Moreau Envelope for Nonconvex Bi-Level Optimization: A Single-Loop and Hessian-Free Solution Strategy

Risheng Liu [1 2] **Zhu Liu** [1] **Wei Yao** [3 4] **Shangzhi Zeng** [5 3] **Jin Zhang** [4 3]

Abstract

This work focuses on addressing two major challenges in the context of large-scale nonconvex Bi-Level Optimization (BLO) problems, which are increasingly applied in machine learning due to their ability to model nested structures. These challenges involve ensuring computational efficiency and providing theoretical guarantees. While recent advances in scalable BLO algorithms have primarily relied on lower-level convexity simplification, our work specifically tackles large-scale BLO problems involving nonconvexity in both the upper and lower levels. We simultaneously address computational and theoretical challenges by introducing an innovative single-loop gradient-based algorithm, utilizing the Moreau envelope-based reformulation, and providing non-asymptotic convergence analysis for general nonconvex BLO problems. Notably, our algorithm relies solely on first-order gradient information, enhancing its practicality and efficiency, especially for large-scale BLO learning tasks. We validate our approach's effectiveness through experiments on various synthetic problems, two typical hyper-parameter learning tasks, and a real-world neural architecture search application, collectively demonstrating its superior performance.

1. Introduction

Bi-Level Optimization (BLO) addresses the challenges posed by nested optimization structures that arise in a wide range of machine learning applications, such as hyper-parameter optimization (Pedregosa, 2016; Franceschi et al., 2018; Mackay et al., 2019), meta learning (Zügner & Günnemann, 2018; Rajeswaran et al., 2019; Ji et al., 2020a), neural architecture search (Liu et al., 2018; Chen et al., 2019; Elsken et al., 2020), etc. Refer to recent survey papers (Liu et al., 2021a; Zhang et al., 2023) for more applications of BLO in machine learning, computer vision and signal processing. The inherent nested nature gives rise to several difficulties and hurdles in effectively solving BLO problems. Over the past decade, a large number of BLO methods have emerged, with a primary emphasis on addressing BLO problems featuring strongly convex lower-level (LL) objective. The LL strong convexity assumption ensures the uniqueness of LL minimizer (a.k.a., LL Singleton), which simplifies both the optimization process and theoretical analysis, see, e.g., (Franceschi et al., 2018; Ghadimi & Wang, 2018; Grazzi et al., 2020; Ji et al., 2020b; Chen et al., 2021; Ji et al., 2022; Hong et al., 2023). To mitigate the restrictive LL Singleton condition, another line of research is dedicated to BLO with convex LL problems, which bring about several challenges such as the presence of multiple LL optimal solutions (a.k.a., Non-Singleton). This may hinder the adoption of implicit-based approaches that rely on the implicit function theorem. To address this concern, recent advances include: aggregation methods (Liu et al., 2020; Li et al., 2020; Liu et al., 2022; 2023b); difference-of-convex algorithm (Gao et al., 2022; Ye et al., 2023); primal-dual algorithms (Sow et al., 2022a); first-order penalty methods (Lu & Mei, 2023).

In this work, we study a BLO problem with a nonconvex LL problem:

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

where $S(x)$ denotes the set of optimal solutions for the LL problem given by

$$\min_{y\in Y} \varphi(x,y) := f(x,y) + g(x,y), \tag{2}$$

[1] School of Software Technology, Dalian University of Technology, Dalian, China [2] Pazhou Laboratory (Huangpu), Guangzhou, China. [3] National Center for Applied Mathematics Shenzhen, Southern University of Science and Technology, Shenzhen, China [4] Department of Mathematics, Southern University of Science and Technology, Shenzhen, China [5] Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia, Canada. Correspondence to: Jin Zhang <zhangj9@sustech.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Table 1. Comparison of our method MEHA with closely related works for addressing **nonconvex-nonconvex BLO** (IAPTT-GM (Liu et al., 2021c), BOME (Ye et al., 2022), V-PBGD (Shen & Chen, 2023), GALET (Xiao et al., 2023), SLM (Lu, 2023)). Different methods employ distinct stationary measures, so we do not delve into complexity comparison here. Below, PL Condition represents the Polyak-Łojasiewicz (PL) condition; L-Smooth means the Lipschitz continuous gradient condition; Bounded and Gradient-Bounded specify that $|F(x,y)| \leq C$ and $\|\nabla_y F(x,y)\| \leq C$ for all (x,y), respectively; F and f are UL and LL objectives, respectively.

Method	**Upper-Level Objective**	**Lower-Level Objective**	**Hessian-Free**	**Single-Loop**	**Non-Asymptotic**
IAPTT-GM	Smooth	L-Smooth & Compactness	✗	✗	✗
GALET	L-Smooth & Gradient-Bounded	PL Condition & L-Smooth	✗	✗	✓
BOME	L-Smooth & Bounded & Gradient-Bounded	PL Condition & L-Smooth	✓	✗	✓
V-PBGD	L-Smooth & Gradient-Bounded	PL Condition & L-Smooth	✓	✗	✓
SLM	L-Smooth & Compactness	PL Condition & L-Smooth	✓	✗	✓
MEHA (smooth case)	L-Smooth	L-Smooth	✓	✓	✓
MEHA (general case)	L-Smooth	L-Smooth Part & Weakly Convex Nonsmooth Part	✓	✓	✓

where X and Y are closed convex sets in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. The function $f(x,y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is smooth, and generally nonconvex, while $g(x,y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is potentially nonsmooth with respect to (w.r.t.) the LL variable y. For specific conditions governing F, f and g, we refer the reader to Assumptions 3.1, 3.2.

While the nonconvex-convex BLO has been extensively studied in the literature, efficient methods for nonconvex-nonconvex BLO remain under-explored. Beyond LL convexity, (Liu et al., 2021c) proposed an iterative differentiation-based BLO method; (Arbel & Mairal, 2022b) extended the approximate implicit differentiation approach; (Liu et al., 2021b) firstly utilized the value function reformulation of BLO to develop algorithms in machine learning. All of these works, however, tend to be complicated and impractical for large-scale BLO problems, and lack a non-asymptotic analysis. When LL objective satisfies the Polyak-Łojasiewicz (PL) or local PL conditions, (Ye et al., 2022) introduced a fully first-order value function-based BLO algorithm with non-asymptotic convergence analysis. Recently, while still considering the PL condition, (Huang, 2023b) introduced a momentum-based BLO algorithm; (Xiao et al., 2023) proposed a generalized alternating method; (Lu, 2023) proposed a smoothed first-order Lagrangian method; (Shen & Chen, 2023) proposed a penalty-based fully first-order BLO algorithm. However, the existing methods still present two significant challenges: ensuring computational efficiency and offering theoretical guarantees in the absence of the PL condition. A concise comparison with works closely related to ours is summarized in Table 1.

1.1. Main Contributions

To the best of our knowledge, this work is the first study to utilize Moreau envelope-based reformulation of BLO, originally presented in (Gao et al., 2023), to design a single-loop and Hessian-free gradient-based algorithm with non-asymptotic convergence analysis for general BLO problems with potentially nonconvex and nonsmooth LL objective functions. This setting encompasses a wide range of machine learning applications, see, e.g., the recent surveys (Liu et al., 2021a; Zhang et al., 2023). Conducting non-asymptotic analysis for our algorithm, which addresses nonconvex LL problem, poses substantial challenges. Existing single-loop gradient-based methods generally require the LL objective to either be strongly convex or satisfy the PL condition, as a mechanism to control the approximation errors incurred when utilizing a single gradient descent step to approximate the real LL optimal solution. Our approach mitigates this limitation by employing Moreau envelope-based reformulation, where the proximal LL problem may exhibit strong convexity even if the original LL problem is nonconvex. Consequently, this enables effective error control and facilitates the algorithm's non-asymptotic convergence analysis for nonconvex LL problem. We summarize our contributions as follows.

- We propose the **M**oreau **E**nvelope based **H**essian-free **A**lgorithm (MEHA), for general BLO problems with nonconvex and probably nonsmooth LL objective functions. MEHA avoids second-order derivative approximations related to the Hessian matrix and can be implemented efficiently in a single-loop manner, enhancing its practicality and efficiency for large-scale BLOs.
- We provide a rigorous analysis of the non-asymptotic convergence of MEHA under milder conditions, avoiding the need for either the convexity assumption or the PL condition on LL problem. In the context of the smooth BLO scenario, our assumption simplifies to UL and LL objective functions being L-smooth.
- We validate the effectiveness and efficiency of MEHA on various synthetic problems, two typicial hyper-parameter learning tasks and the real-world neural architecture search application. These experiments collectively substantiate its superior performance.

1.2. Related Work

We provide a brief review of recent works closely related to ours, with a detailed review available in Section A.9.

Nonconvex-Nonconvex BLO. Beyond LL convexity, (Huang, 2023b) introduced a momentum-based implicit gradient approach with convergence under PL conditions and nondegenerate LL Hessian. (Xiao et al., 2023) proposed a new stationary metric for nonconvex-PL BLOs, alongside a generalized alternating method. These methods, however, necessitate complex Hessian-related computations. Conversely, (Ye et al., 2022) offered a Hessian-free algorithm leveraging value function reformulation. (Lu, 2023) proposed a smoothed first-order Lagrangian method focused on the relaxation of the value function reformulation, also avoiding Hessian computations. (Shen & Chen, 2023) introduced a penalty-based algorithm applicable to both unconstrained and constrained BLOs. These Hessian-free BLO methods all have non-asymptotic convergence under PL conditions, but featuring double-loop structures.

Moreau Envelope based Reformulation for BLO. The authors in the work (Gao et al., 2023) pioneered the utilization of Moreau envelope of LL problem for investigating BLOs. They devised an inexact proximal Difference-of-Weakly-Convex algorithm in a double-loop manner for solving constrained BLO problems with convex LL problem. Subsequently, through the application of Moreau envelope-based reformulation of BLO, the study (Bai et al., 2023) derived new optimality condition results. More recently, while focusing on the convex smooth LL problem scenario, (Yao et al., 2024) introduced a min-max version of Moreau envelope for convex constrained LL problem and developed a single-loop Hessian-free gradient-based algorithm for the constrained BLO problem.

2. A Single-Loop and Hessian-Free Solution Strategy

2.1. Moreau Envelope based Reformulation

In this work, we focus on developing a single-loop algorithm for solving BLO with a nonconvex LL problem. Our approach builds upon the Moreau envelope based reformulation of BLO, which is initially proposed for convex LL scenarios in (Gao et al., 2023). The reformulation is expressed as follows:

$$\min_{(x,y)\in X\times Y} F(x,y) \quad \text{s.t.} \quad \varphi(x,y) - v_\gamma(x,y) \le 0, \tag{3}$$

where $v_\gamma(x,y)$ represents the Moreau envelope associated with the LL problem, defined as:

$$v_\gamma(x,y) := \inf_{\theta\in Y}\left\{\varphi(x,\theta) + \frac{1}{2\gamma}\|\theta - y\|^2\right\}, \tag{4}$$

with $\gamma > 0$. It is important to note that $\varphi(x,y) \ge v_\gamma(x,y)$ holds for all $(x,y)\in X\times Y$. For the convex LL scenarios where $\varphi(x,y)$ is convex in $y \in Y$ for any $x\in X$, the equivalence between the reformulated and the original BLO problems is established in Theorem 1 of (Gao et al., 2023). In scenarios where φ lacks convexity, yet $\varphi(x,\cdot)$ is ρ_{φ_2}-weakly convex [1] on Y and $\gamma \in (0, 1/\rho_{\varphi_2})$, we establish in Theorem A.1 that the reformulation (3) is equivalent to a relaxed version of BLO problem (1),

$$\min_{x\in X, y\in Y} F(x,y) \quad \text{s.t.} \quad y \in \tilde{S}(x), \tag{5}$$

where $\tilde{S}(x) := \{y \mid 0 \in \nabla_y f(x,y) + \partial_y g(x,y) + \mathcal{N}_Y(y)\}$, $\partial_y g(x,y)$ denotes the partial Fréchet (regular) subdifferential of g w.r.t. LL variable at (x,y), and $\mathcal{N}_Y(y)$ signifies the normal cone to Y at y. The stationary condition characterizing $\tilde{S}(x)$ is the first-order optimality conditions of LL problem within the context of this work, specifically, Assumption 3.2. This can be validated through the application of subdifferential sum rules, see, e.g., Proposition 1.30, Theorem 2.19 in (Mordukhovich, 2018).

Specifically, the reformulated problem (3) becomes equivalent to the original BLO problem when the set $\tilde{S}(x)$ coincides with $S(x)$. This equivalence holds, for instances, when $\varphi(x,\cdot)$ is convex or $\varphi(x,y) \equiv f(x,y)$ and it satisfies the PL condition, that is, there exists $\mu > 0$ such

[1] A function $h : \mathbb{R}^p \to \mathbb{R}\cup\{\infty\}$ is ρ-weakly convex if $h(z) + \frac{\rho}{2}\|z\|^2$ is convex. In the context where $z = (x,y)$, we always say h is (ρ_1,ρ_2)-weakly convex if $h(x,y) + \frac{\rho_1}{2}\|x\|^2 + \frac{\rho_2}{2}\|y\|^2$ is convex.

that for any $x \in X$, the inequality $\|\nabla_y f(x,y)\|^2 \geq 2\mu\big(f(x,y) - \inf_{\theta\in\mathbb{R}^m} f(x,\theta)\big)$ holds for all $y \in \mathbb{R}^m$.

We next introduce an approximation problem to the Moreau envelope based reformulation (3), constructed by penalizing the constraint $\varphi(x,y) - v_\gamma(x,y) \leq 0$ in (3). This approximation problem is formulated as:

$$\min_{(x,y)\in X\times Y} \psi_{c_k}(x,y) := F(x,y) + c_k\big(\varphi(x,y) - v_\gamma(x,y)\big), \tag{6}$$

where $\varphi(x,y) = f(x,y) + g(x,y)$, and c_k acts as the penalty parameter. Importantly, as $c_k \to \infty$, any limit point of the sequence of solutions to the approximation problem (6), associated with varying values of c_k, is a solution to the Moreau envelope based reformulation (3). This convergence is formally established in Theorem A.3.

2.2. Gradient of Moreau Envelope $v_\gamma(x,y)$

Before presenting our proposed method, we briefly review some relevant preliminary results related to $v_\gamma(x,y)$, with a special focus on its gradient. Assuming that $\varphi(x,y)$ is $(\rho_{\varphi_1}, \rho_{\varphi_2})$-weakly convex on $X \times Y$, we demonstrate that for $\gamma \in (0, \frac{1}{2\rho_{\varphi_2}})$, the function $v_\gamma(x,y) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{v_2}}{2}\|y\|^2$ is convex on $X \times \mathbb{R}^m$ when $\rho_{v_1} \geq \rho_{\varphi_1}$ and $\rho_{v_2} \geq 1/\gamma$. That is, $v_\gamma(x,y)$ is weakly convex, as detailed in Lemma A.4. Lastly, we define

$$S_\gamma(x,y) := \mathrm{argmin}_{\theta\in Y}\left\{\varphi(x,\theta) + \frac{1}{2\gamma}\|\theta - y\|^2\right\}. \tag{7}$$

For $\gamma \in (0, \frac{1}{2\rho_{\varphi_2}})$, the solution set $S_\gamma(x,y) = \{\theta_\gamma^*(x,y)\}$ is a singleton. Furthermore, when the gradients $\nabla_x f(x,y)$ and $\nabla_x g(x,y)$ exist, $v_\gamma(x,y)$ is differentiable and the gradient ∇v_γ of $v_\gamma(x,y)$ at (x,y) can be expressed as follows,

$$\nabla v_\gamma = \begin{pmatrix} \nabla_x f(x,\theta_\gamma^*(x,y)) + \nabla_x g(x,\theta_\gamma^*(x,y)) \\ \big(y - \theta_\gamma^*(x,y)\big)/\gamma \end{pmatrix}, \tag{8}$$

which is established in Lemma A.5.

2.3. Moreau Envelope based Hessian-Free Algorithm (MEHA)

In this part, we present a single-loop algorithm for the general BLO problem (1), via solving its approximation problem (6). At each iteration, we intend to employ the alternating proximal gradient method to the approximation problem (6) for updating the variables (x,y), starting from the current iterate (x^k,y^k). However, this process encounters certain challenges. Specifically, as outlined in (8), to calculate the gradient of the Moreau envelope $v_\gamma(x,y)$ at a given iterate (x^k,y^k), we have to know $\theta_\gamma^*(x^k,y^k)$, which is the unique solution to the proximal LL problem (4) with $(x,y) = (x^k,y^k)$. Exact computation of $\theta_\gamma^*(x^k,y^k)$ is computationally intensive. To mitigate this, we introduce a new iterative sequence $\{\theta^k\}$, where each θ^{k+1} approximates $\theta_\gamma^*(x^k,y^k)$. At each iteration, θ^{k+1} is generated by applying a single proximal gradient step to the proximal LL problem (4) with $(x,y) = (x^k,y^k)$, starting from the current θ^k. This update is formalized as:

Algorithm 1 Moreau Envelope based Hessian-Free Algorithm (MEHA)

```
Input: Initialize $x^0, y^0, \theta^0$, stepsizes $\alpha_k, \beta_k, \eta_k$, proximal parameter $\gamma$, penalty parameter $c_k$;
for $k = 0, 1, \ldots, K-1$ do
    update $\theta^{k+1}$ by (9);
    calculate $d_x^k$ as in (10) and update $x^{k+1}$ by (11);
    calculate $d_y^k$ as in (12) and update $y^{k+1}$ by (13).
end for
```

$$\theta^{k+1} = \mathrm{Prox}_{\eta_k\tilde{g}(x^k,\cdot)}\big(\theta^k - \eta_k\big(\nabla_y f(x^k,\theta^k) + \frac{\theta^k - y^k}{\gamma}\big)\big), \tag{9}$$

where η_k is the stepsize, and $\tilde{g}(x,y) := g(x,y) + \delta_Y(y)$ represents the nonsmooth part of LL problem. Here, $\mathrm{Prox}_h(y)$ denotes the proximal mapping of a function $h : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, $\mathrm{Prox}_h(y) := \arg\min_{\theta\in\mathbb{R}^m}\big\{h(\theta) + \|\theta - y\|^2/2\big\}$. Leveraging the formula (8), an approximation for $\nabla v_\gamma(x^k,y^k)$ is constructed using θ^{k+1} as an approximation for $\theta_\gamma^*(x^k,y^k)$,

$$(\nabla_x f(x^k,\theta^{k+1}) + \nabla_x g(x^k,\theta^{k+1}), (y^k - \theta^{k+1})/\gamma).$$

Subsequently, for the update of variable x, we construct an approximation for $[\nabla_x \psi_{c_k}(x^k,y^k)]/c_k$, serving as the update direction d_x^k:

$$\begin{aligned} d_x^k := {} & \frac{1}{c_k}\nabla_x F(x^k,y^k) + \nabla_x f(x^k,y^k) + \nabla_x g(x^k,y^k) \\ & - \nabla_x f(x^k,\theta^{k+1}) - \nabla_x g(x^k,\theta^{k+1}), \end{aligned} \tag{10}$$

and update x^{k+1} using the direction d_x^k:

$$x^{k+1} = \mathrm{Proj}_X\left(x^k - \alpha_k d_x^k\right), \tag{11}$$

where $\alpha_k > 0$ is the stepsize, and Proj_X denotes the Euclidean projection operator. Next, for variable y, we construct an approximation for $[\nabla_y(\psi_{c_k} - g)(x^{k+1},y^k)]/c_k$ as the update direction d_y^k:

$$d_y^k := \frac{1}{c_k}\nabla_y F(x^{k+1},y^k) + \nabla_y f(x^{k+1},y^k) - \frac{y^k - \theta^{k+1}}{\gamma}. \tag{12}$$

And y^{k+1} is updated via a proximal gradient step along d_y^k:

$$y^{k+1} = \mathrm{Prox}_{\beta_k\tilde{g}(x^{k+1},\cdot)}\left(y^k - \beta_k d_y^k\right), \tag{13}$$

where $\beta_k > 0$ is the stepsize.

The complete algorithm is outlined in Algorithm 1.

3. Theoretical Investigations

3.1. General Assumptions

Throughout this work, we assume the following standing assumptions on F, f and g hold.

Assumption 3.1. The UL objective F is bounded below on $X \times Y$, denoted by $\underline{F} := \inf_{(x,y)\in X\times Y} F(x,y) > -\infty$. Furthermore, F is L_F-smooth[2] on $X \times Y$. The smooth component of LL objective $f(x,y)$ is L_f-smooth on $X \times Y$.

Assumption 3.2. The nonsmooth component of LL objective, $g(x,y)$, satisfies one of the following conditions:

(i) $g(x,y) = \hat{g}(y)$ with $\hat{g}(y)$ being weakly convex on Y;

(ii) $g(x,y) = x\|y\|_1$ with $X = \mathbb{R}_+$ and $Y = \mathbb{R}^m$;

(iii) $g(x,y) = \sum_{j=1}^{J} x_j \|y^{(j)}\|_2$, where $\{1,\dots,m\}$ is divided into J groups, $y^{(j)}$ denotes the corresponding j-th group of y, and $X = \mathbb{R}_+^J$, $Y = \mathbb{R}^m$;

(iv) The nonsmooth component $g(x,y)$ is (ρ_{g_1}, ρ_{g_2})-weakly convex on $X \times Y$, i.e., $g(x,y) + \frac{\rho_{g_1}}{2}\|x\|^2 + \frac{\rho_{g_2}}{2}\|y\|^2$ is convex on $X \times Y$. Additionally, the gradient $\nabla_x g(x,y)$ exists and is L_g-Lipschitz continuous on $X \times Y$. Moreover, let $\tilde{g}(x,y) := g(x,y) + \delta_Y(y)$, there exist positive constants $L_{\tilde{g}}, \bar{s}$ such that for any $x, x' \in X$, $\theta \in Y$ and $s \in (0, \bar{s}]$,

$$\left\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}(\theta) - \mathrm{Prox}_{s\tilde{g}(x',\cdot)}(\theta)\right\| \le L_{\tilde{g}}\|x - x'\|. \tag{14}$$

The demand for weak convexity in Assumption 3.2 (i) is relatively lenient; a broad spectrum of functions meet this requirement. This encompasses conventional nonconvex regularizers like the Smoothly Clipped Absolute Deviation (SCAD) and the Minimax Concave Penalty (MCP) (refer to Section 2.1 of (Böhm & Wright, 2021)). It is worth noting that Assumptions 3.2 (i), (ii), and (iii) represent specific instances of Assumption 3.2(iv). Specifically, in Section A.8 of Appendix, we provide comprehensive proofs demonstrating that Assumptions 3.2 (ii) and (iii) are special cases of Assumption 3.2 (iv).

These assumptions considerably alleviate LL problem's smoothness requirements prevalent in BLO literature. Even within the context of smooth BLO, our assumptions only require that UL and LL objectives are both L-smooth, without imposing any conditions on the boundedness of $\nabla_y F(x,y)$, as illustrated in Table 1. Consequently, our problem setting encompasses a broad range of practical scenarios, see, e.g., the learning models in (Grazzi et al., 2020).

Finally, leveraging the descent lemma (e.g., Lemma 5.7 of (Beck, 2017)), it can be obtained that any function featuring a Lipschitz-continuous gradient is inherently weakly convex. Thus, under Assumption 3.1, $f(x,y)$ is (ρ_{f_1}, ρ_{f_2})-weakly convex over $X \times Y$, with $\rho_{f_1} = \rho_{f_2} = L_f$. This leads us to the following result.

Lemma 3.3. *Under Assumptions 3.1 and 3.2, LL objective $\varphi(x,y)$ is $(\rho_{\varphi_1}, \rho_{\varphi_2})$-weakly convex on $X \times Y$, where $\rho_{\varphi_1} = \rho_{f_1} + \rho_{g_1}$ and $\rho_{\varphi_2} = \rho_{f_2} + \rho_{g_2}$.*

3.2. Non-Asymptotic Convergence Results

We consider the following residual function, denoted as $R_k(x,y)$, for the approximation problem (6),

$$R_k := \mathrm{dist}\Big(0, \big[\nabla F + c_k\big(\nabla f + \partial g - \nabla v_\gamma\big) + \mathcal{N}_{X\times Y}\big]\Big). \tag{15}$$

This residual function is a stationarity measure for the approximation problem (6). In particular, it is evident that $R_k(x,y) = 0$ if and only if $0 \in \partial\psi_{c_k}(x,y) + \mathcal{N}_{X\times Y}(x,y)$, i.e., the point (x,y) is a stationary point to the approximation problem (6).

Theorem 3.4. *Under Assumptions 3.1 and 3.2, suppose $\gamma \in (0, \frac{1}{2\rho_{f_2}+2\rho_{g_2}})$, $c_k = \underline{c}(k+1)^p$ with $p \in [0, 1/2)$, $\underline{c} > 0$ and $\eta_k \in [\underline{\eta}, (1/\gamma - \rho_{f_2})/(L_f + 1/\gamma)^2] \cap [\underline{\eta}, 1/\rho_{g_2})$ with $\underline{\eta} > 0$, then there exists $c_\alpha, c_\beta > 0$ such that when $\alpha_k \in (\underline{\alpha}, c_\alpha)$ and $\beta_k \in (\underline{\beta}, c_\beta)$ with $\underline{\alpha}, \underline{\beta} > 0$, the sequence of (x^k, y^k, θ^k) generated by MEHA (Algorithm 1) satisfies*

$$\min_{0\le k\le K} \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\| = O\left(\frac{1}{K^{1/2}}\right),$$

and

$$\min_{0\le k\le K} R_k(x^{k+1}, y^{k+1}) = O\left(\frac{1}{K^{(1-2p)/2}}\right).$$

Furthermore, if $p \in (0, 1/2)$ and the sequence $\psi_{c_k}(x^k, y^k)$ is upper-bounded, the sequence of (x^k, y^k) satisfies

$$\varphi(x^K, y^K) - v_\gamma(x^K, y^K) = O\left(\frac{1}{K^p}\right).$$

Remark 3.5. In the case where the penalty parameter c_k remains fixed at a constant value $\underline{c}$, applying $p = 0$ in Theorem 3.4 yields a non-asymptotic $O(1/\sqrt{K})$ convergence rate result of MEHA in solving the approximation problem (6), with $c_k = \underline{c}$, based on its stationarity measure.

The proof of Theorem 3.4, which establishes the non-asymptotic convergence, relies on the monotonically decreasing property of the merit function:

$$V_k := \phi_{c_k}(x^k, y^k) + C_V \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\|^2,$$

where $\theta^*_\gamma(x,y)$ is the unique solution to problem (7), the constant $C_V := (L_f + L_g)^2 + 1/\gamma^2$, and

$$\phi_{c_k}(x,y) := \frac{1}{c_k}\big(F(x,y) - \underline{F}\big) + (f + g - v_\gamma)(x,y). \tag{16}$$

[2] A function h is said to be L-smooth on $X \times Y$ if h is continuously differentiable and its gradient ∇h is L-Lipschitz continuous on $X \times Y$.

Lemma 3.6. *Under Assumptions 3.1 and 3.2, suppose $\gamma \in (0, \frac{1}{2\rho_{f_2}+2\rho_{g_2}})$, $c_{k+1} \geq c_k$ and $\eta_k \in [\underline{\eta}, (1/\gamma - \rho_{f_2})/(L_f + 1/\gamma)^2] \cap [\underline{\eta}, 1/\rho_{g_2})$ with $\underline{\eta} > 0$, then there exists $c_\alpha, c_\beta, c_\theta > 0$ such that when $\alpha_k \in (0, c_\alpha]$ and $\beta_k \in (0, c_\beta]$, the sequence of (x^k, y^k, θ^k) generated by MEHA (Algorithm 1) satisfies*

$$
\begin{aligned}
V_{k+1} - V_k \leq & - c_\theta \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\|^2 \\
& - \frac{\|x^{k+1} - x^k\|^2}{4\alpha_k} - \frac{\|y^{k+1} - y^k\|^2}{4\beta_k}.
\end{aligned}
\tag{17}
$$

We outline the pivotal steps leading to Lemma 3.6.

Step 1: First, we consider the Moreau envelope $v_\gamma(x, y)$, focusing on two of its critical properties: its weak convexity (referenced in Lemma A.4) and the associated gradient formulas (outlined in Lemma A.5). These properties enable us to derive an upper bound for the descent of the penalized objective value $\phi_{c_k}(x, y)$ with incorporating the error term $\left(\frac{\alpha_k}{2}(L_f + L_g)^2 + \frac{\beta_k}{\gamma^2}\right) \left\|\theta^{k+1} - \theta^*_\gamma(x^k, y^k)\right\|^2$, as formulated in Equation (35) in Lemma A.10.

Step 2: The subsequent step involves leveraging the Lipschitz continuity of the Moreau envelope solution $\theta^*_\gamma(x, y)$ (as per Lemma A.8) along with the contraction properties of θ^k towards this solution (discussed in Lemma A.9). This approach is instrumental in controlling the error term $\left\|\theta^{k+1} - \theta^*_\gamma(x^k, y^k)\right\|^2$ as presented in Equation (35).

Ultimately, these steps culminate in confirming the monotonically decreasing property of the merit function V_k as in Lemma 3.6. The detailed proof is provided in Section A.5.

The norm of the hyper-gradient, denoted as $\nabla\Phi(x)$, where $\Phi(x) := F(x, y^*(x))$ and $y^*(x)$ representes the solution of the LL problem, is widely employed as a measure of stationary for BLO problems in the existing literature, see, e.g., (Ghadimi & Wang, 2018; Ji et al., 2020b; Chen et al., 2021; Ji et al., 2022; Hong et al., 2023; Kwon et al., 2023b). Nevertheless, it is crucial to note that this stationarity measure may not be suitable for BLO problems investigated in this work, as $\nabla\Phi(x)$ may not be well-defined. If we focus on a specific class of BLO problems, where $\nabla\Phi(x)$ is well-defined, we can establish the connection between the residual function $R_k(x, y)$ and the hypergradient $\nabla\Phi(x)$, as demonstrated in Lemma A.13. Consequently, Theorem 3.4 offers a non-asymptotic convergence result of MEHA based on the norm of $\nabla\Phi(x)$.

Theorem 3.7. *Under the same conditions as detailed in Theorem 3.4, and with the additional assumptions that $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, $g(x, y) = 0$, $f(x, y)$ is μ-strongly convex with respect to y for any x, F and f satisfy certain smoothness conditions (detailed in Lemma A.13), $p \in (0, 1/2)$, and $\gamma > 1/\mu$, the sequence of (x^k, y^k, θ^k) generated by MEHA (Algorithm 1) satisfies*

$$
\min_{0 \leq k \leq K} \|\nabla\Phi(x^{k+1})\| = O(\frac{1}{K^{1/2-p}} + \frac{1}{K^p}).
$$

4. Experimental Results

In this section, we verify the convergence behaviors and applicable practicality of MEHA on a series of different synthetic problems. Then we compare the performances on three real-world applications, including few-shot learning, data hyper-cleaning and neural architecture search. The code is available under `https://github.com/vis-opt-group/MEHA/`.

4.1. Synthetic Numerical Verification

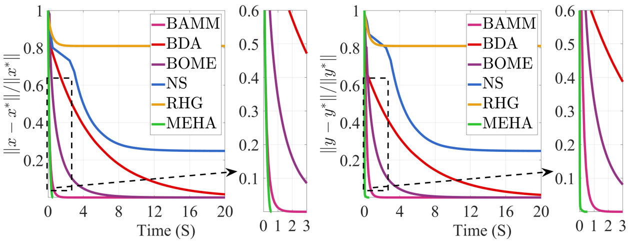

Figure 1. Illustrating the convergence curves of advanced BLO schemes and MEHA by the criteria, $\|x - x^*\|/\|x^*\|$ and $\|y - y^*\|/\|y^*\|$, under LL merely convex case.

LL Merely Convex Case. We first demonstrate the high efficiency of the proposed method on a toy example in LL merely convex case, expressed as follows:

$$
\begin{aligned}
& \min_{x \in \mathbb{R}^n, y=(y_1,y_2) \in \mathbb{R}^{2n}} \frac{1}{2}\left\|x - y_2\right\|^2 + \frac{1}{2}\left\|y_1 - \mathbf{e}\right\|^2 \\
& \text{s.t.} \quad y = (y_1, y_2) \in \arg \min_{(y_1,y_2) \in \mathbb{R}^{2n}} \frac{1}{2}\left\|y_1\right\|^2 - x^\top y_1.
\end{aligned}
\tag{18}
$$

The unique solution of this BLO is $(\mathbf{e}, \mathbf{e}, \mathbf{e})$. We plot the convergence curves on the Figure 1. Convergence criteria for numerical experiments are reported in Section A.11. Our proposed MEHA achieves the fastest convergence compared with existing BLO scheme, while some other methods fail to converge to the correct solution.

LL Non-Convex Case. We next demonstrate the effectiveness in handling the LL non-convex case by considering a toy example given in the form of

$$
\begin{aligned}
& \min_{x \in \mathbb{R}, y \in \mathbb{R}^n} \|x - a\|^2 + \|y - a\mathbf{e} - c\|^2 \\
& \text{s.t.} \quad y_i \in \arg\min_{y_i \in \mathbb{R}} \sin(x + y_i - c_i) \quad \forall i,
\end{aligned}
\tag{19}
$$

where constants $c \in \mathbb{R}^n$ and $a \in \mathbb{R}$. As shown the literature (Liu et al., 2021b), the optimal solution

Table 2. Comparison of total iterative time with representative BLO methods on the LL non-convex case with different dimensions.

	Dimension =1			Dimension =10		
Methods	BOME	IAPTT	MEHA	BOME	IAPTT	MEHA
Time (S)	5.062	21.23	**0.774**	6.348	35.73	**2.342**
	Dimension =50			Dimension =100		
Methods	BOME	IAPTT	MEHA	BOME	IAPTT	MEHA
Time (S)	11.02	154.54	**9.552**	16.48	290.82	**12.26**

Table 3. Comparison of LL non-smooth case utilizing lasso regression under diverse dimensions.

	Dimension=2				Dimension=100			Dimension=1000		
Methods	Grid	Random	TPE	MEHA	Random	TPE	MEHA	Random	TPE	MEHA
Time (S)	14.87	17.11	3.32	**1.91**	86.27	232.61	**2.34**	700.74	2244.44	**22.83**
	×7.79	×8.96	×1.74	-	×36.87	×99.4	-	×30.69	×98.3	-

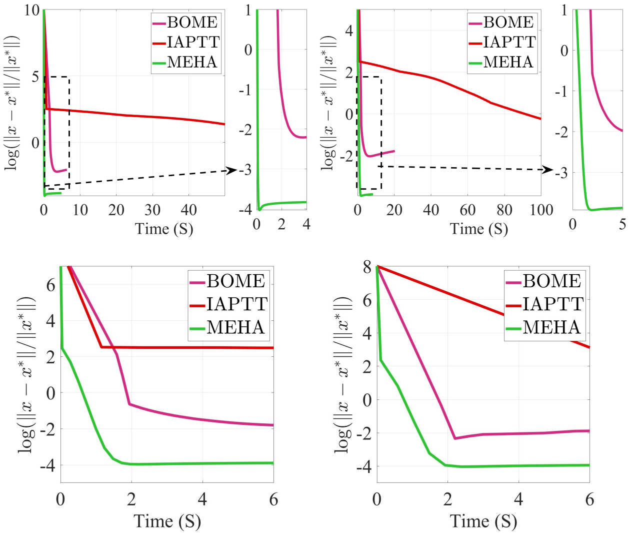

Figure 2. Convergence curves of advanced BLO methods and MEHA by the criteria, $\log(\|x-x^*\|/\|x^*\|)$, under the LL non-convex case with different dimensions (10, 50, 100, and 200).

Table 4. Comparison of computation efficiency under large scale dimension on the LL non-convex case.

Dimension	BOME	IAPTT	MEHA
200	44.47	622.77	**33.41**
300	48.73	1164.16	**37.85**
500	52.83	1657.46	**40.25**
1000	68.65	3537.19	**47.45**

is derived as $x^* = \frac{(1-n)a+nC}{1+n}$ and $y_i^* = C + c_i - x^*$ for $i = 1,\dots,n$. Here, C is defined as $C = \arg\min_k \left\{\|C_k - 2a\| : C_k = -\frac{\pi}{2} + 2k\pi, k \in \mathbb{Z}\right\}$. The optimal value is $F^* = \frac{n(C-2a)^2}{1+n}$. When $n = 1$, given $a = 2$ and $c = 2$, the concrete solution is $x^* = 3\pi/4$ and $y^* = 3\pi/4 + 2$. Given the initialization point $x_0, y_0 = (-6, 0)$, we compare the performance of MEHA with several Bi-Level Optimization (BLO) schemes (*i.e.,* BOME, and IAPTT) across different dimensions in Figure 2. Numerical comparisons, particularly the iterative time with advanced competitors, are reported in Table 2. Our method MEHA achieves the fastest convergence speed, outperforming BOME.

Computation Efficiency under Large Scale. We present the convergence time under large scale setting in Table. 4. MEHA demonstrates rapid convergence across various high-dimensional non-convex scenarios. In 1000 dimensions, MEHA achieves a time reduction of 30.9%.

LL Non-Smooth Case. We validate the computational efficiency of MEHA through a toy lasso regression example:

$$\min_{x\in\mathbb{R}^n, 0\le x\le 1, y\in\mathbb{R}^n} \sum_{i=1}^n y_i \\ \text{s.t.} \quad y \in \arg\min_{y\in\mathbb{R}^n} \frac{1}{2}\|y-v\|^2 + \sum_{i=1}^n x_i \|y_i\|_1, \tag{20}$$

where $v := \left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}, -\frac{1}{n}, -\frac{1}{n}, \cdots - \frac{1}{n}\right) \in \mathbb{R}^n$. The number of positive and negative values is $\frac{n}{2}$. The optimal solution can be calculated as $x_i \in [\frac{1}{n}, 1]$, $y_i = 0$ when $i = 1, \cdots \frac{n}{2}$ and $x_i = 0$, $y_i = -\frac{1}{n}$ when $i = \frac{n}{2} + 1, \cdots n$. The solving time of various methods with different dimensions is reported in Table 3. In comparison to grid search, random search, and Bayesian optimization-based TPE (Bergstra et al., 2013), our approach demonstrates superior efficiency, requiring the least search time to identify optimal solutions across varied dimensions, particularly notable in large-scale scenarios, such as 1000 dimensions. The convergence curves, depicting the accuracy of MEHA, are illustrated in Figure 6 in Appendix A.10.

Furthermore, we assess the algorithm's performance on the more challenging group lasso hyperparamter selection

Table 5. Comparison of convergence speed for few-shot learning (10-way and 20-way) and accuracy for data hyper-cleaning tasks.

Methods	10-Way		20-Way		FashionMNIST		MNIST	
	Acc. (%)	Time (S)	Acc. (%)	Time (S)	Acc. (%)	F1 score	Acc. (%)	F1 score
RHG	89.77	557.71	**90.17**	550.51	80.87	88.67	86.93	89.46
BDA	89.61	869.98	89.78	1187.63	80.93	87.75	85.83	89.20
CG	89.56	363.76	89.39	602.24	80.15	88.32	84.02	81.39
BAMM	90.57	180.48	90.13	255.99	81.64	88.16	88.60	89.91
IAPTT	89.66	299.27	89.57	3450.80	81.87	89.25	87.57	91.10
BOME	89.76	191.25	89.49	273.19	82.16	85.01	88.19	87.95
V-PBGD	90.28	315.59	90.00	649.93	81.72	89.93	89.10	90.64
MEHA	**91.17**	**130.81**	89.96	**235.77**	**83.19**	**89.94**	**89.55**	**91.26**

Table 6. Comparisons of sparse group lasso problem on the synthetic data. m represents the feature dimension.

Methods	m=600		m=1200	
	Test Err.	Time (S)	Test Err.	Time (S)
Grid	84.2	4.1	87.9	11.8
Random	176.5	4.5	170.4	10.7
TPE	123.9	11.7	157.2	58.2
IGJO	96.7	0.5	107.8	6.4
VF-iDCA	74.7	0.5	85.2	8.8
MEHA	**71.9**	**0.3**	**84.3**	**2.3**
Methods	m=2400		m=3600	
	Test Err.	Time (S)	Test Err.	Time (S)
Grid	89.5	25.8	91.3	38.3
Random	141.9	22.3	132.1	42.6
TPE	146.1	89.8	110.2	157.8
IGJO	115.5	12.9	120.4	24.6
VF-iDCA	85.2	16.9	90.0	49.2
MEHA	**85.1**	**2.2**	**89.5**	**2.5**

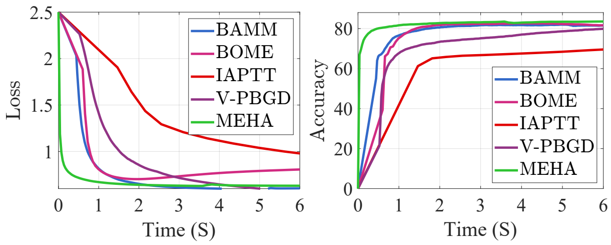

Figure 3. Comparison of data hyper-cleaning on FashionMNIST.

problem using synthetic data, formulated as follows:

$$\min_{x\in\mathbb{R}^n_+, y\in\mathbb{R}^m} l_{val}(y) \tag{21}$$
$$\text{s.t.} \quad y \in \arg\min_{y\in\mathbb{R}^m} l_{tr}(y) + \sum_{i=1}^{n} x_i \|y^{(i)}\|_2,$$

where l_{val} and l_{tr} denote the loss $\frac{1}{2}\sum_i \left|b_i - y^T a_i\right|^2$ on the validation and training datasets, respectively. Here, x denotes hyperparameters with n dimensions, while $\mathbf{a}$ and $\mathbf{b}$ stand for the inputs and their corresponding labels, respectively. Further details regarding data generation are elaborated in Section A.11. Competitive methods, including implicit differentiation IGJO (Feng & Simon, 2018) and VF-iDCA (Gao et al., 2022), are included for comparison. Numerical results repeated five times are presented in Table 6. Our proposed MEHA outperforms in high-dimensional tasks, yielding the most precise results.

4.2. Real-world Applications

Few-Shot Learning. The goal of N-way M-shot classification is to improve the adaptability of learnable model, facilitating rapid adaptation to new tasks. In our experiments using the Omniglot dataset (Finn et al., 2017), specifically in 10-way 1-shot and 20-way 1-shot scenarios with a LL convex formulation, we provide a runtime comparison in Table 5 to achieve consistent accuracy levels (90%) for both scenarios. Notably, our method attains comparable accuracy while significantly reducing computational time.

Data Hyper-Cleaning. In the right section of Table 5, we depict the accuracy and F1-score, for diverse methods (LL nonconvex case) on FashionMNIST and MNIST datasets. Remarkably, our proposed method significantly attains the desired solution with the highest accuracy. Furthermore, Figure 3 visually represents the validation loss and test accuracy across various methods. It is evident that our approach not only demonstrates the fastest convergence rate but also maintains its superior performance.

Neural Architecture Search. The main goal is to discover high-performance neural network structures. Our specific focus on differentiable NAS methods, representing a LL non-convex case. To showcase consistent performance, accuracy results are presented across different stages in Table 7. Remarkably, MEHA consistently outperforms specialized designs for NAS, establishing its superiority.

These methods mostly require massive engineering to construct search strategies and spaces. MEHA based on BLO formulation can better improve the performances with theoretical guarantees. Note that our theory can handle the neural architecture search problem with smooth activation functions, such as Swish. The results are reported in Table 9 of Section A.10.

Table 7. Comparing Top-1 accuracy in searching and evaluation stages for the specially designed schemes, including DARTS (Liu et al., 2018), P-DARTS (Chen et al., 2019), PC-DARTS (Xu et al., 2019).

Methods	Strategy	Searching		Evaluation		
		Train	Valid	Train	Valid	Test
DARTS	Single-Layer Approximation	98.320	88.940	99.481	95.639	95.569
P-DARTS	Progressive Search	96.168	90.488	**99.802**	95.701	95.710
PC-DARTS	Partially-Connected Search	84.821	83.516	98.163	95.630	95.540
MEHA	Bi-Level Optimization	**99.060**	**99.764**	99.419	**96.150**	**96.070**

5. Conclusions

By utilizing the Moreau envelope-based reformulation for general nonconvex and nonsmooth BLOs, we propose a provably single-loop and Hessian-free gradient-based algorithm, named MEHA. We validate the effectiveness and efficiency of MEHA for large-scale nonconvex-nonconvex BLO in both synthetic problems and various practical machine learning tasks. By leveraging the simplicity of our approach and integrating techniques such as variance reduction and momentum, we would be interested in studying the stochastic algorithms in the future.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

This work is partially supported by the National Key R&D Program of China (No. 2022YFA1004101), the National Natural Science Foundation of China (Nos. U22B2052, 12326605, 12222106, 62331014, 12371305), Guangdong Basic and Applied Basic Research Foundation (No. 2022B1515020082). Authors listed in alphabetical order.

References

Arbel, M. and Mairal, J. Amortized implicit differentiation for stochastic bilevel optimization. In *ICLR*, 2022a.

Arbel, M. and Mairal, J. Non-convex bilevel games with critical point selection maps. In *NeurIPS*, volume 35, pp. 8013–8026, 2022b.

Bai, K., Ye, J. J., and Zeng, S. Optimality conditions for bilevel programs via Moreau envelope reformulation. *arXiv preprint arXiv:2311.14857*, 2023.

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

Bergstra, J., Yamins, D., and Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *ICML*, pp. 115–123, 2013.

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

Bertrand, Q., Klopfenstein, Q., Massias, M., Blondel, M., Vaiter, S., Gramfort, A., and Salmon, J. Implicit differentiation for fast hyperparameter selection in non-smooth convex learning. *JMLR*, 23(1):6680–6722, 2022.

Böhm, A. and Wright, S. J. Variable smoothing for weakly convex composite functions. *JOTA*, 188:628–649, 2021.

Chen, L., Ma, Y., and Zhang, J. Near-optimal fully first-order algorithms for finding stationary points in bilevel optimization. *arXiv preprint arXiv:2306.14853*, 2023.

Chen, T., Sun, Y., and Yin, W. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *NeurIPS*, volume 34, pp. 25294–25307, 2021.

Chen, X., Xie, L., Wu, J., and Tian, Q. Progressive differentiable architecture search: Bridging the depth gap between search and evaluation. In *ICCV*, pp. 1294–1303, 2019.

Chen, Z., Kailkhura, B., and Zhou, Y. A fast and convergent proximal algorithm for regularized nonconvex and nonsmooth bi-level optimization. *arXiv preprint arXiv:2203.16615*, 2022.

Dagréou, M., Ablin, P., Vaiter, S., and Moreau, T. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *NeurIPS*, volume 35, pp. 26698–26710, 2022.

Elsken, T., Staffler, B., Metzen, J. H., and Hutter, F. Meta-learning of neural architectures for few-shot learning. In *CVPR*, pp. 12365–12375, 2020.

Feng, J. and Simon, N. Gradient-based regularization parameter selection for problems with nonsmooth penalty functions. *JCGS*, 27(2):426–435, 2018.

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

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

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *ICML*, pp. 1568–1577, 2018.

Gao, L. L., Ye, J. J., Yin, H., Zeng, S., and Zhang, J. Value function based difference-of-convex algorithm for bilevel hyperparameter selection problems. In *ICML*, pp. 7164–7182, 2022.

Gao, L. L., Ye, J. J., Yin, H., Zeng, S., and Zhang, J. Moreau envelope based difference-of-weakly-convex reformulation and algorithm for bilevel programs. *arXiv preprint arXiv:2306.16761*, 2023.

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

Grazzi, R., Franceschi, L., Pontil, M., and Salzo, S. On the iteration complexity of hypergradient computation. In *ICML*, pp. 3748–3758, 2020.

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIOPT*, 33(1): 147–180, 2023.

Huang, F. Adaptive mirror descent bilevel optimization. *arXiv preprint arXiv:2311.04520*, 2023a.

Huang, F. On momentum-based gradient methods for bilevel optimization with nonconvex lower-level. *arXiv preprint arXiv:2303.03944*, 2023b.

Huang, F., Li, J., Gao, S., and Huang, H. Enhanced bilevel optimization via Bregman distance. In *NeurIPS*, volume 35, pp. 28928–28939, 2022.

Ji, K. and Liang, Y. Lower bounds and accelerated algorithms for bilevel optimization. *JMLR*, 23:1–56, 2022.

Ji, K., Lee, J. D., Liang, Y., and Poor, H. V. Convergence of meta-learning with task-specific adaptation over partial parameters. In *NeurIPS*, volume 33, pp. 11490–11500, 2020a.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Nonasymptotic analysis and faster algorithms. *arXiv preprint arXiv:2010.07962*, 2020b.

Ji, K., Liu, M., Liang, Y., and Ying, L. Will bilevel optimizers benefit from loops. In *NeurIPS*, volume 35, pp. 3011–3023, 2022.

Kwon, J., Kwon, D., Wright, S., and Nowak, R. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. *arXiv preprint arXiv:2309.01753*, 2023a.

Kwon, J., Kwon, D., Wright, S., and Nowak, R. D. A fully first-order method for stochastic bilevel optimization. In *ICML*, pp. 18083–18113, 2023b.

Li, J., Gu, B., and Huang, H. Improved bilevel model: Fast and optimal algorithm with theoretical guarantee. *arXiv preprint arXiv:2009.00690*, 2020.

Li, J., Gu, B., and Huang, H. A fully single loop algorithm for bilevel optimization without hessian inverse. In *AAAI*, volume 36, pp. 7426–7434, 2022.

Liu, H., Simonyan, K., and Yang, Y. DARTS: Differentiable architecture search. In *ICLR*, 2018.

Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *ICML*, pp. 6305–6315, 2020.

Liu, R., Gao, J., Zhang, J., Meng, D., and Lin, Z. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE TPAMI*, 44(12):10045–10067, 2021a.

Liu, R., Liu, X., Yuan, X., Zeng, S., and Zhang, J. A value-function-based interior-point method for non-convex bi-level optimization. In *ICML*, pp. 6882–6892, 2021b.

Liu, R., Liu, Y., Zeng, S., and Zhang, J. Towards gradient-based bilevel optimization with non-convex followers and beyond. In *NeurIPS*, volume 34, pp. 8662–8675, 2021c.

Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A general descent aggregation framework for gradient-based bi-level optimization. *IEEE TPAMI*, 45(1):38–57, 2022.

Liu, R., Gao, J., Liu, X., and Fan, X. Learning with constraint learning: New perspective, solution strategy and various applications. *arXiv preprint arXiv:2307.15257*, 2023a.

Liu, R., Liu, Y., Yao, W., Zeng, S., and Zhang, J. Averaged method of multipliers for bi-level optimization without lower-level strong convexity. In *ICML*, pp. 21839–21866, 2023b.

Liu, R., Liu, Y., Zeng, S., and Zhang, J. Augmenting iterative trajectory for bilevel optimization: Methodology, analysis and extensions. *arXiv preprint arXiv:2303.16397*, 2023c.

Lu, S. SLM: A smoothed first-order Lagrangian method for structured constrained nonconvex optimization. In *NeurIPS*, 2023.

Lu, Z. and Mei, S. First-order penalty methods for bilevel optimization. *arXiv preprint arXiv:2301.01716*, 2023.

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

Mairal, J., Bach, F., and Ponce, J. Task-driven dictionary learning. *IEEE TPAMI*, 34(4):791–804, 2011.

Mordukhovich, B. S. *Variational analysis and applications*, volume 30. 2018.

Okuno, T., Takeda, A., and Kawana, A. Hyperparameter learning via bilevel nonsmooth optimization. *arXiv preprint arXiv:1806.01520*, 2018.

Pedregosa, F. Hyperparameter optimization with approximate gradient. In *ICML*, pp. 737–746, 2016.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. In *NeurIPS*, volume 32, 2019.

Rockafellar, R. T. *Conjugate duality and optimization*. SIAM, 1974.

Shen, H. and Chen, T. On penalty-based bilevel gradient descent method. In *ICML*, pp. 30992–31015, 2023.

Sow, D., Ji, K., Guan, Z., and Liang, Y. A constrained optimization approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022a.

Sow, D., Ji, K., and Liang, Y. On the convergence theory for hessian-free bilevel algorithms. In *NeurIPS*, volume 35, pp. 4136–4149, 2022b.

Xiao, Q., Lu, S., and Chen, T. An alternating optimization method for bilevel problems under the Polyak-Łojasiewicz condition. In *NeurIPS*, 2023.

Xu, Y., Xie, L., Zhang, X., Chen, X., Qi, G.-J., Tian, Q., and Xiong, H. Pc-darts: Partial channel connections for memory-efficient architecture search. In *ICLR*, 2019.

Yang, H., Luo, L., Li, C. J., and Jordan, M. I. Accelerating inexact hypergradient descent for bilevel optimization. *arXiv preprint arXiv:2307.00126*, 2023a.

Yang, Y., Xiao, P., and Ji, K. Achieving $O(\epsilon^{-1.5})$ complexity in Hessian/Jacobian-free stochastic bilevel optimization. In *NeurIPS*, 2023b.

Yao, W., Yu, C., Zeng, S., and Zhang, J. Constrained bi-level optimization: Proximal lagrangian value function approach and hessian-free algorithm. In *ICLR*, 2024.

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

Ye, J. J., Yuan, X., Zeng, S., and Zhang, J. Difference of convex algorithms for bilevel programs with applications in hyperparameter selection. *MP*, 198(2):1583–1616, 2023.

Ye, M., Liu, B., Wright, S., Stone, P., and Liu, Q. Bome! bilevel optimization made easy: A simple first-order approach. In *NeurIPS*, volume 35, pp. 17248–17262, 2022.

Zhang, Y., Khanduri, P., Tsaknakis, I., Yao, Y., Hong, M., and Liu, S. An introduction to bi-level optimization: Foundations and applications in signal processing and machine learning. *arXiv preprint arXiv:2308.00788*, 2023.

Zügner, D. and Günnemann, S. Adversarial attacks on graph neural networks via meta learning. In *ICLR*, 2018.

A. Appendix

The appendix is organized as follows:

- The equivalent result of the reformulated problem 3 is provided in Section A.1.
- The convergence of approximation problem (6) is established in Section A.2.
- We prove the weakly convexity and derive the gradient formula of Moreau Envelope in Section A.3.
- Some useful auxiliary lemmas are provided in Section A.4.
- The proof of Lemma 3.6 is given in Section A.5.
- The proof of Theorem 3.4 is provided in Section A.6.
- The proof of of Theorem 3.7 is provided in Section A.7.
- We verify Assumptions 3.2(ii) and (iii) are special cases of Assumption 3.2(iv) in Section A.8.
- Expanded related work is provided in Section A.9.
- Additional experimental results are provided in Section A.10.
- Experimental details are provided in Section A.11.

A.1. Equivalence of Moreau Envelope Based Reformulation

The following theorem establishes the equivalence between the Moreau Envelope based reformulation problem (3) and the relaxed bilevel optimization problem (5). The proof is inspired by the one of Theorem 1 in (Gao et al., 2023). For the convenience of the reader, we restate problems (3) and (5) as follows:

$$\min_{(x,y)\in X\times Y} F(x,y) \quad \text{s.t.} \quad \varphi(x,y) - v_\gamma(x,y) \le 0, \tag{3}$$

where $v_\gamma(x,y) := \inf_{\theta\in Y}\left\{\varphi(x,\theta) + \frac{1}{2\gamma}\|\theta - y\|^2\right\}$, $\varphi(x,y) = f(x,y) + g(x,y)$, and

$$\min_{x\in X, y\in Y} F(x,y) \quad \text{s.t.} \quad 0 \in \nabla_y f(x,y) + \partial_y g(x,y) + \mathcal{N}_Y(y). \tag{5}$$

Theorem A.1. *Suppose that $\varphi(x,\cdot)$ is ρ_{φ_2}-weakly convex on Y for all x, i.e., $\varphi(x,\cdot) + \frac{\rho_{\varphi_2}}{2}\|\cdot\|^2$ is convex on Y for all x. Then for $\gamma \in (0, 1/\rho_{\varphi_2})$, the Moreau Envelope based reformulation problem (3) is equivalent to the relaxed BLO problem (5).*

Proof. First, given any feasible point (x,y) of problem (3), it necessarily belongs to $X\times Y$ and satisfies

$$\varphi(x,y) \le v_\gamma(x,y) := \inf_{\theta\in Y}\left\{\varphi(x,\theta) + \frac{1}{2\gamma}\|\theta - y\|^2\right\} \le \varphi(x,y).$$

From which, it follows that $\varphi(x,y) = v_\gamma(x,y)$ and thus $y \in \operatorname{argmin}_{\theta\in Y}\left\{\varphi(x,\theta) + \frac{1}{2\gamma}\|\theta - y\|^2\right\}$. This leads to

$$0 \in \nabla_y f(x,y) + \partial_y g(x,y) + \mathcal{N}_Y(y),$$

implying that (x,y) is feasible for problem (5).

Conversely, consider that (x,y) is an feasible point of problem (5). This implies that $(x,y)\in X\times Y$, $0 \in \nabla_y f(x,y) + \partial_y g(x,y) + \mathcal{N}_Y(y)$. Given that $\varphi(x,\cdot) : \mathbb{R}^m \to \mathbb{R}$ is ρ_{φ_2}-weakly convex on Y, then when $\gamma \in (0, 1/\rho_{\varphi_2})$, the function $\varphi(x,\cdot) + \frac{1}{2\gamma}\|\cdot - y\|^2$ is convex on Y, making it lower regular. Clearly, $\delta_Y(\cdot)$ is lower regular since Y is a closed convex set.

By leveraging the subdifferential sum rules for two lower regular l.s.c. functions Theorem 2.19 of (Mordukhovich, 2018), we arrive at

$$\partial\left(\varphi(x,\cdot)+\frac{1}{2\gamma}\|\cdot-y\|^2+\delta_Y(\cdot)\right)=\nabla_y f(x,\cdot)+\partial_y g(x,\cdot)+(\cdot-y)/\gamma+\mathcal{N}_Y(\cdot).$$

With the right-hand set-valued mapping at y containing 0, we can deduce that

$$0\in\partial_\theta\left(\varphi(x,\theta)+\frac{1}{2\gamma}\|\theta-y\|^2+\delta_Y(\theta)\right)\bigg|_{\theta=y}.$$

Thus, invoking the first-order optimally condition for convex functions, we infer

$$y\in\operatorname{argmin}_{\theta\in Y}\left\{\varphi(x,\theta)+\frac{1}{2\gamma}\|\theta-y\|^2\right\}.$$

This implies $\varphi(x,y)=v_\gamma(x,y)$, confirming (x,y) as an feasible point to problem (3). □

A.2. Convergence of Approximation Problem (6)

For the convenience of the reader, we restate the approxiimation problem (6) as follows:

$$\min_{(x,y)\in X\times Y}\psi_{c_k}(x,y):=F(x,y)+c_k\Big(f(x,y)+g(x,y)-v_\gamma(x,y)\Big). \tag{6}$$

We will show that, as $c_k\to\infty$, any limit point of the sequence of solutions to the approximation problem (6), associated with varying values of c_k, is a solution to the Moreau envelope based reformulation (3). Using the same proof technique as presented in Lemma 2 of (Liu et al., 2020), we can establish the following result.

Lemma A.2. *If $\varphi(\cdot,y)$ is continuous on X for any $y\in Y$, then $v_\gamma(x,y)$ is upper semi-continuous on $X\times\mathbb{R}^m$.*

We are now prepared to establish the convergence of the approximation problem (6).

Theorem A.3. *Assume that X and Y are closed and F, f and g are all continuous on $X\times Y$. Additionally, suppose $c_k\to\infty$ and let*

$$(x_k,y_k)\in\underset{(x,y)\in X\times Y}{\operatorname{argmin}}\ \psi_{c_k}(x,y).$$

Then, for any limit point $(\bar{x},\bar{y})$ of the sequence $\{(x_k,y_k)\}$, $(\bar{x},\bar{y})$ is a solution to the Moreau envelope based reformulation (3).

Proof. Let $(\bar{x},\bar{y})$ be any limit point of the sequence $\{(x_k,y_k)\}$ and $\{(x_j,y_j)\}\subseteq\{(x_k,y_k)\}$ be the subsequence such that $(x_j,y_j)\to(\bar{x},\bar{y})$. Due to the closedness of both X and Y, we have $(\bar{x},\bar{y})\in X\times Y$. For any $\epsilon>0$, consider $(x_\epsilon,y_\epsilon)\in X\times Y$ as a point that satisfies $\varphi(x_\epsilon,y_\epsilon)-v_\gamma(x_\epsilon,y_\epsilon)\le 0$ and

$$F(x_\epsilon,y_\epsilon)<\inf_{x\in X,y\in Y}\{F(x,y)\ \text{ s.t. }\ \varphi(x,y)-v_\gamma(x,y)\le 0\}+\epsilon.$$

Then, since $(x_k,y_k)\in\arg\min_{(x,y)\in X\times Y}\psi_{c_k}(x,y)$, we have

$$\begin{aligned}&F(x_k,y_k)+c_k\Big(f(x_k,y_k)+g(x_k,y_k)-v_\gamma(x_k,y_k)\Big)\\ \le\ &F(x_\epsilon,y_\epsilon)<\inf_{x\in X,y\in Y}\{F(x,y)\ \text{ s.t. }\ \varphi(x,y)-v_\gamma(x,y)\le 0\}+\epsilon.\end{aligned} \tag{22}$$

By taking $k=j$ and letting $j\to\infty$ in (22), and as $c_k\to\infty$, we obtain

$$\limsup_{j\to\infty}\ f(x_j,y_j)+g(x_j,y_j)-v_\gamma(x_j,y_j)\le 0.$$

Now, considering that, as shown in Lemma A.2, $v_\gamma(x,y)$ is upper semi-continuous at $(\bar{x},\bar{y})$, and $f(x,y)$, and $g(x,y)$ are continuous at $(\bar{x},\bar{y})$, we can conclude that

$$f(\bar{x},\bar{y})+g(\bar{x},\bar{y})-v_\gamma(\bar{x},\bar{y})\le 0. \tag{23}$$

From (22), as $f(x_k, y_k) + g(x_k, y_k) - v_\gamma(x_k, y_k) \geq 0$, we have

$$F(x_k, y_k) < \inf_{x \in X, y \in Y} \{F(x, y) \text{ s.t. } \varphi(x, y) - v_\gamma(x, y) \leq 0\} + \epsilon.$$

By taking $k = j$ and letting $j \to \infty$ in the above inequality, and considering that $F(x, y)$ is continuous at $(\bar{x}, \bar{y})$, we obtain

$$F(\bar{x}, \bar{y}) \leq \inf_{x \in X, y \in Y} \{F(x, y) \text{ s.t. } \varphi(x, y) - v_\gamma(x, y) \leq 0\} + \epsilon.$$

Due to the arbitrariness of ϵ, we can conclude that

$$F(\bar{x}, \bar{y}) \leq \inf_{x \in X, y \in Y} \{F(x, y) \text{ s.t. } \varphi(x, y) - v_\gamma(x, y) \leq 0\},$$

and then the conclusion follows from $(\bar{x}, \bar{y}) \in X \times Y$ and (23). □

A.3. Properties of Moreau Envelope

By invoking Theorem 1 of (Rockafellar, 1974), we have that when the LL problem is fully convex, the Moreau Envelope $v_\gamma(x, y)$ is also convex. We further generalize this finding in the subsequent lemma, showing that $v_\gamma(x, y)$ retains weak convexity when the LL problem exhibits weak convexity. The foundation for this proof draws inspiration from Theorem 2 in (Gao et al., 2023).

Lemma A.4. *Suppose that $\varphi(x, y)$ is $(\rho_{\varphi_1}, \rho_{\varphi_2})$-weakly convex on $X \times Y$. Then for $\gamma \in (0, \frac{1}{2\rho_{\varphi_2}})$, $\rho_{v_1} \geq \rho_{\varphi_1}$ and $\rho_{v_2} \geq \frac{1}{\gamma}$, the function*

$$v_\gamma(x, y) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{v_2}}{2}\|y\|^2$$

is convex on $X \times \mathbb{R}^m$.

Proof. We first extend the definition of the Moreau envelope $v_\gamma(x, y)$ from $x \in X$ to $x \in \mathbb{R}^n$ by

$$v_\gamma(x, y) := \inf_{\theta \in \mathbb{R}^m} \left\{ \varphi(x, \theta) + \frac{1}{2\gamma}\|\theta - y\|^2 + \delta_{X \times Y}(x, \theta) \right\} \quad \forall x \in \mathbb{R}^n, y \in \mathbb{R}^m.$$

It follows that $v_\gamma(x, y) = +\infty$ for $x \notin X$. For any $\rho_{v_1}, \rho_{v_2} > 0$, the function $v_\gamma(x, y) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{v_2}}{2}\|y\|^2$ can be rewritten as

$$\begin{aligned} & v_\gamma(x, y) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{v_2}}{2}\|y\|^2 \\ = & \inf_{\theta \in \mathbb{R}^m} \left\{ \phi_{\gamma, \rho_v}(x, y, \theta) := \varphi(x, \theta) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{v_2}}{2}\|y\|^2 + \frac{1}{2\gamma}\|\theta - y\|^2 + \delta_{X \times Y}(x, \theta) \right\}. \end{aligned}$$

By direct computations, we obtain the following equation,

$$\begin{aligned} & \phi_{\gamma, \rho_v}(x, y, \theta) \\ = & \varphi(x, \theta) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{\varphi_2}}{2}\|\theta\|^2 + \delta_{X \times Y}(x, \theta) + \left(\frac{1}{2\gamma} - \frac{\rho_{\varphi_2}}{2} \right)\|\theta\|^2 + \frac{1 + \gamma\rho_{v_2}}{2\gamma}\|y\|^2 - \frac{1}{\gamma}\langle \theta, y \rangle. \end{aligned}$$

Given that $\rho_{v_1} \geq \rho_{\varphi_1}$, the convexity of $\varphi(x, \theta) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{\varphi_2}}{2}\|\theta\|^2 + \delta_{X \times Y}(x, \theta)$ can be immediately inferred, given that $\varphi(x, y)$ is $(\rho_{\varphi_1}, \rho_{\varphi_2})$-weakly convex on $X \times Y$.

Further, when $\gamma \in (0, \frac{1}{2\rho_{\varphi_2}})$ and $\rho_{v_2} \geq \frac{1}{\gamma}$, it can be shown that both conditions, $\frac{1}{4\gamma} - \frac{\rho_{\varphi_2}}{2} > 0$ and $\frac{1 + \gamma\rho_{v_2}}{2} \geq 1$, hold. This implies that the function

$$\begin{aligned} & \left(\frac{1}{2\gamma} - \frac{\rho_{\varphi_2}}{2} \right)\|\theta\|^2 + \frac{1 + \gamma\rho_{v_2}}{2\gamma}\|y\|^2 - \frac{1}{\gamma}\langle \theta, y \rangle \\ = & \left(\frac{1}{4\gamma} - \frac{\rho_{\varphi_2}}{2} \right)\|\theta\|^2 + \frac{1}{\gamma}\left(\frac{1}{4}\|\theta\|^2 + \frac{1 + \gamma\rho_{v_2}}{2}\|y\|^2 - \langle \theta, y \rangle \right), \end{aligned}$$

is convex with respect to (y, θ). Therefore, under the conditions $\gamma \in (0, \frac{1}{2\rho_{\varphi_2}})$, $\rho_{v_1} \geq \rho_{\varphi_1}$ and $\rho_{v_2} \geq \frac{1}{\gamma}$, the extended-valued function $\phi_{\gamma,\rho_v}(x, y, \theta)$ is convex with respect to (x, y, θ) over $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$. This, in turn , establishes the convexity of

$$v_\gamma(x, y) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{v_2}}{2}\|y\|^2 = \inf_{\theta \in \mathbb{R}^m} \phi_{\gamma,\rho_v}(x, y, \theta)$$

over $X \times \mathbb{R}^m$ by leveraging Theorem 1 of (Rockafellar, 1974). □

Next we develop a calculus for the Moreau Envelope $v_\gamma(x, y)$, providing formulas for its gradient. These results immediately give insights into the proposed algorithm. The proof closely follows that of Theorem 5 in (Gao et al., 2023).

Lemma A.5. *Under Assumption of Lemma A.4, suppose that the gradient $\nabla_x g(x, y)$ exists and is continuous on $X \times Y$. The for $\gamma \in (0, \frac{1}{2\rho_{\varphi_2}})$, $S_\gamma(x, y) = \{\theta^*_\gamma(x, y)\}$ is a singleton. Furthermore,*

$$\nabla v_\gamma(x, y) = \Big(\nabla_x f(x, \theta^*_\gamma(x, y)) + \nabla_x g(x, \theta^*_\gamma(x, y)),\ \big(y - \theta^*_\gamma(x, y)\big)/\gamma\Big). \tag{24}$$

Proof. Considering $\gamma \in (0, \frac{1}{2\rho_{\varphi_2}})$ and the weakly convexity of $\varphi(x, y)$, the function $\varphi(x, \theta) + \frac{1}{2\gamma}\|\theta - y\|^2 + \delta_Y(\theta)$ is shown to be $(\frac{1}{\gamma} - \frac{1}{\rho_{\varphi_2}})$-strongly convex with respect to θ. Consequently, $S_\gamma(x, y) = \{\theta^*_\gamma(x, y)\}$ is a singleton.

Further, for $\gamma \in (0, \frac{1}{2\rho_{\varphi_2}})$, we have $\rho_{v_1} \geq \rho_{\varphi_1}$ and $\rho_{v_2} \geq \frac{1}{\gamma}$. Leveraging Lemma A.4 and its subsequent proof, the function $v_\gamma(x, y) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{v_2}}{2}\|y\|^2$ is established as convex, and for any $(x, y) \in X \times Y$, the following holds

$$v_\gamma(x, y) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{v_2}}{2}\|y\|^2 = \inf_{\theta \in Y} \left\{\varphi(x, \theta) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{v_2}}{2}\|y\|^2 + \frac{1}{2\gamma}\|\theta - y\|^2\right\},$$

where $\varphi(x, \theta) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{v_2}}{2}\|y\|^2 + \frac{1}{2\gamma}\|\theta - y\|^2$ is convex with respect to (x, y, θ). By applying Theorem 3 of (Ye et al., 2023) and exploiting the continuously differentiable property of $g(x, y)$ with respect to x, the desired formulas are derived. □

A.4. Auxiliary Lemmas

In this section, we present auxiliary lemmas crucial for the non-asymptotic convergence analysis.

Lemma A.6. *Let $\gamma \in (0, \frac{1}{2\rho_{\varphi_2}})$, $(\bar{x}, \bar{y}) \in X \times \mathbb{R}^m$. Then for any $\rho_{v_1} \geq \rho_{\varphi_1}$, $\rho_{v_2} \geq \frac{1}{\gamma}$ and (x, y) on $X \times \mathbb{R}^m$, the following inequality holds:*

$$-v_\gamma(x, y) \leq -v_\gamma(\bar{x}, \bar{y}) - \langle \nabla v_\gamma(\bar{x}, \bar{y}), (x, y) - (\bar{x}, \bar{y})\rangle + \frac{\rho_{v_1}}{2}\|x - \bar{x}\|^2 + \frac{\rho_{v_2}}{2}\|y - \bar{y}\|^2. \tag{25}$$

Proof. According to Lemma A.4, $v_\gamma(x, y) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{v_2}}{2}\|y\|^2$ is convex on $X \times \mathbb{R}^m$. As a result, for any (x, y) on $X \times \mathbb{R}^m$,

$$\begin{aligned} & v_\gamma(x, y) + \frac{\rho_{v_1}}{2}\|x\|^2 + \frac{\rho_{v_2}}{2}\|y\|^2 \\ \geq\ & v_\gamma(\bar{x}, \bar{y}) + \frac{\rho_{v_1}}{2}\|\bar{x}\|^2 + \frac{\rho_{v_2}}{2}\|\bar{y}\|^2 + \langle \nabla v_\gamma(\bar{x}, \bar{y}) + (\rho_{v_1}\bar{x}, \rho_{v_2}\bar{y}), (x, y) - (\bar{x}, \bar{y})\rangle. \end{aligned}$$

Consequently, the conclusion follows directly. □

Lemma A.7. *For any $0 < s < 1/\rho_{g_2}$, and $\theta, \theta' \in \mathbb{R}^m$, the following inequality is satisfied:*

$$\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}(\theta) - \mathrm{Prox}_{s\tilde{g}(x,\cdot)}(\theta')\| \leq 1/(1 - s\rho_{g_2})\|\theta - \theta'\|. \tag{26}$$

Proof. Let us denote $\mathrm{Prox}_{s\tilde{g}(x,\cdot)}(\theta)$ and $\mathrm{Prox}_{s\tilde{g}(x,\cdot)}(\theta')$ by θ^+ and θ'^+, respectively. From the definitions, we have

$$0 \in \partial_y \tilde{g}(x, \theta^+) + \frac{1}{s}(\theta^+ - \theta),$$

and

$$0 \in \partial_y \tilde{g}(x, \theta'^+) + \frac{1}{s}(\theta'^+ - \theta').$$

Given the ρ_{g_2}-weakly convexity of $\tilde{g}(x,\cdot)$, it implies

$$\left\langle -\frac{1}{s}(\theta^+ - \theta) + \frac{1}{s}({\theta'}^+ - \theta'), \theta^+ - {\theta'}^+ \right\rangle \geq -\rho_{g_2}\|\theta^+ - {\theta'}^+\|^2.$$

From the above, the desired conclusion follows directly. □

Lemma A.8. *Let* $\gamma \in (0, \frac{1}{\rho_{f_2}+2\rho_{g_2}})$*. Then, there exists* $L_\theta > 0$ *such that for any* $(x,y),(x',y') \in X \times \mathbb{R}^m$*, the following inequality holds:*

$$\|\theta^*_\gamma(x,y) - \theta^*_\gamma(x',y')\| \leq L_\theta\|(x,y) - (x',y')\|. \tag{27}$$

Proof. Given that $\theta^*_\gamma(x,y)$ is optimal for the convex optimization problem $\min_{\theta\in Y}\ \varphi(x,\theta) + \frac{1}{2\gamma}\|\theta - y\|^2$, we have

$$\begin{aligned} &0 \in \nabla_y f(x, \theta^*_\gamma(x,y)) + \partial_y g(x, \theta^*_\gamma(x,y)) + (\theta^*_\gamma(x,y) - y)/\gamma + \mathcal{N}_Y(\theta^*_\gamma(x,y)),\\ &0 \in \nabla_y f(x', \theta^*_\gamma(x',y')) + \partial_y g(x', \theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y')/\gamma + \mathcal{N}_Y(\theta^*_\gamma(x',y')). \end{aligned}$$

Due to the ρ_{g_2}-weakly convexity of $\tilde{g}(x,y) := g(x,y) + \delta_Y(y)$ with respect to y, we obtain

$$\begin{aligned} \theta^*_\gamma(x,y) &= \mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x,y) - s\left(\nabla_y f(x,\theta^*_\gamma(x,y)) + (\theta^*_\gamma(x,y) - y)/\gamma\right)\right),\\ \theta^*_\gamma(x',y') &= \mathrm{Prox}_{s\tilde{g}(x',\cdot)}\left(\theta^*_\gamma(x',y') - s\left(\nabla_y f(x',\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y')/\gamma\right)\right), \end{aligned} \tag{28}$$

when $0 < s < 1/\rho_{g_2}$. Consequently, we deduce that

$$\begin{aligned} &\|\theta^*_\gamma(x,y) - \theta^*_\gamma(x',y')\|\\ =& \big\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x,y) - s\left(\nabla_y f(x,\theta^*_\gamma(x,y)) + (\theta^*_\gamma(x,y) - y)/\gamma\right)\right)\\ &\quad - \mathrm{Prox}_{s\tilde{g}(x',\cdot)}\left(\theta^*_\gamma(x',y') - s\left(\nabla_y f(x',\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y')/\gamma\right)\right)\big\|\\ \leq& \big\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x,y) - s\left(\nabla_y f(x,\theta^*_\gamma(x,y)) + (\theta^*_\gamma(x,y) - y)/\gamma\right)\right)\\ &\quad - \mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x',y') - s\left(\nabla_y f(x',\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y')/\gamma\right)\right)\big\|\\ &+ \big\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x',y') - s\left(\nabla_y f(x',\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y')/\gamma\right)\right)\\ &\quad - \mathrm{Prox}_{s\tilde{g}(x',\cdot)}\left(\theta^*_\gamma(x',y') - s\left(\nabla_y f(x',\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y')/\gamma\right)\right)\big\|\\ \leq& \big\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x,y) - s\left(\nabla_y f(x,\theta^*_\gamma(x,y)) + (\theta^*_\gamma(x,y) - y)/\gamma\right)\right)\\ &\quad - \mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x',y') - s\left(\nabla_y f(x,\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y)/\gamma\right)\right)\big\|\\ &+ \big\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x',y') - s\left(\nabla_y f(x,\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y)/\gamma\right)\right)\\ &\quad - \mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x',y') - s\left(\nabla_y f(x',\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y')/\gamma\right)\right)\big\|\\ &+ L_{\tilde{g}}\|x - x'\|, \end{aligned} \tag{29}$$

where the second inequality is a consequence of Assumption 3.2 (iv), which states that $\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}(\theta) - \mathrm{Prox}_{s\tilde{g}(x',\cdot)}(\theta)\| \leq L_{\tilde{g}}\|x - x'\|$ for any $\theta \in Y$ and $s \in (0,\bar{s}]$. Invoking Lemma A.7, for $0 < s < 1/\rho_{g_2}$, we derive

$$\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}(\theta) - \mathrm{Prox}_{s\tilde{g}(x,\cdot)}(\theta')\| \leq 1/(1 - s\rho_{g_2})\|\theta - \theta'\| \quad \forall \theta,\theta' \in \mathbb{R}^m. \tag{30}$$

Given that $f(x,\theta) + \frac{1}{2\gamma}\|\theta - y\|^2$ is $(\frac{1}{\gamma} - \rho_{f_2})$-strongly convex with respect to θ on Y, we have

$$\begin{aligned} &\left\langle \nabla_y f(x,\theta^*_\gamma(x,y)) + (\theta^*_\gamma(x,y) - y)/\gamma - \nabla_y f(x,\theta^*_\gamma(x',y')) - (\theta^*_\gamma(x',y') - y)/\gamma, \theta^*_\gamma(x,y) - \theta^*_\gamma(x',y')\right\rangle\\ &\geq \left(\frac{1}{\gamma} - \rho_{f_2}\right)\|\theta^*_\gamma(x,y) - \theta^*_\gamma(x',y')\|^2, \end{aligned}$$

which implies that when $0 < s \leq (1/\gamma - \rho_{f_2})/(L_f + 1/\gamma)^2$,

$$\begin{aligned} &\big\|\theta^*_\gamma(x,y) - s\left(\nabla_y f(x,\theta^*_\gamma(x,y)) + (\theta^*_\gamma(x,y) - y)/\gamma\right) - \theta^*_\gamma(x',y')\\ &\quad + s\left(\nabla_y f(x,\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y)/\gamma\right)\big\|^2\\ \leq& \left[1 - 2s\left(1/\gamma - \rho_{f_2}\right) + s^2(L_f + 1/\gamma)^2\right]\|\theta^*_\gamma(x,y) - \theta^*_\gamma(x',y')\|^2\\ \leq& \left[1 - s\left(1/\gamma - \rho_{f_2}\right)\right]\|\theta^*_\gamma(x,y) - \theta^*_\gamma(x',y')\|^2. \end{aligned}$$

Combining this with (30), we infer that

$$\begin{aligned}
&\big\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x,y) - s\left(\nabla_y f(x,\theta^*_\gamma(x,y)) + (\theta^*_\gamma(x,y) - y)/\gamma\right)\right) \\
&\qquad - \mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x',y') - s\left(\nabla_y f(x,\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y)/\gamma\right)\right)\big\| \\
\le\,& 1/(1 - s\rho_{g_2})\big\|\theta^*_\gamma(x,y) - s\left(\nabla_y f(x,\theta^*_\gamma(x,y)) + (\theta^*_\gamma(x,y) - y)/\gamma\right) \\
&\qquad\qquad - \theta^*_\gamma(x',y') + s\left(\nabla_y f(x,\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y)/\gamma\right)\big\| \\
\le\,& \sqrt{1 - s\left(1/\gamma - \rho_{f_2}\right)}/(1 - s\rho_{g_2})\|\theta^*_\gamma(x,y) - \theta^*_\gamma(x',y')\|.
\end{aligned} \tag{31}$$

Next, utilizing Lemma A.7, for $0 < s < 1/\rho_{g_2}$, it follows that

$$\begin{aligned}
&\big\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x',y') - s\left(\nabla_y f(x,\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y)/\gamma\right)\right) \\
&\qquad - \mathrm{Prox}_{s\tilde{g}(x,\cdot)}\left(\theta^*_\gamma(x',y') - s\left(\nabla_y f(x',\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y')/\gamma\right)\right)\big\| \\
\le\,& 1/(1 - s\rho_{g_2})\big\|\theta^*_\gamma(x',y') - s\left(\nabla_y f(x,\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y)/\gamma\right) \\
&\qquad\qquad - \theta^*_\gamma(x',y') + s\left(\nabla_y f(x',\theta^*_\gamma(x',y')) + (\theta^*_\gamma(x',y') - y')/\gamma\right)\big\| \\
\le\,& s/(1 - s\rho_{g_2})\left(\|\nabla_y f(x,\theta^*_\gamma(x',y')) - \nabla_y f(x',\theta^*_\gamma(x',y'))\| + \|y - y'\|/\gamma\right) \\
\le\,& s/(1 - s\rho_{g_2})\left(L_f\|x - x'\| + \frac{1}{\gamma}\|y - y'\|\right).
\end{aligned} \tag{32}$$

From estimates (29), (31) and (32), we deduce that, for any $s > 0$ satisfying $s \le (1/\gamma - \rho_{f_2})/(L_f + 1/\gamma)^2$, $s \le \bar{s}$ and $s < 1/\rho_{g_2}$, the following condition holds

$$\begin{aligned}
\|\theta^*_\gamma(x,y) - \theta^*_\gamma(x',y')\| \le\,& \sqrt{1 - s\left(1/\gamma - \rho_{f_2}\right)}/(1 - s\rho_{g_2})\|\theta^*_\gamma(x,y) - \theta^*_\gamma(x',y')\| \\
&+ s/(1 - s\rho_{g_2})\left(L_f\|x - x'\| + \frac{1}{\gamma}\|y - y'\|\right) + L_{\tilde{g}}\|x - x'\|.
\end{aligned} \tag{33}$$

Given that $\gamma < \frac{1}{\rho_{f_2} + 2\rho_{g_2}}$, it can be inferred that $1/\gamma - \rho_{f_2} > 2\rho_{g_2}$. This implies $1 - 2s\rho_{g_2} > 1 - s(1/\gamma - \rho_{f_2})$, leading to $1 - s(1/\gamma - \rho_{f_2}) < (1 - s\rho_{g_2})^2$. Consequently, we deduce $\sqrt{1 - s\left(1/\gamma - \rho_{f_2}\right)}/(1 - s\rho_{g_2}) < 1$. From these derivations, the desired conclusion is evident. □

Lemma A.9. *Suppose* $\gamma \in (0, \frac{1}{\rho_{f_2} + 2\rho_{g_2}})$ *and* $\eta_k \in (0, (1/\gamma - \rho_{f_2})/(L_f + 1/\gamma)^2] \cap (0, 1/\rho_{g_2})$*, the sequence of* (x^k, y^k, θ^k) *generated by Algorithm 1 satisfies*

$$\|\theta^{k+1} - \theta^*_\gamma(x^k,y^k)\| \le \sigma_k\|\theta^k - \theta^*_\gamma(x^k,y^k)\|, \tag{34}$$

where $\sigma_k := \sqrt{1 - \eta_k\left(1/\gamma - \rho_{f_2}\right)}/(1 - \eta_k\rho_{g_2}) < 1$.

Proof. Recalling (28) from Lemma A.8 that when $\eta_k < 1/\rho_{g_2}$,

$$\theta^*_\gamma(x^k,y^k) = \mathrm{Prox}_{\eta_k\tilde{g}(x^k,\cdot)}\left(\theta^*_\gamma(x^k,y^k) - \eta_k\left(\nabla_y f(x^k,\theta^*_\gamma(x,y)) + (\theta^*_\gamma(x^k,y^k) - y^k)/\gamma\right)\right).$$

Considering the update rule for θ^{k+1} as defined in (9) and using arguments analogous to those in the derivation of (31) from Lemma A.8, when $\eta_k \le (1/\gamma - \rho_{f_2})/(L_f + 1/\gamma)^2$, it follows

$$\|\theta^{k+1} - \theta^*_\gamma(x^k,y^k)\| \le \sigma_k\|\theta^k - \theta^*_\gamma(x^k,y^k)\|,$$

where $\sigma_k := \sqrt{1 - \eta_k\left(1/\gamma - \rho_{f_2}\right)}/(1 - \eta_k\rho_{g_2})$. Notably, $\sigma_k < 1$ is a consequence of $\gamma < \frac{1}{\rho_{f_2} + 2\rho_{g_2}}$. □

The update of variables (x, y) in (11) and (13) can be interpreted as inexact alternating proximal gradient from (x^k, y^k) on $\min_{(x,y)\in X\times Y} \phi_{c_k}(x,y)$, in which ϕ_{c_k} is defined in (16) as

$$\phi_{c_k}(x,y) := \frac{1}{c_k}\left(F(x,y) - \underline{F}\right) + f(x,y) + g(x,y) - v_\gamma(x,y).$$

The subsequent lemma illustrates that the function $\phi_{c_k}(x,y)$ exhibits a monotonic decreasing behavior with errors at each iteration.

Lemma A.10. *Under Assumptions 3.1 and 3.2, suppose* $\gamma \in (0, \frac{1}{2\rho_{f_2}+2\rho_{g_2}})$ *and* $\beta_k < 1/\rho_{g_2}$*, the sequence of* (x^k, y^k, θ^k) *generated by Algorithm 1 satisfies*

$$\begin{aligned}\phi_{c_k}(x^{k+1}, y^{k+1}) \le \phi_{c_k}(x^k, y^k) &- \left(\frac{1}{2\alpha_k} - \frac{L_{\phi_k}}{2} - \frac{\beta_k L_\theta^2}{\gamma^2}\right)\|x^{k+1}-x^k\|^2 \\ &- \left(\frac{1}{2\beta_k} - \frac{\rho_{g_2}}{2} - \frac{L_{\phi_k}}{2}\right)\|y^{k+1}-y^k\|^2 \\ &+ \left(\frac{\alpha_k}{2}(L_f+L_g)^2 + \frac{\beta_k}{\gamma^2}\right)\left\|\theta^{k+1} - \theta^*_\gamma(x^k, y^k)\right\|^2,\end{aligned} \tag{35}$$

where $L_{\phi_k} := L_F/c_k + L_f + L_g + \max\{\rho_{\varphi_1}, 1/\gamma\}$.

Proof. Under the conditions of Assumption 3.1, the functions F and f exhibit L_F- and L_f-smooth on $X \times Y$, respectively. Further, according to Assumption 3.2, the function $g(\cdot, y^k)$ is L_g-smooth on X. Leveraging these assumptions and invoking Lemma A.6, we deduce

$$\phi_{c_k}(x^{k+1}, y^k) \le \phi_{c_k}(x^k, y^k) + \langle \nabla_x \phi_{c_k}(x^k, y^k), x^{k+1} - x^k\rangle + \frac{L_{\phi_k}}{2}\|x^{k+1}-x^k\|^2, \tag{36}$$

with $L_{\phi_k} := L_F/c_k + L_f + L_g + \max\{\rho_{\varphi_1}, 1/\gamma\}$. Considering the update rule for the variable x as defined in (11) and leveraging the property of the projection operator Proj_X, it follows that

$$\langle x^k - \alpha_k d_x^k - x^{k+1}, x^k - x^{k+1}\rangle \le 0,$$

leading to

$$\langle d_x^k, x^{k+1} - x^k\rangle \le -\frac{1}{\alpha_k}\|x^{k+1}-x^k\|^2.$$

Combining this inequality with (36), if can be deduced that

$$\begin{aligned}\phi_{c_k}(x^{k+1}, y^k) \le \phi_{c_k}(x^k, y^k) &- \left(\frac{1}{\alpha_k} - \frac{L_{\phi_k}}{2}\right)\|x^{k+1}-x^k\|^2 \\ &+ \left\langle \nabla_x \phi_{c_k}(x^k, y^k) - d_x^k, x^{k+1} - x^k\right\rangle.\end{aligned} \tag{37}$$

Given the expression for $\nabla v_\gamma(x, y)$ as derived in Lemma A.5 and the definition of d_x^k provided in (10), we obtain

$$\begin{aligned}&\left\|\nabla_x \phi_{c_k}(x^k, y^k) - d_x^k\right\|^2 \\ =& \left\|\nabla_x f(x^k, \theta^*_\gamma(x^k, y^k)) + \nabla_x g(x^k, \theta^*_\gamma(x^k, y^k)) - \nabla_x f(x^k, \theta^{k+1}) - \nabla_x g(x^k, \theta^{k+1})\right\|^2 \\ \le& (L_f + L_g)^2\left\|\theta^{k+1} - \theta^*_\gamma(x^k, y^k)\right\|^2.\end{aligned} \tag{38}$$

This yields

$$\begin{aligned}&\left\langle \nabla_x \phi_{c_k}(x^k, y^k) - d_x^k, x^{k+1} - x^k\right\rangle \\ \le& \frac{\alpha_k}{2}(L_f+L_g)^2\left\|\theta^{k+1} - \theta^*_\gamma(x^k, y^k)\right\|^2 + \frac{1}{2\alpha_k}\|x^{k+1}-x^k\|^2,\end{aligned}$$

which combining with (37) leads to

$$\begin{aligned}\phi_{c_k}(x^{k+1}, y^k) \le \phi_{c_k}(x^k, y^k) &- \left(\frac{1}{2\alpha_k} - \frac{L_{\phi_k}}{2}\right)\|x^{k+1}-x^k\|^2 \\ &+ \frac{\alpha_k}{2}(L_f+L_g)^2\left\|\theta^{k+1} - \theta^*_\gamma(x^k, y^k)\right\|^2.\end{aligned} \tag{39}$$

Considering the update rule for variable y given by (13), and the ρ_{g_2}-weakly convex property of $g(x^{k+1}, \cdot)$ over Y, it follows that for $\beta_k < 1/\rho_{g_2}$,

$$\langle d_y^k, y^{k+1} - y^k\rangle + g(x^{k+1}, y^{k+1}) + \left(\frac{1}{\beta_k} - \frac{\rho_{g_2}}{2}\right)\|y^{k+1}-y^k\|^2 \le g(x^{k+1}, y^k). \tag{40}$$

Under Assumption 3.1, where F and f are L_F- and l_f-smooth on $X \times Y$, respectively, and invoking Lemma A.6, we deduce

$$\begin{aligned}&\phi_{c_k}(x^{k+1}, y^{k+1}) - g(x^{k+1}, y^{k+1}) \\ &\le \phi_{c_k}(x^{k+1}, y^k) - g(x^{k+1}, y^k) + \langle \nabla_y \left(\phi_{c_k} - g\right)(x^{k+1}, y^k), y^{k+1} - y^k \rangle + \frac{L_{\phi_k}}{2}\|y^{k+1} - y^k\|^2.\end{aligned} \tag{41}$$

Combining this inequality with (40), we obtain

$$\begin{aligned}&\phi_{c_k}(x^{k+1}, y^{k+1}) \\ &\le \phi_{c_k}(x^{k+1}, y^k) - \left(\frac{1}{\beta_k} - \frac{\rho_{g_2}}{2} - \frac{L_{\phi_k}}{2}\right)\|y^{k+1} - y^k\|^2 \\ &\quad + \left\langle \nabla_y \left(\phi_{c_k} - g\right)(x^{k+1}, y^k) - d_y^k, y^{k+1} - y^k \right\rangle.\end{aligned} \tag{42}$$

Given the expression for $\nabla v_\gamma(x, y)$ as derived in Lemma A.5 and the definition of d_y^k from (12), we deduce

$$\begin{aligned}\left\|\nabla_y \left(\phi_{c_k} - g\right)(x^{k+1}, y^k) - d_y^k\right\|^2 &= \left\|(y^k - \theta_\gamma^*(x^{k+1}, y^k))/\gamma - (y^k - \theta^{k+1})/\gamma\right\|^2 \\ &= \frac{1}{\gamma^2}\left\|\theta^{k+1} - \theta_\gamma^*(x^{k+1}, y^k)\right\|^2,\end{aligned} \tag{43}$$

and thus

$$\left\langle \nabla_y \left(\phi_{c_k} - g\right)(x^{k+1}, y^k) - d_y^k, y^{k+1} - y^k \right\rangle \le \frac{\beta_k}{2\gamma^2}\left\|\theta^{k+1} - \theta_\gamma^*(x^{k+1}, y^k)\right\|^2 + \frac{1}{2\beta_k}\|y^{k+1} - y^k\|^2.$$

Consequently, we have from (42) that

$$\begin{aligned}&\phi_{c_k}(x^{k+1}, y^{k+1}) \\ &\le \phi_{c_k}(x^{k+1}, y^k) - \left(\frac{1}{2\beta_k} - \frac{\rho_{g_2}}{2} - \frac{L_{\phi_k}}{2}\right)\|y^{k+1} - y^k\|^2 + \frac{\beta_k}{2\gamma^2}\left\|\theta^{k+1} - \theta_\gamma^*(x^{k+1}, y^k)\right\|^2 \\ &\le \phi_{c_k}(x^{k+1}, y^k) - \left(\frac{1}{2\beta_k} - \frac{\rho_{g_2}}{2} - \frac{L_{\phi_k}}{2}\right)\|y^{k+1} - y^k\|^2 + \frac{\beta_k}{\gamma^2}\left\|\theta^{k+1} - \theta_\gamma^*(x^k, y^k)\right\|^2 \\ &\quad + \frac{\beta_k L_\theta^2}{\gamma^2}\left\|x^{k+1} - x^k\right\|^2,\end{aligned} \tag{44}$$

where the last inequality follows from Lemma A.8. The conclusion follows by combining this with (39). □

A.5. Proof of Lemma 3.6

With the auxiliary lemmas from the preceding section, we demonstrate the decreasing property of the merit function V_k.

Lemma A.11. *Under Assumptions 3.1 and 3.2, suppose* $\gamma \in (0, \frac{1}{2\rho_{f_2} + 2\rho_{g_2}})$, $c_{k+1} \ge c_k$ *and* $\eta_k \in [\underline{\eta}, (1/\gamma - \rho_{f_2})/(L_f + 1/\gamma)^2] \cap [\underline{\eta}, 1/\rho_{g_2})$ *with* $\underline{\eta} > 0$, *then there exists* $c_\alpha, c_\beta, c_\theta > 0$ *such that when* $0 < \alpha_k \le c_\alpha$ *and* $0 < \beta_k \le c_\beta$, *the sequence of* (x^k, y^k, θ^k) *generated by MEHA (Algorithm 1) satisfies*

$$V_{k+1} - V_k \le -\frac{1}{4\alpha_k}\|x^{k+1} - x^k\|^2 - \frac{1}{4\beta_k}\|y^{k+1} - y^k\|^2 - c_\theta\left\|\theta^k - \theta_\gamma^*(x^k, y^k)\right\|^2, \tag{45}$$

where $c_\theta = \frac{1}{2}\left(\frac{\eta\rho_{g_2}}{1 - \underline{\eta}\rho_{g_2}}\right)^2\left((L_f + L_g)^2 + 1/\gamma^2\right)$.

Proof. Let us first recall (35) from Lemma A.10, which states that

$$\begin{aligned}\phi_{c_k}(x^{k+1}, y^{k+1}) \le \phi_{c_k}(x^k, y^k) &- \left(\frac{1}{2\alpha_k} - \frac{L_{\phi_k}}{2} - \frac{\beta_k L_\theta^2}{\gamma^2}\right)\|x^{k+1} - x^k\|^2 \\ &- \left(\frac{1}{2\beta_k} - \frac{\rho_{g_2}}{2} - \frac{L_{\phi_k}}{2}\right)\|y^{k+1} - y^k\|^2 \\ &+ \left(\frac{\alpha_k}{2}(L_f + L_g)^2 + \frac{\beta_k}{\gamma^2}\right)\left\|\theta^{k+1} - \theta_\gamma^*(x^k, y^k)\right\|^2,\end{aligned} \tag{46}$$

when $\beta_k < 1/\rho_{g_2}$. Since $c_{k+1} \geq c_k$, we can infer that $(F(x^{k+1}, y^{k+1}) - \underline{F})/c_{k+1} \leq (F(x^{k+1}, y^{k+1}) - \underline{F})/c_k$. Combining this with (46) leads to

$$
\begin{aligned}
V_{k+1} - V_k = & \, \phi_{c_{k+1}}(x^{k+1}, y^{k+1}) - \phi_{c_k}(x^k, y^k) \\
& + \left((L_f + L_g)^2 + 1/\gamma^2\right) \left\|\theta^{k+1} - \theta^*_\gamma(x^{k+1}, y^{k+1})\right\|^2 \\
& - \left((L_f + L_g)^2 + 1/\gamma^2\right) \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\|^2 \\
\leq & \, \phi_{c_k}(x^{k+1}, y^{k+1}) - \phi_{c_k}(x^k, y^k) \\
& + \left((L_f + L_g)^2 + 1/\gamma^2\right) \left\|\theta^{k+1} - \theta^*_\gamma(x^{k+1}, y^{k+1})\right\|^2 \\
& - \left((L_f + L_g)^2 + 1/\gamma^2\right) \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\|^2 \\
\leq & - \left(\frac{1}{2\alpha_k} - \frac{L_{\phi_k}}{2} - \frac{\beta_k L_\theta^2}{\gamma^2}\right) \|x^{k+1} - x^k\|^2 \\
& - \left(\frac{1}{2\beta_k} - \frac{\rho_{g_2}}{2} - \frac{L_{\phi_k}}{2}\right) \|y^{k+1} - y^k\|^2 \\
& + \left((L_f + L_g)^2 + 1/\gamma^2\right) \left\|\theta^{k+1} - \theta^*_\gamma(x^{k+1}, y^{k+1})\right\|^2 \\
& - \left((L_f + L_g)^2 + 1/\gamma^2\right) \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\|^2 \\
& + \left(\frac{\alpha_k}{2}(L_f + L_g)^2 + \frac{\beta_k}{\gamma^2}\right) \left\|\theta^{k+1} - \theta^*_\gamma(x^k, y^k)\right\|^2 .
\end{aligned}
\tag{47}
$$

We can demonstrate that

$$
\begin{aligned}
& \left\|\theta^{k+1} - \theta^*_\gamma(x^{k+1}, y^{k+1})\right\|^2 - \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\|^2 + \frac{\alpha_k}{2} \left\|\theta^{k+1} - \theta^*_\gamma(x^k, y^k)\right\|^2 \\
\leq & (1 + \epsilon_k + \frac{\alpha_k}{2}) \left\|\theta^{k+1} - \theta^*_\gamma(x^k, y^k)\right\|^2 - \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\|^2 \\
& + (1 + \frac{1}{\epsilon_k}) \|\theta^*_\gamma(x^{k+1}, y^{k+1}) - \theta^*_\gamma(x^k, y^k)\|^2 \\
\leq & (1 + \epsilon_k + \frac{\alpha_k}{2}) \sigma_k^2 \|\theta^k - \theta^*_\gamma(x^k, y^k)\|^2 - \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\|^2 \\
& + (1 + \frac{1}{\epsilon_k}) L_\theta^2 \left\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\right\|^2 ,
\end{aligned}
$$

for any $\epsilon_k > 0$, where the second inequality is a consequence of Lemmas A.8 and A.9. Since $\gamma < \frac{1}{\rho_{f_2} + 2\rho_{g_2}}$, we have $1 - 2\eta_k \rho_{g_2} > 1 - \eta_k(1/\gamma - \rho_{f_2})$, and thus $\sigma_k^2 = (1 - \eta_k\left(1/\gamma - \rho_{f_2}\right))/(1 - \eta_k \rho_{g_2})^2 \leq 1 - \left(\frac{\eta_k \rho_{g_2}}{1 - \eta_k \rho_{g_2}}\right)^2$. By setting $\epsilon_k = \frac{1}{4}\left(\frac{\eta_k \rho_{g_2}}{1 - \eta_k \rho_{g_2}}\right)^2$ in the above inequality, we deduce that when $\alpha_k \leq \frac{1}{2}\left(\frac{\eta_k \rho_{g_2}}{1 - \eta_k \rho_{g_2}}\right)^2$, it follows that

$$
\begin{aligned}
& \left\|\theta^{k+1} - \theta^*_\gamma(x^{k+1}, y^{k+1})\right\|^2 - \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\|^2 + \frac{\alpha_k}{2} \left\|\theta^{k+1} - \theta^*_\gamma(x^k, y^k)\right\|^2 \\
\leq & - \frac{1}{2}\left(\frac{\eta_k \rho_{g_2}}{1 - \eta_k \rho_{g_2}}\right)^2 \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\|^2 \\
& + \left(1 + 4\left(\frac{1 - \eta_k \rho_{g_2}}{\eta_k \rho_{g_2}}\right)^2\right) L_\theta^2 \left\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\right\|^2 .
\end{aligned}
\tag{48}
$$

Similarly, we can show that, when $\beta_k \leq \frac{1}{4}\left(\frac{\eta_k \rho_{g_2}}{1 - \eta_k \rho_{g_2}}\right)^2$, it holds that

$$
\begin{aligned}
& \left\|\theta^{k+1} - \theta^*_\gamma(x^{k+1}, y^{k+1})\right\|^2 - \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\|^2 + \beta_k \left\|\theta^{k+1} - \theta^*_\gamma(x^k, y^k)\right\|^2 \\
\leq & - \frac{1}{2}\left(\frac{\eta_k \rho_{g_2}}{1 - \eta_k \rho_{g_2}}\right)^2 \left\|\theta^k - \theta^*_\gamma(x^k, y^k)\right\|^2 \\
& + \left(1 + 4\left(\frac{1 - \eta_k \rho_{g_2}}{\eta_k \rho_{g_2}}\right)^2\right) L_\theta^2 \left\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\right\|^2 .
\end{aligned}
\tag{49}
$$

Combining (47), (48) and (49), we have

$$\begin{aligned}
&V_{k+1} - V_k \\
\le& -\left[\frac{1}{2\alpha_k} - \frac{L_{\phi_k}}{2} - \frac{\beta_k L_\theta^2}{\gamma^2} - \left(1 + 4\left(\frac{1-\eta_k\rho_{g_2}}{\eta_k\rho_{g_2}}\right)^2\right) L_\theta^2\left((L_f+L_g)^2 + 1/\gamma^2\right)\right]\|x^{k+1}-x^k\|^2 \\
&-\left[\frac{1}{2\beta_k} - \frac{\rho_{g_2}}{2} - \frac{L_{\phi_k}}{2} - \left(1 + 4\left(\frac{1-\eta_k\rho_{g_2}}{\eta_k\rho_{g_2}}\right)^2\right) L_\theta^2\left((L_f+L_g)^2 + 1/\gamma^2\right)\right]\|y^{k+1}-y^k\|^2 \\
&-\frac{1}{2}\left(\frac{\eta_k\rho_{g_2}}{1-\eta_k\rho_{g_2}}\right)^2\left((L_f+L_g)^2+1/\gamma^2\right)\left\|\theta^k - \theta_\gamma^*(x^k,y^k)\right\|^2.
\end{aligned} \tag{50}$$

When $c_{k+1} \ge c_k$, $\eta_k \ge \underline{\eta} > 0$, $\alpha_k \le \frac{1}{2}\left(\frac{\underline{\eta}\rho_{g_2}}{1-\underline{\eta}\rho_{g_2}}\right)^2$ and $\beta_k \le \frac{1}{4}\left(\frac{\underline{\eta}\rho_{g_2}}{1-\underline{\eta}\rho_{g_2}}\right)^2$ and , it holds that, for any k, $\alpha_k \le \frac{1}{2}\left(\frac{\eta_k\rho_{g_2}}{1-\eta_k\rho_{g_2}}\right)^2$, $\beta_k \le \frac{1}{4}\left(\frac{\eta_k\rho_{g_2}}{1-\eta_k\rho_{g_2}}\right)^2$,

$$\begin{aligned}
&\frac{L_{\phi_k}}{2} + \frac{\beta_k L_\theta^2}{\gamma^2} + \left(1 + 4\left(\frac{1-\eta_k\rho_{g_2}}{\eta_k\rho_{g_2}}\right)^2\right) L_\theta^2\left((L_f+L_g)^2+1/\gamma^2\right) \\
\le& \frac{L_{\phi_0}}{2} + \frac{L_\theta^2}{4\gamma^2}\left(\frac{\underline{\eta}\rho_{g_2}}{1-\underline{\eta}\rho_{g_2}}\right)^2 + \left(1 + 4\left(\frac{1-\underline{\eta}\rho_{g_2}}{\underline{\eta}\rho_{g_2}}\right)^2\right) L_\theta^2\left((L_f+L_g)^2+1/\gamma^2\right) =: C_\alpha,
\end{aligned} \tag{51}$$

and

$$\begin{aligned}
&\frac{\rho_{g_2}}{2} + \frac{L_{\phi_k}}{2} + \left(1 + 4\left(\frac{1-\eta_k\rho_{g_2}}{\eta_k\rho_{g_2}}\right)^2\right) L_\theta^2\left((L_f+L_g)^2+1/\gamma^2\right) \\
\le& \frac{\rho_{g_2}}{2} + \frac{L_{\phi_0}}{2} + \left(1 + 4\left(\frac{1-\underline{\eta}\rho_{g_2}}{\underline{\eta}\rho_{g_2}}\right)^2\right) L_\theta^2\left((L_f+L_g)^2+1/\gamma^2\right) =: C_\beta,
\end{aligned} \tag{52}$$

Consequently, since $\frac{1}{4C_\beta} < \frac{1}{2\rho_{g_2}}$, if $c_\alpha, c_\beta > 0$ satisfies

$$c_\alpha \le \min\left\{\frac{1}{2}\left(\frac{\underline{\eta}\rho_{g_2}}{1-\underline{\eta}\rho_{g_2}}\right)^2, \frac{1}{4C_\alpha}\right\}, \quad c_\beta \le \min\left\{\frac{1}{4}\left(\frac{\underline{\eta}\rho_{g_2}}{1-\underline{\eta}\rho_{g_2}}\right)^2, \frac{1}{4C_\beta}\right\}, \tag{53}$$

then, when $0 < \alpha_k \le c_\alpha$ and $0 < \beta_k \le c_\beta$, it holds that

$$\frac{1}{2\alpha_k} - \frac{L_{\phi_k}}{2} - \frac{\beta_k L_\theta}{\gamma^2} - \left(1 + 4\left(\frac{1-\eta_k\rho_{g_2}}{\eta_k\rho_{g_2}}\right)^2\right) L_\theta^2\left((L_f+L_g)^2+1/\gamma^2\right) \ge \frac{1}{4\alpha_k},$$

and

$$\frac{1}{2\beta_k} - \frac{\rho_{g_2}}{2} - \frac{L_{\phi_k}}{2} - \left(1 + 4\left(\frac{1-\eta_k\rho_{g_2}}{\eta_k\rho_{g_2}}\right)^2\right) L_\theta^2\left((L_f+L_g)^2+1/\gamma^2\right) \ge \frac{1}{4\beta_k}.$$

Consequently, the conclusion follows from (50). □

A.6. Proof of Theorem 3.4

By leveraging the monotonically decreasing property of the merit function V_k, we can establish the non-asymptotic convergence for the sequence (x^k, y^k, θ^k) generated by the proposed MEHA.

Theorem A.12. *Under Assumptions 3.1 and 3.2, suppose* $\gamma \in (0, \frac{1}{2\rho_{f_2}+2\rho_{g_2}})$, $c_k = \underline{c}(k+1)^p$ *with* $p \in [0, 1/2)$, $\underline{c} > 0$ *and* $\eta_k \in [\underline{\eta}, (1/\gamma - \rho_{f_2})/(L_f + 1/\gamma)^2] \cap [\underline{\eta}, 1/\rho_{g_2})$ *with* $\underline{\eta} > 0$, *then there exists* $c_\alpha, c_\beta > 0$ *such that when* $\alpha_k \in (\underline{\alpha}, c_\alpha)$ *and* $\beta_k \in (\underline{\beta}, c_\beta)$ *with* $\underline{\alpha}, \underline{\beta} > 0$, *the sequence of* (x^k, y^k, θ^k) *generated by Algorithm 1 satisfies*

$$\min_{0\le k\le K}\left\|\theta^k - \theta_\gamma^*(x^k,y^k)\right\| = O\left(\frac{1}{K^{1/2}}\right),$$

and

$$\min_{0\le k\le K} R_k(x^{k+1},y^{k+1}) = O\left(\frac{1}{K^{(1-2p)/2}}\right).$$

Furthermore, if $p\in(0,1/2)$ and there exists $M>0$ such that $\psi_{c_k}(x^k,y^k)\le M$ for any k, the sequence of (x^k,y^k) satisfies

$$\varphi(x^K,y^K)-v_\gamma(x^K,y^K) = O\left(\frac{1}{K^p}\right).$$

Proof. First, Lemma 3.6 ensures the existence of $c_\alpha, c_\beta>0$ for which (17) is valid under the conditions $\alpha_k\le c_\alpha$, $\beta_k\le c_\beta$. Upon telescoping (17) over the range $k=0,1,\dots,K-1$, we derive

$$\begin{aligned}&\sum_{k=0}^{K-1}\Bigg(\frac{1}{4\alpha_k}\|x^{k+1}-x^k\|^2+\frac{1}{4\beta_k}\|y^{k+1}-y^k\|^2\\ &\qquad+\frac{1}{2}\left(\frac{\underline{\eta}\rho_{g_2}}{1-\underline{\eta}\rho_{g_2}}\right)^2\left((L_f+L_g)^2+1/\gamma^2\right)\left\|\theta^k-\theta^*_\gamma(x^k,y^k)\right\|^2\Bigg)\\ &\le V_0-V_K\le V_0,\end{aligned}\tag{54}$$

where the last inequality is valid because V_K is nonnegative. Thus, we have

$$\sum_{k=0}^{\infty}\left\|\theta^k-\theta^*_\gamma(x^k,y^k)\right\|^2<\infty,$$

and then

$$\min_{0\le k\le K}\left\|\theta^k-\theta^*_\gamma(x^k,y^k)\right\| = O\left(\frac{1}{K^{1/2}}\right).$$

According to the update rule of variables (x,y) as defined in (11) and (13), we have that

$$\begin{aligned}&0\in c_kd_x^k+\mathcal{N}_X(x^{k+1})+\frac{c_k}{\alpha_k}\left(x^{k+1}-x^k\right),\\ &0\in c_kd_y^k+c_k\partial_y g(x^{k+1},y^{k+1})+\mathcal{N}_Y(y^{k+1})+\frac{c_k}{\beta_k}\left(y^{k+1}-y^k\right).\end{aligned}\tag{55}$$

From the definitions of d_x^k and d_y^k provided in (10) and (12), and given $\nabla_x g(x^{k+1},y^{k+1})\times\partial_y g(x^{k+1},y^{k+1})\subseteq\partial g(x^{k+1},y^{k+1})$, a result stemming from the weakly convexity of g and its continuously differentiable property with respect to x as outlined in Assumption 3.2 and corroborated by Theorem 5 of (Gao et al., 2023), we deduce

$$\begin{aligned}(e_x^k,e_y^k)\in&\nabla F(x^{k+1},y^{k+1})+c_k\left(\nabla f(x^{k+1},y^{k+1})+\partial g(x^{k+1},y^{k+1})-\nabla v_\gamma(x^{k+1},y^{k+1})\right)\\ &+\mathcal{N}_{X\times Y}(x^{k+1},y^{k+1}),\end{aligned}$$

with

$$\begin{aligned}&e_x^k:=\nabla_x\psi_{c_k}(x^{k+1},y^{k+1})-c_kd_x^k-\frac{c_k}{\alpha_k}\left(x^{k+1}-x^k\right),\\ &e_y^k:=\nabla_y\left(\psi_{c_k}-c_kg\right)(x^{k+1},y^{k+1})-c_kd_y^k-\frac{c_k}{\beta_k}\left(y^{k+1}-y^k\right).\end{aligned}\tag{56}$$

Next, we estimate $\|e_x^k\|$. We have

$$\|e_x^k\|\le\|\nabla_x\psi_{c_k}(x^{k+1},y^{k+1})-\nabla_x\psi_{c_k}(x^k,y^k)\|+\|\nabla_x\psi_{c_k}(x^k,y^k)-c_kd_x^k\|+\frac{c_k}{\alpha_k}\left\|x^{k+1}-x^k\right\|.$$

Considering the first term on the right hand side of the preceding inequality, and invoking Assumptions 3.1 and 3.2 alongside Lemma A.5, A.6, A.8, we establish the existence of $L_{\psi_1}>0$ such that

$$\|\nabla_x\psi_{c_k}(x^{k+1},y^{k+1})-\nabla_x\psi_{c_k}(x^k,y^k)\|\le c_kL_{\psi_1}\|(x^{k+1},y^{k+1})-(x^k,y^k)\|.$$

Using (38) and Lemma A.9, we deduce

$$\|\nabla_x \psi_{c_k}(x^k,y^k) - c_k d_x^k\| = c_k \left\|\nabla_x \phi_{c_k}(x^k,y^k) - d_x^k\right\| \le c_k(L_f+L_g)\left\|\theta^k - \theta_\gamma^*(x^k,y^k)\right\|. \tag{57}$$

Hence, we have

$$\|e_x^k\| \le c_k L_{\psi_1}\|(x^{k+1},y^{k+1})-(x^k,y^k)\| + \frac{c_k}{\alpha_k}\left\|x^{k+1}-x^k\right\| + c_k(L_f+L_g)\left\|\theta^k - \theta_\gamma^*(x^k,y^k)\right\|.$$

For $\|e_y^k\|$, it follows that

$$\begin{aligned}\|e_y^k\| \le& \|\nabla_y\left(\psi_{c_k}-c_k g\right)(x^{k+1},y^{k+1}) - \nabla_y\left(\psi_{c_k}-c_k g\right)(x^{k+1},y^k)\| + \frac{c_k}{\beta_k}\left\|y^{k+1}-y^k\right\| \\ &+ \|\nabla_y\left(\psi_{c_k}-c_k g\right)(x^{k+1},y^k) - c_k d_y^k\|.\end{aligned}$$

Analogously, invoking Assumptions 3.1 and 3.2 together with Lemmas A.5, A.6, and A.8, we have the existence of $L_{\psi_2} := L_F + L_f + \frac{1}{\gamma} + L_\theta > 0$ such that

$$\|\nabla_y\left(\psi_{c_k}-c_k g\right)(x^{k+1},y^{k+1}) - \nabla_y\left(\psi_{c_k}-c_k g\right)(x^{k+1},y^k)\| \le c_k L_{\psi_2}\|y^{k+1}-y^k\|.$$

Using (43), Lemma A.8 and Lemma A.9, we obtain

$$\begin{aligned}\|\nabla_y\left(\psi_{c_k}-c_k g\right)(x^{k+1},y^k) - c_k d_y^k\| =& c_k\left\|\nabla_y\left(\phi_{c_k}-g\right)(x^{k+1},y^k) - d_y^k\right\| \\ \le& \frac{c_k}{\gamma}\left(\left\|\theta^k-\theta_\gamma^*(x^k,y^k)\right\| + L_\theta\|x^{k+1}-x^k\|\right).\end{aligned}$$

Therefore, we have

$$\|e_y^k\| \le c_k L_{\psi_2}\|y^{k+1}-y^k\| + \frac{c_k}{\beta_k}\left\|y^{k+1}-y^k\right\| + \frac{c_k}{\gamma}\left(\left\|\theta^k-\theta_\gamma^*(x^k,y^k)\right\| + L_\theta\|x^{k+1}-x^k\|\right).$$

With the estimations of $\|e_x^k\|$ and $\|e_y^k\|$, we obtain the existence of $L_\psi > 0$ such that

$$\begin{aligned}R_k(x^{k+1},y^{k+1}) \le& c_k L_\psi\|(x^{k+1},y^{k+1})-(x^k,y^k)\| + \left(\frac{c_k}{\alpha_k}+\frac{c_k L_\theta}{\gamma}\right)\left\|x^{k+1}-x^k\right\| \\ &+ \frac{c_k}{\beta_k}\left\|y^{k+1}-y^k\right\| + c_k(L_f+L_g+\frac{1}{\gamma})\left\|\theta^k-\theta_\gamma^*(x^k,y^k)\right\|.\end{aligned}$$

Employing the aforementioned inequality and given that $\alpha_k \ge \underline{\alpha}$ and $\beta_k \ge \underline{\beta}$ for some positive constants $\underline{\alpha}, \underline{\beta}$, we demonstrate the existence of $C_R > 0$ such that

$$\begin{aligned}&\frac{1}{c_k^2}R_k(x^{k+1},y^{k+1})^2 \\ \le& C_R\Big(\frac{1}{4\alpha_k}\|x^{k+1}-x^k\|^2 + \frac{1}{4\beta_k}\|y^{k+1}-y^k\|^2 \\ &+ \frac{1}{2}\left(\frac{\underline{\eta}\rho_{g_2}}{1-\underline{\eta}\rho_{g_2}}\right)^2\left((L_f+L_g)^2+1/\gamma^2\right)\left\|\theta^k-\theta_\gamma^*(x^k,y^k)\right\|^2\Big).\end{aligned} \tag{58}$$

Combining this with (54) implies that

$$\sum_{k=0}^{\infty}\frac{1}{c_k^2}R_k(x^{k+1},y^{k+1})^2 < \infty. \tag{59}$$

Because $2p < 1$, it holds that

$$\sum_{k=0}^{K}\frac{1}{c_k^2} = \frac{1}{\underline{c}^2}\sum_{k=0}^{K}\left(\frac{1}{k+1}\right)^{2p} \ge \frac{1}{\underline{c}^2}\int_1^{K+2}\frac{1}{t^{2p}}dt \ge \frac{(K+2)^{1-2p}-1}{(1-2p)\underline{c}^2},$$

and we can conclude from (59) that

$$\min_{0\le k\le K} R_k(x^{k+1},y^{k+1}) = O\left(\frac{1}{K^{(1-2p)/2}}\right).$$

In fact, since

$$\sum_{k=\lfloor K/2\rfloor-1}^{K}\frac{1}{c_k^2} = \frac{1}{\underline{c}^2}\sum_{k=\lfloor K/2\rfloor-1}^{K}\left(\frac{1}{k+1}\right)^{2p} \ge \frac{1}{\underline{c}^2}\int_{\lfloor K/2\rfloor}^{2\lfloor K/2\rfloor}\frac{1}{t^{2p}}dt \ge \frac{2^{1-2p}-1}{(1-2p)\underline{c}^2}\lfloor K/2\rfloor^{1-2p},$$

we have

$$\min_{\lfloor K/2\rfloor-1\le k\le K} R_k(x^{k+1},y^{k+1}) = O\left(\frac{1}{K^{(1-2p)/2}}\right). \tag{60}$$

Finally, since $\psi_{c_k}(x^k,y^k)\le M$ and $F(x^k,y^k)\ge \underline{F}$ for any k, we have

$$c_k\Big(\varphi(x^k,y^k)-v_\gamma(x^k,y^k)\Big)\le M-\underline{F}, \quad \forall k,$$

and we can obtain from $c_k=\underline{c}(k+1)^p$ that

$$\varphi(x^K,y^K)-v_\gamma(x^K,y^K) = O\left(\frac{1}{K^p}\right).$$

□

A.7. Proof of Theorem 3.7

In the typical scenario where $\varphi(x,y)=f(x,y)$ being a strongly convex smooth function in y for each x, and $X=\mathbb{R}^n$, $Y=\mathbb{R}^m$, it can be established that the residual function $R_k(x,y)$ is closely related to the norm of the hyper-gradient $\nabla\Phi(x)$ associated with the hyper-objective $\Phi(x):=F(x,y^*(x))$, where $y^*(x)$ is the unique LL optimal solution.

Recall that this residual function is a stationarity measure for the following penalized version of the constrained problem (3), with c_k serving as the penalty parameter:

$$\min_{(x,y)\in X\times Y}\psi_{c_k}(x,y) := F(x,y)+c_k\big(f(x,y)-v_\gamma(x,y)\big).$$

By the expression (8) of $\nabla v_\gamma(x,y)$ and the optimality condition of $\theta_\gamma^* := \theta_\gamma^*(x,y)$ given as

$$\nabla_y f(x,\theta_\gamma^*)+\frac{\theta_\gamma^*-y}{\gamma}=0,$$

the residual function $R_k(x,y)$ in (15) becomes

$$R_k(x,y)=\left\|\begin{pmatrix}\nabla_x F(x,y)+c_k\left[\nabla_x f(x,y)-\nabla_x f(x,\theta_\gamma^*(x,y))\right]\\ \nabla_y F(x,y)+c_k\left[\nabla_y f(x,y)-\nabla_y f(x,\theta_\gamma^*(x,y))\right]\end{pmatrix}\right\| =: \left\|\begin{pmatrix}R_k^{(1)}(x,y)\\ R_k^{(2)}(x,y)\end{pmatrix}\right\|.$$

Lemma A.13. *Under Assumptions 3.1 and 3.2, suppose that $X=\mathbb{R}^n$, $Y=\mathbb{R}^m$, and the lower-level objective $\varphi(x,\cdot)=f(x,\cdot)$ is a μ-strongly convex smooth function for each x. Let $\gamma>1/\mu$, then*

$$\|y-y^*(x)\|\le\frac{4L_{F,0}+4\|R_k^{(2)}(x,y)\|}{c_k\mu}, \tag{61}$$

where $L_{F,0}$ is the upper bound of $\|\nabla_y F(x,y^*(x))\|$ on X. Additionally, suppose $\|R_k^{(2)}(x,y)\|\le L_{F,0}$, then $c_k\|y-y^*(x)\|\le 8L_{F,0}/\mu$. If further $\nabla_{xy}^2 f(x,\cdot)$, $\nabla_{yy}^2 f(x,\cdot)$ are $L_{f,2}$-Lipschitz continuous on $X\times Y$, then

$$\|\nabla\Phi(x)-R_k^{(1)}(x,y)\|\le\frac{L_\mu}{c_k}+\frac{L_f}{\mu}\|R_k^{(2)}(x,y)\|, \tag{62}$$

where $L_\mu := \frac{8L_{F,0}}{\mu}\left(1+\frac{L_f}{\mu}\right)\left(L_F+\frac{8L_{f,2}L_{F,0}}{\mu}\right)+\frac{8L_\Phi L_{F,0}}{\mu^2\gamma}$.

Proof. First, by the optimality condition of $\theta_\gamma^* := \theta_\gamma^*(x,y)$, we have

$$\left[\nabla_{yy}^2 f(x,\theta_\gamma^*)\right] \nabla_y \theta_\gamma^* + \frac{1}{\gamma}\left(\nabla_y \theta_\gamma^* - I\right) = 0,$$

which implies that $\nabla_y \theta_\gamma^* = \frac{1}{\gamma}\left[\nabla_{yy}^2 f(x,\theta_\gamma^*) + \frac{1}{\gamma} I\right]^{-1}$. Note that $L_f I \succeq \nabla_{yy}^2 f \succeq \mu I$. Then $\frac{1}{\gamma}\left(\mu + \frac{1}{\gamma}\right)^{-1} I \succeq \nabla_y \theta_\gamma^* \succeq \frac{1}{\gamma}\left(L_f + \frac{1}{\gamma}\right)^{-1} I$. Thus $\nabla_{yy}^2 v_\gamma = \frac{1}{\gamma}(I - \nabla_y \theta_\gamma^*) \preceq \frac{\mu}{\mu\gamma+1} I$. This implies that $f(x,\theta) - v_\gamma(x,\theta)$ is $\mu/2$-strongly convex in θ when $\gamma > 1/\mu$. Thus, $\arg\min_\theta \big(f(x,\theta) - v_\gamma(x,\theta)\big)$ has a unique solution, denoted by $y_\gamma^*(x)$. We claim that $y_\gamma^*(x) = y^*(x) =: y^*$. Here, we denote $y^*(x)$ as y^* for brevity. Indeed, the expression (8) of $\nabla v_\gamma(x,y)$ shows that

$$\nabla_y (f - v_\gamma)(x,y^*) = \nabla_y f(x,y^*) - \frac{1}{\gamma}\big(y^* - \theta_\gamma^*(x,y^*)\big) = 0,$$

by using the facts $\nabla_y f(x,y^*) = 0$ and $\theta_\gamma^*(x,y^*) = y^*$.

Second, we demonstrate the validity of estimate (61). Since $f(x,\theta) - v_\gamma(x,\theta)$ is a $\mu/2$-strongly convex smooth function in θ, the function $\psi_{c_k}(x,\theta) = F(x,\theta) + c_k\big(f(x,\theta) - v_\gamma(x,\theta)\big)$ is $(-L_F + c_k\mu/2)$-strongly convex in θ. Let $c_k > 4L_F/\mu$, then it is $c_k\mu/4$-strongly convex. Hence, $\arg\min_{\theta\in\mathbb{R}^m} \psi_{c_k}(x,\theta)$ has a unique solution, denoted by $y_{c_k}^*(x)$. By the first-order optimality condition, we have

$$\nabla_y F(x, y_{c_k}^*(x)) + c_k\big(\nabla_y f(x, y_{c_k}^*(x)) - \nabla_y v_\gamma(x, y_{c_k}^*(x))\big) = 0.$$

By the coercivity property of $\mu/2$-strongly convex functions $f(x,\cdot) - v_\gamma(x,\cdot)$,

$$\big\langle \nabla_y (f - v_\gamma)(x, y^*(x)) - \nabla_y (f - v_\gamma)(x, y_{c_k}^*(x)), \theta_\gamma^*(x) - y_{c_k}^*(x) \big\rangle \geq \frac{\mu}{2}\|\theta_\gamma^*(x) - y_{c_k}^*(x)\|^2.$$

Recall that $\nabla_y (f - v_\gamma)(x, y^*(x)) = 0$, we have

$$\|y^*(x) - y_{c_k}^*(x)\| \leq \frac{2}{\mu}\|\nabla_y (f - v_\gamma)(x, y_{c_k}^*(x))\| \leq \frac{2}{c_k\mu}\|\nabla_y F(x, y_{c_k}^*(x))\|.$$

On the other hand, by the definition of $y^*(x)$,

$$\|\nabla_y F(x, y_{c_k}^*(x))\| \leq \|\nabla_y F(x, y_{c_k}^*(x)) - \nabla_y F(x, y^*(x))\| + \|\nabla_y F(x, y^*(x))\| \leq L_F\|y_{c_k}^*(x) - y^*(x)\| + L_{F,0}.$$

Hence, we obtain

$$\|y^*(x) - y_{c_k}^*(x)\| \leq \frac{2L_{F,0}}{c_k\mu(1 - \frac{2L_F}{c_k\mu})}. \tag{63}$$

Since $c_k > 4L_F/\mu$, we derive

$$\|y^*(x) - y_{c_k}^*(x)\| \leq \frac{4L_{F,0}}{c_k\mu}. \tag{64}$$

Similarly, by the coercivity property of $c_k\mu/4$-strongly convex functions $F(x,\cdot) + c_k\big(f(x,\cdot) - v_\gamma(x,\cdot)\big)$,

$$\big\langle \nabla_y F(x,y) + c_k\big(\nabla_y f(x,y) - \nabla_y v_\gamma(x,y)\big), y - y_{c_k}^*(x) \big\rangle \geq \frac{c_k\mu}{4}\|y - y_{c_k}^*(x)\|^2.$$

This implies that

$$\|y - y_{c_k}^*(x)\| \leq \frac{4}{c_k\mu}\|\nabla_y F(x,y) + c_k\big(\nabla_y f(x,y) - \nabla_y v_\gamma(x,y)\big)\| = \frac{4\|R_k^{(2)}(x,y)\|}{c_k\mu}. \tag{65}$$

By combining the estimates (64) and (65), we have successfully established estimate (61). Hence, when $\|R_k^{(2)}(x,y)\| \leq L_{F,0}$, we get $c_k\|y - y^*(x)\| \leq 8L_{F,0}/\mu$.

Next, we shall validate the estimate (62). Recall that

$$\nabla\Phi(x) = \nabla_x F(x,y^*) - \nabla_{xy}^2 f(x,y^*)\left[\nabla_{yy}^2 f(x,y^*)\right]^{-1} \nabla_y F(x,y^*). \tag{66}$$

Define

$$\bar{\nabla}\Phi(x,y) = \nabla_x F(x,\theta_\gamma^*) - \nabla_{xy}^2 f(x,\theta_\gamma^*)\left[\nabla_{yy}^2 f(x,\theta_\gamma^*)\right]^{-1}\nabla_y F(x,\theta_\gamma^*). \tag{67}$$

Then by Lemma 2.2 of (Ghadimi & Wang, 2018), there is a positive constant L_Φ such that

$$\|\nabla\Phi(x) - \bar{\nabla}\Phi(x,y)\| \le L_\Phi \|y^*(x) - \theta_\gamma^*(x,y)\|. \tag{68}$$

By $\nabla_y f(x,\theta_\gamma^*) + \frac{1}{\gamma}(\theta_\gamma^* - y) = 0$ and the strongly convexity of $f(x,\cdot)$, we get

$$\left\langle \frac{1}{\gamma}(\theta_\gamma^* - y), y^* - \theta_\gamma^* \right\rangle \ge \mu\|y^* - \theta_\gamma^*\|^2,$$

which implies that

$$\|y^* - \theta_\gamma^*\| \le \frac{1}{\mu\gamma}\|\theta_\gamma^* - y\|. \tag{69}$$

We claim that $\|\theta_\gamma^* - y\| \le \|y - y^*\|$. In fact, by the optimality of θ_γ^*, we have

$$f(x,\theta_\gamma^*) + \frac{1}{2\gamma}\|\theta_\gamma^* - y\|^2 \le f(x,y^*) + \frac{1}{2\gamma}\|y^* - y\|^2.$$

The desired result follows from the fact that $f(x,\theta_\gamma^*) - f(x,y^*) \ge \frac{\mu}{2}\|\theta_\gamma^* - y^*\|^2 \ge 0$. Hence,

$$\|\nabla\Phi(x) - \bar{\nabla}\Phi(x,y)\| \le \frac{L_\Phi}{\mu\gamma}\|y - y^*(x)\|. \tag{70}$$

Since $R_k^{(1)}(x,y) = \nabla_x F(x,y) + c_k\left[\nabla_x f(x,y) - \nabla_x f(x,\theta_\gamma^*(x))\right]$, we have

$$\begin{aligned}
&\bar{\nabla}\Phi(x,y) - R_k^{(1)}(x,y)\\
=&\nabla_x F(x,\theta_\gamma^*) - \nabla_x F(x,y)\\
&- \nabla_{xy}^2 f(x,\theta_\gamma^*)\left[\nabla_{yy}^2 f(x,\theta_\gamma^*)\right]^{-1}\left(\nabla_y F(x,\theta_\gamma^*) - \nabla_y F(x,y)\right)\\
&- \nabla_{xy}^2 f(x,\theta_\gamma^*)\left[\nabla_{yy}^2 f(x,\theta_\gamma^*)\right]^{-1}\left(\nabla_y F(x,y) + c_k\big(\nabla_y f(x,y) - \nabla_y f(x,\theta_\gamma^*)\big)\right)\\
&+ c_k\nabla_{xy}^2 f(x,\theta_\gamma^*)\left[\nabla_{yy}^2 f(x,\theta_\gamma^*)\right]^{-1}\left(\nabla_y f(x,y) - \nabla_y f(x,\theta_\gamma^*) - \nabla_{yy}^2 f(x,\theta_\gamma^*)(y - \theta_\gamma^*)\right)\\
&- c_k\left(\nabla_x f(x,y) - \nabla_x f(x,\theta_\gamma^*) - \nabla_{xy}^2 f(x,\theta_\gamma^*)(y - \theta_\gamma^*)\right).
\end{aligned}$$

By Assumptions 3.1 and 3.2 (i), we have

$$\begin{aligned}
\|\nabla_x F(x,y^*) - \nabla_x F(x,y)\| \le& L_F\|y^* - y\|,\\
\|\nabla_y F(x,y^*) - \nabla_y F(x,y)\| \le& L_F\|y^* - y\|.
\end{aligned}$$

If further $\nabla_{xy}^2 f(x,\cdot)$, $\nabla_{yy}^2 f(x,\cdot)$ are $L_{f,2}$-Lipschitz continuous on $X \times Y$, then

$$\begin{aligned}
\|\nabla_y f(x,y) - \nabla_y f(x,y^*) - \nabla_{yy}^2 f(x,y^*)(y - y^*)\| \le& L_{f,2}\|y - y^*\|^2,\\
\|\nabla_x f(x,y) - \nabla_x f(x,y^*) - \nabla_{xy}^2 f(x,y^*)(y - y^*)\| \le& L_{f,2}\|y - y^*\|^2.
\end{aligned}$$

Therefore, by the μ-strongly convex of $f(x,\cdot)$ and $c_k\|y - y^*(x)\| \le 8L_{F,0}/\mu$, we have

$$\begin{aligned}
\|\bar{\nabla}\Phi(x,y) - R_k^{(1)}(x,y)\| \le& L_F\|y^* - y\| + \frac{L_f}{\mu}L_F\|y^* - y\| + \frac{L_f}{\mu}\|R_k^{(2)}(x,y)\|\\
&+ c_k\frac{L_f}{\mu}L_{f,2}\|y - y^*\|^2 + c_k L_{f,2}\|y - y^*\|^2\\
\le&\left(1 + \frac{L_f}{\mu}\right)\|y^* - y\|\left(L_F + L_{f,2}c_k\|y^* - y\|\right) + \frac{L_f}{\mu}\|R_k^{(2)}(x,y)\|\\
\le&\left(1 + \frac{L_f}{\mu}\right)\|y^* - y\|\left(L_F + \frac{8L_{f,2}L_{F,0}}{\mu}\right) + \frac{L_f}{\mu}\|R_k^{(2)}(x,y)\|\\
\le&\frac{8L_{F,0}}{\mu}\left(1 + \frac{L_f}{\mu}\right)\left(L_F + \frac{8L_{f,2}L_{F,0}}{\mu}\right)\frac{1}{c_k} + \frac{L_f}{\mu}\|R_k^{(2)}(x,y)\|.
\end{aligned}$$

Therefore,

$$\begin{aligned}\|\nabla\Phi(x) - R_k^{(1)}(x,y)\| \le & \|\bar{\nabla}\Phi(x,y) - R_k^{(1)}(x,y)\| + \|\nabla\Phi(x) - \bar{\nabla}\Phi(x,y)\| \\ \le & \frac{8L_{F,0}}{\mu}\left(1+\frac{L_f}{\mu}\right)\left(L_F + \frac{8L_{f,2}L_{F,0}}{\mu}\right)\frac{1}{c_k} + \frac{L_f}{\mu}\|R_k^{(2)}(x,y)\| + \frac{8L_\Phi L_{F,0}}{\mu^2\gamma}\frac{1}{c_k},\end{aligned}$$

which implies the desired outcome and thereby concludes the proof. □

We can obtain Theorem 3.7 by combining Lemma A.13 and Theorem 3.4.

A.8. Verifying Assumptions 3.2(ii) and (iii) are special cases of Assumption 3.2(iv)

In this section, we prove that Assumptions 3.2(ii) and (iii) are special cases of Assumption 3.2(iv).

Lemma A.14. *The function $g(x,y) = x\|y\|_1$ satisfies Assumption 3.2(iv) when $X = \mathbb{R}_+$ and $Y = \mathbb{R}^m$.*

Proof. Initially, as depicted in Section 6.1 of (Gao et al., 2023), for $x \in \mathbb{R}_+$ and $y \in \mathbb{R}^m$,

$$x\|y\|_1 + \frac{\sqrt{p}}{2}x^2 + \frac{\sqrt{p}}{2}\|y\|^2 = \sum_{i=1}^m \frac{1}{2\sqrt{p}}\left(x + \sqrt{p}|y_i|\right)^2, \tag{71}$$

which is convex with respect to $(x,y) \in \mathbb{R}_+ \times \mathbb{R}^m$. Consequently, $g(x,y)$ is $\sqrt{p}$-weakly convex. Further, for any given $s \in (0, \bar{s}]$, we have

$$\mathrm{Prox}_{s\tilde{g}(x,\cdot)}(\theta) = \mathrm{Prox}_{sx\|\cdot\|_1}(\theta) = \mathcal{T}_{sx}(\theta) = \left(\mathcal{T}_{sx}(\theta_i)\right)_{i=1}^m = \left([|\theta_i| - sx]_+ \cdot \mathrm{sgn}(\theta_i)\right)_{i=1}^m. \tag{72}$$

This results in

$$\left\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}(\theta) - \mathrm{Prox}_{s\tilde{g}(x',\cdot)}(\theta)\right\| \le s\|x - x'\| \le \bar{s}\|x - x'\|. \tag{73}$$

In summary, Assumption 3.2(iv) is satisfied by $g(x,y) = x\|y\|_1$ when $X = \mathbb{R}_+$ and $Y = \mathbb{R}^m$. □

Lemma A.15. *The function $g(x,y) = \sum_{j=1}^J x_j\|y^{(j)}\|_2$, where $\{1,\dots,m\}$ is divided into J groups, $y^{(j)}$ denotes the corresponding j-th group of y, satisfies Assumption 3.2(iv) when $X = \mathbb{R}_+^J$ and $Y = \mathbb{R}^m$.*

Proof. Initially, as depicted in Section 6.2 of (Gao et al., 2023), for $x \in \mathbb{R}_+^J$ and $y \in \mathbb{R}^m$,

$$\sum_{j=1}^J x_j\|y^{(j)}\|_2 + \sum_{j=1}^J \frac{1}{2}x_j^2 + \frac{1}{2}\|y\|_2^2 = \sum_{j=1}^J \frac{1}{2}\left(x_j + \|y^{(j)}\|_2\right)^2, \tag{74}$$

which is convex with respect to $(x,y) \in \mathbb{R}_+^J \times \mathbb{R}^m$. Consequently, $g(x,y)$ is 1-weakly convex. Further, for any given $s \in (0, \bar{s}]$, we have

$$\begin{aligned}\mathrm{Prox}_{s\tilde{g}(x,\cdot)}(\theta) &= \mathrm{Prox}_{s\sum_{j=1}^J x_j\|\cdot^{(j)}\|_2}(\theta) \\ &= \bigtimes_{j=1}^J \mathrm{Prox}_{sx_j\|\cdot\|_2}(\theta^{(j)}) \\ &= \bigtimes_{j=1}^J \begin{cases} [\|\theta^{(j)}\|_2 - sx_j]_+ \frac{\theta^{(j)}}{\|\theta^{(j)}\|_2}, & \theta^{(j)} \neq 0, \\ 0, & \theta^{(j)} = 0. \end{cases}\end{aligned} \tag{75}$$

This results in

$$\left\|\mathrm{Prox}_{s\tilde{g}(x,\cdot)}(\theta) - \mathrm{Prox}_{s\tilde{g}(x',\cdot)}(\theta)\right\| \le s\|x - x'\| \le \bar{s}\|x - x'\|. \tag{76}$$

In summary, Assumption 3.2 (iv) is satisfied by $g(x,y) = \sum_{j=1}^J x_j\|y^{(j)}\|_2$ when $X = \mathbb{R}_+^J$ and $Y = \mathbb{R}^m$. □

A.9. Expanded Related Work

In this section, we provide an extensive review of recent studies closely related to ours.

Nonconvex-Convex BLO. The LL strong convexity significantly contributes to the development of efficient BLO algorithms, see, e.g., (Ghadimi & Wang, 2018; Ji et al., 2020b; Chen et al., 2021; Ji et al., 2022; Hong et al., 2023; Kwon et al., 2023b). It guarantees the uniqueness of the LL minimizer (Lower-Level Singleton), which facilitates the demonstration of asymptotic convergence for the iterative differentiation-based approach (Franceschi et al., 2018). If further the LL objective is twice differentiable, the gradient of the UL objective (hyper-gradient) can be expressed using the implicit function theorem. Then the uniformly LL strong convexity implies both the smoothness and the Lipschitz continuity properties of the LL solution mapping. These essential properties facilitates the demonstration of non-asymptotic convergence for both the iterative differentiation and the approximate implicit differentiation approaches with rapid convergence rates, see e.g., (Ghadimi & Wang, 2018; Ji et al., 2020b; Chen et al., 2021; Sow et al., 2022b; Ji et al., 2022; Ji & Liang, 2022; Arbel & Mairal, 2022a; Li et al., 2022; Dagréou et al., 2022; Hong et al., 2023; Yang et al., 2023a). Due to the implicit gradient, the methods mentioned above necessitate costly manipulation involving the Hessian matrix, making them all second-order methods. Recently, (Kwon et al., 2023b) developed stochastic and deterministic fully first-order BLO algorithms based on the value function approach (Ye & Zhu, 1995), and established their non-asymptotic convergence guarantees, while an improved convergence analysis is provided in the recent work (Chen et al., 2023). By using a projection-aided finite-difference Hessian/Jacobian-vector approximation, and momentum-based updates, a simple fully single-loop Hessian/Jacobian-free stochastic BLO algorithm has been proposed in (Yang et al., 2023b) with an $\widetilde{O}(\epsilon^{-1.5})$ sample complexity.

In the absence of strong convexity, additional challenges may arise, including the presence of multiple LL solutions (Non-Singleton), which can hinder the application of implicit-based approaches involved in the study of nonconvex-strongly-convex BLOs. To tackle Non-Singleton, sequential averaging methods (also referred to as aggregation methods) were proposed in (Liu et al., 2020; Li et al., 2020; Liu et al., 2022; 2023b). Recent advances include value function based difference- of-convex algorithm (Gao et al., 2022; Ye et al., 2023); primal-dual algorithms (Sow et al., 2022a); first-order penalty methods using a novel minimax optimization reformulation (Lu & Mei, 2023).

Nonconvex-Nonconvex BLO. While the nonconvex-convex BLO has been extensively studied in the literature, the efficient methods for nonconvex-nonconvex BLO remain under-explored. Beyond the LL convexity, the authors in (Liu et al., 2021c) develop a method with initialization auxiliary and pessimistic trajectory truncation; the study (Arbel & Mairal, 2022b) extends implicit differentiation to a class of nonconvex LL functions with possibly degenerate critical points and then develops unrolled optimization algorithms. However, these works requires second-order gradient information and do not provide finite-time convergence guarantees. Still with the second-order gradient information but providing non-asymptotic analysis, the recent works (Huang, 2023b) and (Xiao et al., 2023) propose a momentum-based BLO algorithm and a generalized alternating method for BLO with a nonconvex LL objective that satisfies the Polyak-Łojasiewicz (PL) condition, respectively. In contrast to these methods discussed above, the value function reformulation of BLO was firstly utilized in (Liu et al., 2021b) to develop BLO algorithms in machine learning, using an interior-point method combined with a smoothed approximation. But it lacks a complete non-asymptotic analysis. Subsequently, (Ye et al., 2022) introduced a fully first-order value function based BLO algorithm. They also established the non-asymptotic convergence results when the LL objective satisfies the PL or local PL conditions. Recently, (Shen & Chen, 2023) proposed a penalty-based fully first-order BLO algorithm and established its finite-time convergence under the PL conditions. Notably, this work relaxed the relatively restrictive assumption on the boundedness of both the UL and LL objectives that was present in (Ye et al., 2022). Other recent advances include penalty method and first-order stochastic approximation in (Kwon et al., 2023a), which primarily explores BLO from the perspective of the hyper-objective, using a penalty value function-based approach; efficient adaptive projection-aid gradient methods based on mirror descent in (Huang, 2023a) for both deterministic and stochastic BLO problems.

Nonsmooth BLO. Despite plenty of research focusing on smooth BLOs, there are relatively fewer studies addressing nonsmooth BLOs, see, e.g., (Mairal et al., 2011; Okuno et al., 2018; Bertrand et al., 2020; 2022). However, these works typically deal with special nonsmooth LL problems, e.g., task-driven dictionary learning with elastic-net (involving the ℓ_1-norm) in (Mairal et al., 2011); the Lasso-type models (including the ℓ_1-norm as well) for hyper-parameter optimization in (Bertrand et al., 2020); ℓ_p-hyperparameter learning with $0 < p < 1$ in (Okuno et al., 2018); non-smooth convex learning with separable non-smooth terms in (Bertrand et al., 2022). Recently, there is a number of works studying BLOs with general nonsmooth LL problems. By decoupling hyperparameters from the regularization, based on the value function approach, (Gao et al., 2022) develop a sequentially convergent Value Function-based Difference-of-Convex Algorithm with inexactness for a specific class of bi-level hyper-parameter selection problems. (Gao et al., 2023) introduces a Moreau envelope-based reformulation of BLOs and develops an inexact proximal Difference-of-weakly-Convex algorithm with sequential convergence, to substantially weaken the underlying assumption in (Ye et al., 2023) from lower level full convexity to weak convexity. There is also a line of works devoted to tackle the nonsmooth UL setting, including: Bregman distance-based method in (Huang et al., 2022); proximal gradient-type algorithm in (Chen et al., 2022).

A.10. Additional Experimental Results

LL Strongly Convex Case. We initially present the convergence results using the toy numerical problem introduced in BDA (Liu et al., 2020) with a lower-level convex objective, expressed as:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x-z_0\|^2 + \frac{1}{2}y^*(x)^\top Ay^*(x) \text{ s.t. } y^*(x) = \arg\min_{y\in\mathbb{R}^n} f(x,y) = \frac{1}{2}y^\top Ay - x^\top y, \tag{77}$$

where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$. We define A has the positive-definite symmetric property and $A \in \mathbb{S}^{n\times n}$, $z_0 \neq 0$ and $z_0 \in \mathbb{R}^n$. Concretely, we set $A = I$ and $z_0 = e$. Thus, the optimal solution is $x^* = y^* = e/2$, where e represents the vector containing all elements equal to one.

This case notably aligns with several convergence assumptions of BLO methods, covering Explicit Gradient-Based Methods (EGBMs) like RHG (Franceschi et al., 2017), BDA, and IAPTT (Liu et al., 2021c), Implicit Gradient-Based Methods (IGBMs) such as CG (Pedregosa, 2016) and NS (Rajeswaran et al., 2019), and contemporary proposed methods (BRC (Liu et al., 2023a), BOME (Ye et al., 2022), F2SA (Kwon et al., 2023b), BAMM (Liu et al., 2023b)). Detailed numerical comparisons are provided in Table 8 in the Appendix, with visual comparisons in Figure 4. Analyzing the behaviors in Figure 4, our method exhibits the fastest convergence among EGBMs, IGBMs, and single-loop methods. From Table 8, it is evident that our method achieves two significant improvements. In comparison to the effective BAMM, our method demonstrates an 85.8% improvement in inference time. Furthermore, the proposed scheme incurs the lowest computational cost, utilizing only 10.35% of the memory used by BAMM.

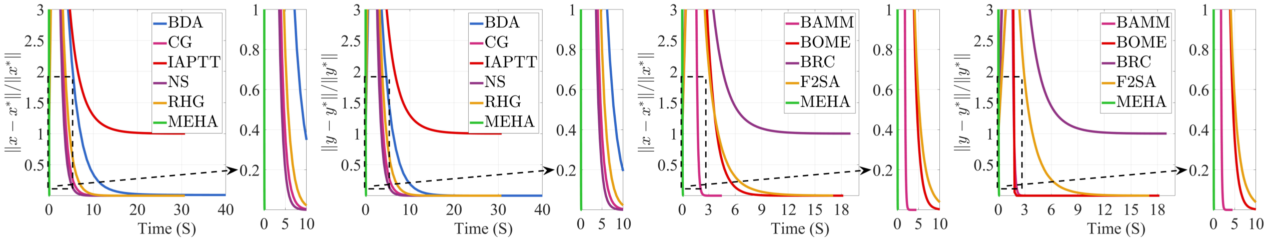

Figure 4. Illustrating the convergence curves of advanced BLO methods and MEHA by the criterion of $\|x-x^*\|/\|x^*\|$ and $\|y-y^*\|/\|y^*\|$ under LL strongly convex case.

Table 8. Time and memory of diverse BLO schemes under LL strongly convex case.

Methods	RHG	BDA	IAPTT	CG	NS	BRC	BOME	F^{2}SA	BAMM	MEHA
Time (S)	14.01	50.13	31.98	15.66	21.10	19.48	11.18	16.21	2.650	**0.166**
Memory (B)	160768	212480	160768	111104	110592	12800	14848	12288	14848	**1536**

LL Non-convex Case. We depict the convergence curves utilizing advanced Bi-Level Optimization (BLO) methods in the non-convex scenario with diverse metrics in Figure 5.

LL Non-Smooth Case. We present the convergence curves of our method, MEHA, across different dimensions in Figure 6. The results indicate that our method effectively identifies the optimal hyper-parameter x^* and the optimal solution $\{x^*, y^*\}$ under diverse high dimensions.

Neural Architecture Search. Additionally, we have also conducted a NAS experiment to test the performance of our proposed method on architectures that use the smooth activation function (Swish). For comparison, the compared version

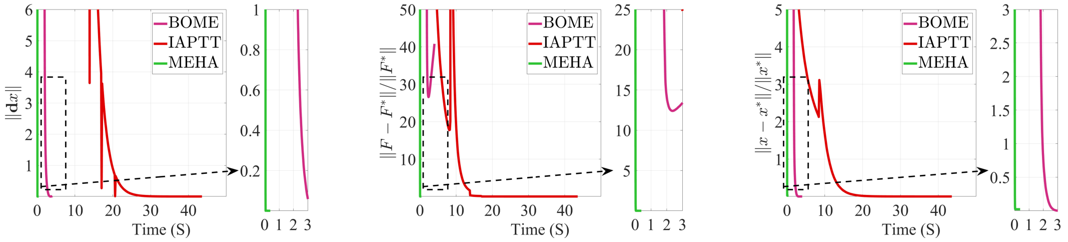

Figure 5. Visualizing the convergence behaviours of BOME, IAPTT and MEHA under LL non-convex case with one dimension, using the metrics of descent direction $\|\mathbf{d}x\|$, UL objective F and reconstruction error with x.

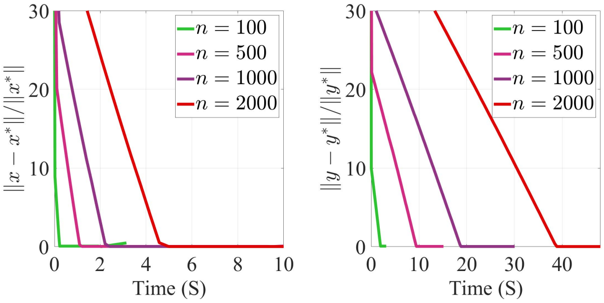

Figure 6. Illustrating the convergence curves by the criterion of $\|x - x^*\|/\|x^*\|$ and $\|y - y^*\|/\|y^*\|$ under the LL non-smooth case with different dimensions.

was searched within the original search space that features non-smooth activation functions. This comparison is detailed in Table 9 that includes two variants of our method: MEHA (non-smooth) and MEHA (smooth).

Table 9. Comparing Top-1 accuracy in searching and evaluation stages with diverse activated functions.

Methods	Searching		Evaluation		
	Train	Valid	Train	Valid	Test
MEHA (Non-Smooth)	99.060	99.764	99.419	96.150	96.070
MEHA (Smooth)	99.292	99.786	99.360	95.870	95.870

A.11. Experimental Details

First, by using the results in Section A.8, it can be verified that all models employed in our experimental studies meet the assumptions assumed.

Second, we conducted the experiments on a PC with Intel i7-8700 CPU (3.2 GHz), 32GB RAM and NVIDIA RTX 1070 GPU. We utilized the PyTorch framework on the 64-bit Windows system.

Synthetic Numerical Examples. We utilize the SGD optimizer to update the UL variable x. We uniformly utilize the $\|x-x^*\|/\|x^*\| \leq 10^{-3}$ criterion for LL merely convex case. We utilize the $\|dx\| \leq 1e^{-8}$ for one-dimension LL non-convex case $\|dx\| \leq 1e^{-3}$ for other high dimension cases. The maximum of iteration steps is 800. The learning steps α, β and η are fixed. Specifically, inverse power learning rate annealing strategy to dynamically adjust the learning rate (α and β) for LL merely-convex case.

As for the LL non-smooth case, We leverage the method (Feng & Simon, 2018) to generate the synthesized data for the group lasso hyper-parameter selection case, including 100 training, validation, and testing pairs, respectively. We set the response b_i as $b_i = v^T a_i + \sigma\epsilon_i$, where $v = [v^{(1)}, v^{(2)}, v^{(3)}]$, $v^{(i)} = (1, 1, 1, \cdots, 1, 0, 0, 0, \cdots, 0)$. The number of elements with value 1 is 50. The parameters σ are sampled from a standard normal distribution, and ϵ is selected to achieve an SNR of 2. The number of groups is set as 30 for 600 dimensions and 300 for other scenarios. $\|x^k - x^{k-1}\|/\|x^k\| \leq 0.2$ is as the stop criterion. As for the experimental settings of other compared methods (including grid search, random search, TPE, IGJO and VF-iDCA), we follow the effective practice (Gao et al., 2022).

For hyper-parameter selection, we set specific values for α_0, β_0, η_0, γ, $\underline{c}$, and p under different convexity scenarios: 1.5, 0.8, 0.8, 10, 33.3, and 0.49 for strongly convex cases; 0.012, 0.1, 0.009, 5, 0.167, and 0.49 for merely convex cases. In non-convex cases, we set 5e-4, 5e-4, 0.001, 200, 0.02, and 0.49 for these parameters, respectively. Additionally, for Equation (20) in non-smooth cases, we used 0.1, 1e-5, 0.1, 10, 2, and 0.49, while for Equation (21), we employed 0.01, 0.05, 0.05, 100, 20, and 0.48 for these hyper-parameters.

Few-Shot Learning. As for this task, the upper-level variables x represent the shared weights for feature extraction. $y := \{y^j\}$ denotes the task-specific parameters. We leverage superscripts to represent task indexes and subscripts to denote element indexes of vectors. Leveraging the cross-entropy loss as the objective l, we provide the bi-level formulation as:

$$\min_{x,y} \sum_j l_{val}(x, y^j; \mathcal{D}^j_{\texttt{val}}) \quad \text{s.t.} \quad y \in \arg\min_{\tilde{y}} \sum_j l_{tr}(x, \tilde{y}^j; \mathcal{D}^j_{\texttt{tr}}). \tag{78}$$

Following with the practice (Liu et al., 2023c), we utilize four layers of convolution blocks (ConvNet-4) to construct the backbone (*i.e., x*), which is widely utilized for few-shot learning tasks. The task-specific classifier y is composited by fully-connection layers with softmax operation. Adam and SGD optimizers are utilized to update x and y for all algorithms fairly. The meta batch size and hidden layers are 16 and 32, respectively. α_0, β_0, η_0, γ, $\underline{c}$ and p are 0.08, 0.05, 0.001, 100, 0.067 and 0.08, respectively. We utilize the inverse power learning rate annealing strategy to dynamically adjust the learning rate (α and β). η and γ are fixed.

Data Hyper-Cleaning. The mathematical formulation can be written as:

$$\min_{x,y} \sum_{u_i, q_i \in \mathcal{D}_{\texttt{val}}} l_{val}(y; u_i, q_i) \quad \text{s.t.} \quad y \in \arg\min_{\tilde{y}} \sum_{u_i, q_i \in \mathcal{D}_{\texttt{tr}}} \sigma(x_i) l_{tr}(\tilde{y}; u_i, q_i), \tag{79}$$

where the upper-level variable x is a vector with the same dimension of the number of corrupted examples. y denotes the target classification model. $\sigma(x)$ is a sigmoid function. In detail, we only utilize two-layer of fully-connection to define y for the noncovex case. Two datasets FashionMNIST and MNIST are utilized to conduct the experiments. We randomly split these datasets to composite the training, validation and testing subsets with 5000, 5000, 10000 examples, respectively. Half of data in the training dataset is tampered. α_0, β_0, η_0, γ, $\underline{c}$ and p are 0.1, 0.15, 0.001, 100, 0.2 and 0.25, respectively. SGD optimizer is utilized to update the UL variable x fairly. We utilize the inverse power learning rate annealing strategy to dynamically adjust the learning rate (α and β). η and γ are fixed.

Neural Architecture Search. The bi-level formulation of neural architecture search is

$$\min_{x,y} l_{val}(y, x; \mathcal{D}_{\texttt{val}}) \quad \text{s.t.} \quad y \in \arg\min_{\tilde{y}} l_{tr}(\tilde{y}, x; \mathcal{D}_{\texttt{tr}}), \tag{80}$$

where the architecture parameters are denoted as the upper-level variable x and the lower-level variable y represents the network weights. l_{val} and l_{tr} are the losses on validation and training datasets. The definition of search space, cells, and experimental hyper-parameters settings are following with the literature (Liu et al., 2018). We leveraged the Cifar-10 dataset to perform the experiments of image classification. As for the super-network, we conducted the search procedure with three layers of cells for 50 epochs. The network for training is increased with 8 layers and trained from scratch with 600 epochs. α_0, β_0, η_0, γ, $\underline{c}$ and p are 8e-5, 0.025, 0.025, 200, 2 and 0.49, respectively. We utilized the cosine decreasing learning rate annealing strategy to dynamically adjust the learning rate (α, β and η). η and γ are fixed.

Our experiment for NAS was designed to evaluate the performance of our proposed algorithm within practical bilevel optimization problems, particularly those involving a nonconvex lower-level objective. To ensure a consistent and fair comparison against existing methods, we adopted the subgradient descent technique for handling the nonsmoothness in neural networks, aligning our approach with that of our competitors (Liu et al., 2018). Specifically, our method utilizes the same search space defined in the referenced works, which incorporates the nonsmooth ReLU activation function in certain operations. For the gradient computation of the ReLU activation function, we define a unit step function Heaviside(x) for ReLU. It takes the value of $x > 0$ when and $x \leq 0$ when to approximate the gradient.