship of Optimal solution (short for "Opt.", the red one in the last subfigure) with the iterative solutions of RHG and BDA in the last subfigure. Thus, we conclude that our BDA can find the optimal point, while RHG converge to a non-optimal point in $\mathcal{S}(\mathbf{x})$. .

4.2 One-stage BDA

Multi-step of the LL iteration modules $\mathcal{T}_k$ 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 [12]. By setting $K=1$ in Eq. (8), the algorithm reads as

$$\mathbf{y}_1(\mathbf{x})=\mathcal{T}_1(\mathbf{x},\mathbf{y}_0)=\mathtt{Proj}_{\mathcal{Y}}\left(\mathbf{y}_0-\rho\partial_{\mathbf{y}}\phi(\mathbf{x},\mathbf{y}_0)\right), \tag{38}$$

where $\phi(\mathbf{x},\mathbf{y}_0)=\alpha F(\mathbf{x},\mathbf{y}_0)+\beta f(\mathbf{x},\mathbf{y}_0)$ and $\alpha,\beta\in(0,1]$ denote the aggregation parameters. Indeed, if $(\mathbf{y}_0-\rho\partial_{\mathbf{y}}\phi(\mathbf{x},\mathbf{y}_0))\in\mathcal{Y}$, with this one-stage simplification, we can simplify the back-propagation calculation with the following finite difference approximation

$$\begin{aligned}\tfrac{d\varphi_1(\mathbf{x})}{d\mathbf{x}}&=\tfrac{\partial F(\mathbf{x},\mathbf{y}_1)}{\partial\mathbf{x}}+\tfrac{\partial F(\mathbf{x},\mathbf{y}_1)}{\partial\mathbf{y}_1}\tfrac{d\mathbf{y}_1}{d\mathbf{x}}\\&\approx\tfrac{\partial F(\mathbf{x},\mathbf{y}_1)}{\partial\mathbf{x}}-\rho\tfrac{\partial_{\mathbf{x}}\phi(\mathbf{x},\mathbf{h}_0^+)-\partial_{\mathbf{x}}\phi(\mathbf{x},\mathbf{h}_0^-)}{2\epsilon},\end{aligned}$$

where $\mathbf{h}_0^{\pm}=\mathbf{y}_0\pm\epsilon\partial F(\mathbf{x},\mathbf{y}_1)/\partial\mathbf{y}_1$ and $\partial_{\mathbf{x}}\phi(\mathbf{x},\mathbf{y})=\alpha\partial_{\mathbf{x}}F(\mathbf{x},\mathbf{y})+\beta\partial_{\mathbf{x}}f(\mathbf{x},\mathbf{y})$. Since $\mathcal{Y}$ can be a big interval, this case (i.e., $(\mathbf{y}_0-\rho\partial_{\mathbf{y}}\phi(\mathbf{x},\mathbf{y}_0))\in\mathcal{Y}$) is often satisfied in general. If $(\mathbf{y}_0-\rho\partial_{\mathbf{y}}\phi(\mathbf{x},\mathbf{y}_0))\notin\mathcal{Y}$, the above back-propagation can be calculated by the following form

$$\begin{aligned}\tfrac{d\varphi_1(\mathbf{x})}{d\mathbf{x}}\approx&\tfrac{\partial F(\mathbf{x},\mathbf{y}_1)}{\partial\mathbf{x}}\\&-\tfrac{\partial_{\mathbf{x}}\phi(\mathbf{x},\mathbf{h}_0^{++})-\partial_{\mathbf{x}}\phi(\mathbf{x},\mathbf{h}_0^{-+})-\left(\partial_{\mathbf{x}}\phi(\mathbf{x},\mathbf{h}_0^{+-})-\partial_{\mathbf{x}}\phi(\mathbf{x},\mathbf{h}_0^{--})\right)}{4\epsilon^{1+\frac{1}{2}}},\end{aligned}$$

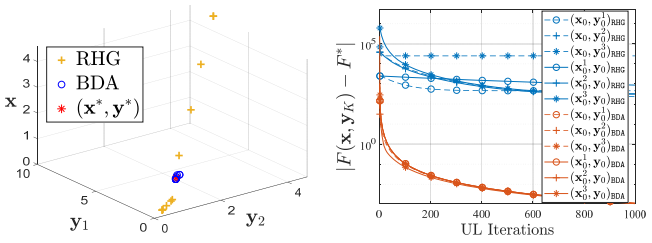

Fig. 4. Comparisons of BDA with RHG on ten different initial points. We set the dimensional $n=50$ and $K=20$. The left subfigure show the iteration solution of different initial points. We select five different initial points and show the UL objective behavior of BDA and RHG on the right subfigure.

where $\mathbf{h}_0^{++}=\mathbf{y}_0\pm\epsilon\mathtt{Proj}_{\mathcal{Y}}\left(\mathbf{z}_0+\epsilon^{1/2}\partial F(\mathbf{x},\mathbf{y}_1)/\partial\mathbf{y}\right)$ and $\mathbf{h}_0^{\pm-}=\mathbf{y}_0\pm\epsilon\mathtt{Proj}_{\mathcal{Y}}\left(\mathbf{z}_0-\epsilon^{1/2}\partial F(\mathbf{x},\mathbf{y}_1)/\partial\mathbf{y}\right)$ with $\mathbf{z}_0=\mathbf{y}_0-\rho\partial_{\mathbf{y}}\phi(\mathbf{x},\mathbf{y}_0)$.

5 Experimental Results

This section first verify the numerical results and then evaluate the performance of our proposed method on different problems. We conducted these experiments on a PC with In tel Core i7-7700 CPU (3.6 GHz), 32GB RAM and an NVIDIA GeForce RTX 2060 6GB GPU.

5.1 Numerical Evaluations

Our numerical results are investigated based on the synthetic BLO described in Subsection 4.1, i.e., Counter Example in Eq. (36). As stated in Subsection 4.1, this deterministic bi-level formulation satisfies all the assumptions required in Section 3, but it cannot meet the LLS condition considered in [29], [6], [3], [28], [4].

To show the influence of the LL iterations (i.e., K) on different methods, we first plotted the convergence behaviors (i.e., $|f(\mathbf{x}^t,\mathbf{y}_k)-f(\mathbf{x}^t,\mathbf{y}^*(\mathbf{x}^t))|$, $\|\mathbf{y}_k(\mathbf{x}^t)-\mathbf{y}^*(\mathbf{x}^t)\|$ and $\|\mathbf{y}_k(\mathbf{x}^t)-\mathbf{y}^*\|$ with $t=0,5,20$) under different given $\mathbf{x}$ (i.e., $\mathbf{x}^0$, $\mathbf{x}^5$, $\mathbf{x}^{20}$) in Figure 3. This figure compare our BDA with the most popular bi-level based methods (i.e., T-RHG and RHG). Note that $t=0,5,20$ are the UL iteration steps during the operation process. From the first and second row of Figure 3, we observed that with fixed UL variable $\mathbf{x}$, the results of RHG and BDA converge to the optimal solution with corresponding given $\mathbf{x}^t$. The third row of Figure 3 plotted the distance between the current iteration step and the optimal solution $\mathbf{y}^*$. As can be seen, after a few UL iteration steps (i.e., $t\geq 5$), BDA is close to the optimal solution $\mathbf{y}^*$ while RHG and T-RHG cannot. In the above figures, we set $\alpha_k=0.5/k$, $k=1,\cdots,K$.

Figure 4 plotted numerical results of the proposed BDA and RHG [6], [3] with ten different initialization points. We considered different numerical metrics, such as the relationship of $(\mathbf{x},\mathbf{y})$ with optimal solution $(\mathbf{x}^*,\mathbf{y}^*)$ and the distance between $F(\mathbf{x},\mathbf{y}_K)$ and F^* (i.e., $|F(\mathbf{x},\mathbf{y}_K)-F^*|$), for evaluations. It needs to be noted that we select five different initial points to show the performance of $|F(\mathbf{x},\mathbf{y}_K)-F^*|$. 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. (36) is not a singleton, which does not

TABLE 3
Data hyper-cleaning accuracy of the compared methods on two different datasets, i.e., MNIST [37] and Fashion MNIST [38]. 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. "Test Acc." and "Val. Acc." denote the averaged accuracy of test and validation sets, respectively.

Methods	MNIST				Fashion MNIST			
	Val. Acc.	Test Acc.	F1-Score	Time(s)	Val. Acc.	Test Acc.	F1-Score	Time(s)
IHG	86.98	87.69	87.62	12.14 ± 0.73	82.66	83.82	83.63	10.92 ± 0.75
RHG	88.08	88.30	88.20	3.72 ± 0.01	85.12	86.14	86.04	1.09 ± 0.01
T-RHG	88.30	86.16	88.10	2.49 ± 0.07	85.12	86.06	86.07	0.63 ± 0.01
O-BDA	**88.84**	88.45	88.37	2.89 ± 0.01	**86.34**	86.16	86.05	0.66 ± 0.01
BDA	88.26	**88.47**	**88.42**	7.82 ± 0.12	85.28	**86.26**	**86.17**	1.91 ± 0.01

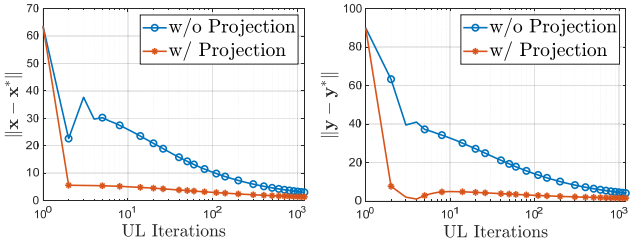

Fig. 5. Comparing of our BDA under different settings, i.e., with and without projection operator (namely w/ Projection and w/o Projection). We set $n = 50$ and $K = 20$.

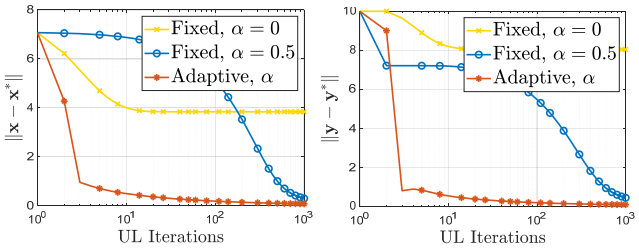

Fig. 6. The iteration curves of the developed BDA with different α settings (i.e., with Fixed $\alpha = 0$, $\alpha = 0.5$ and Adaptive $\alpha = 0.5/k$). We set $n = 50$ and $K = 20$.

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. (8), we report in Figure 5 the results (i.e., $\|\mathbf{x} - \mathbf{x}^*\|$ and $\|\mathbf{y} - \mathbf{y}^*\|$) 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 $\mathcal{Y}$. 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 α_k. 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 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

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 [6], [28]) 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 softmax regression (with parameters $\mathbf{y}$) as our classifier and introduce hyper-parameters $\mathbf{x}$ to weight samples for training. We 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, $\ell(\mathbf{y}; \mathbf{u}_i, \mathbf{v}_i)$ means 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. 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.,

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

In particular, the UL and LL objective F and f w.r.t. $\mathbf{y}$ 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) [26], RHG and Truncated RHG (T-RHG) [28]. We first conduct the experiment on two datasets (MNIST dataset [37] and Fashion MNIST dataset [38]) that each with 5000 training examples (i.e., $\mathcal{D}_{\mathtt{tr}}$), 5000 validation examples (i.e., $\mathcal{D}_{\mathtt{val}}$) and a test set with the remaining 60000 samples. We randomly chose 2500 training samples from $\mathcal{D}_{\mathtt{tr}}$ 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 3, the developed method perform the best both on MNIST and Fashion MNIST dataset. Besides, the developed O-BDA still perform better when comparing with the existing bi-level based methods.

TABLE 4
Averaged accuracy scores ± standard deviation of various methods (model-based methods and gradient-based bi-level methods) on few-shot classification classification problems (1-shot and 5-shot, i.e., $M = 1, 5$, $N = 5, 20, 30, 40$) on Omniglot.

Method	5-way		20-way		30-way		40-way	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MAML	98.70 ± 0.40%	**99.91** ± 0.10%	95.80 ± 0.30%	98.90 ± 0.20%	86.86 ± 0.49%	96.86 ± 0.19%	85.98 ± 0.45%	94.46 ± 0.13%
Meta-SGD	97.97 ± 0.70%	98.96 ± 0.20%	93.98 ± 0.43%	98.42 ± 0.11%	89.91 ± 0.04%	96.21 ±0.15%	87.39 ± 0.43%	95.10 ± 0.15%
Reptile	97.68 ± 0.04%	99.48 ± 0.06%	89.43 ± 0.14%	97.12 ± 0.32%	85.40 ± 0.30%	95.28 ± 0.30%	82.50 ± 0.30%	92.79 ± 0.33%
iMAML,GD	**99.16** ± 0.35%	99.67 ± 0.12%	94.46 ± 0.42%	98.69 ± 0.10%	89.52 ± 0.20%	96.51 ± 0.08%	87.28 ± 0.21%	95.27 ± 0.08%
RHG	98.64 ± 0.21%	99.58 ± 0.12%	96.13 ± 0.20%	99.09 ± 0.08%	93.92 ± 0.18%	98.43 ± 0.08%	90.78 ± 0.20%	96.79 ± 0.10%
T-RHG	98.74 ± 0.21%	99.71 ± 0.07%	95.82 ± 0.20%	98.95 ± 0.07%	94.02 ± 0.18%	98.39 ± 0.07%	90.73 ± 0.20%	96.79 ± 0.10%
BDA	99.04 ± 0.18%	99.74 ± 0.05%	**96.50** ± 0.16%	**99.19** ± 0.07%	**94.37** ± 0.18%	**98.53** ± 0.07%	**92.49** ± 0.18%	**97.12** ± 0.09 %

TABLE 5
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. ± Var. (Acc.)" denotes the averaged accuracy and the corresponding variance.

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

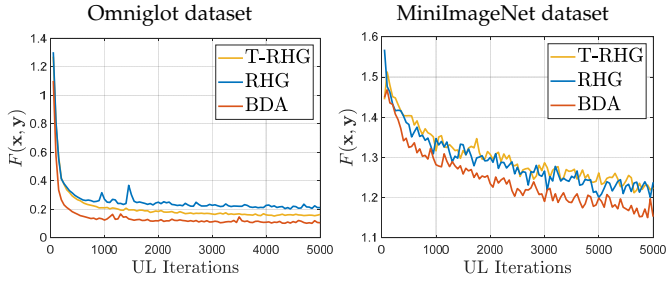

Fig. 7. Illustrating the validation loss (i.e., UL objectives $F(\mathbf{x}, \mathbf{y})$) for three bi-level based methods on few-shot classification task.

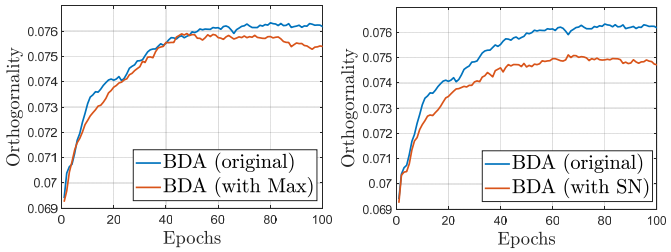

Fig. 8. Evaluating the orthogonality under different constraints (i.e., with Max-norm regularization and Spectral Normalization (SN for short)) on few-shot application.

5.3 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 [39], [40], where each task is to discriminate N separate classes 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 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 $\mathbf{x}$) 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 Ominglot [41] and MiniImageNet [39] benchmarks. We compared our BDA to several state-of-the-art approaches, such as MAML [29], Meta-SGD [42], Reptile [30], iMAML [4], RHG, and T-RHG. As shown in Table 4, 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 over other methods. Besides, we evaluate the performance of BDA and bi-level based methods (i.e., RHG and T-RHG) on the more challenging MiniImageNet data set and the corresponding results are listed in Table 5. As shown in the second column of Table 5 that the developed BDA perform better than RHG and T-RHG. The rightmost two columns demonstrate that BDA needed the fewest iterations to achieve almost the same accuracy ($\approx$ 44%). The corresponding validation loss on Omniglot and MiniImageNet about 5-way 1-shot are shown in Figure 7.

Moreover, to evaluate the effectiveness of the projection operator, we conduct an experiment evaluating orthogonal features of the network with two different strategies (i.e., max-norm regularization and spectral normalization). Note that we compute the orthogonality following [43]. As shown in Figure 8, BDA with both Max and SN training schemes show the lower orthogonal sum. This experiment implies that the projection operator can help obtain a better network.

6 Conclusion

This work established a flexible descent aggregation framework with task-tailored iteration dynamics modules to solve bi-level tasks by formulating BLO in Eqs. (1)-(2) from the viewpoint of optimistic bi-level. Embedded with a specific gradient-aggregation-based iterative module, the developed method is applicable to capture a variety of learning tasks. Then, this work strictly proved the convergence of the developed framework without the LLS assumption and the strong convexity in the UL objective. Furthermore, we provided an one-stage technique to accelerate the back-propagation calculation. We constructed a counter-example to illustrate the advance of our method and explicitly indicates the importance of the LLS condition for existing gradient-based bi-level methods. 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, 61733002, 61672125, 61772105 and 11971220), LiaoNing Revitalization Talents Program (XLYC1807088), and the Fundamental Research Funds for the Central Universities. Foundation of Guangdong Province 2019A1515011152. This work was also supported by the General Research Fund 12302318 from Hong Kong Research Grants Council.

References

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

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

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

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

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

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

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

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

[9] 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.

[10] C. Gao, Y. Chen, S. Liu, Z. Tan, and S. Yan, "Adversarialnas: Adversarial neural architecture search for gans," in *IEEE CVPR*, 2020, pp. 5680–5689.

[11] Y. Tian, L. Shen, G. Su, Z. Li, and W. Liu, "Alphagan: Fully differentiable architecture search for generative adversarial networks," *CoRR, abs/2006.09134*, 2020.

[12] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," in *ICLR*, 2019.

[13] 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. 10 734–10 742.

[14] K. Nakai, T. Matsubara, and K. Uehara, "Att-darts: Differentiable neural architecture search for attention," in *IEEE IJCNN*, 2020, pp. 1–8.

[15] Y. Hu, X. Wu, and R. He, "Tf-nas: Rethinking three search freedoms of latency-constrained differentiable neural architecture search," in *ECCV*, 2020.

[16] 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.

[17] J. C. De los Reyes, C.-B. Schönlieb, and T. Valkonen, "Bilevel parameter learning for higher-order total variation regularisation models," *Journal of Mathematical Imaging and Vision*, vol. 57, no. 1, pp. 1–25, 2017.

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

[19] R. Liu, P. Mu, J. Chen, X. Fan, and Z. Luo, "Investigating task-driven latent feasibility for nonconvex image modeling," *IEEE TIP*, vol. 29, pp. 7629–7640, 2020.

[20] R. Liu, J. Liu, Z. Jiang, X. Fan, and Z. Luo, "A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion," *IEEE TIP*, vol. 30, pp. 1261–1274, 2021.

[21] G. Kunapuli, K. P. Bennett, J. Hu, and J.-S. Pang, "Classification model selection via bilevel programming," *Optimization Methods & Software*, vol. 23, no. 4, pp. 475–489, 2008.

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

[23] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo, "On the iteration complexity of hypergradient computation," in *ICML*, 2020.

[24] 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.

[25] 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.

[26] F. Pedregosa, "Hyperparameter optimization with approximate gradient," in *ICML*, 2016, pp. 737–746.

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

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

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

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

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

[32] 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.

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

[34] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011, vol. 408.

[35] A. Cabot, "Proximal point algorithm controlled by a slowly vanishing term: Applications to hierarchical minimization," *SIAM Journal on Optimization*, vol. 15, no. 2, pp. 555–572, 2005.

[36] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, "Automatic differentiation in machine learning: a survey," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 5595–5637, 2017.

[37] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[38] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *CoRR, abs/1708.07747*, 2017.

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

[40] 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.

[41] 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.
[42] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-sgd: Learning to learn quickly for few-shot learning," in *ICML*, 2018.
[43] A. Prakash, J. Storer, D. Florencio, and C. Zhang, "Repr: Improved training of convolutional filters," in *IEEE CVPR*, 2019, pp. 10 666–10 675.