Averaged Method of Multipliers for Bi-Level Optimization without Lower-Level Strong Convexity

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

Abstract

Gradient methods have become mainstream techniques for Bi-Level Optimization (BLO) in learning fields. The validity of existing works heavily rely on either a restrictive Lower- Level Strong Convexity (LLSC) condition or on solving a series of approximation subproblems with high accuracy or both. In this work, by averaging the upper and lower level objectives, we propose a single loop Bi-level Averaged Method of Multipliers (sl-BAMM) for BLO that is simple yet efficient for large-scale BLO and gets rid of the limited LLSC restriction. We further provide non-asymptotic convergence analysis of sl-BAMM towards KKT stationary points, and the comparative advantage of our analysis lies in the absence of strong gradient boundedness assumption, which is always required by others. Thus our theory safely captures a wider variety of applications in deep learning, especially where the upper-level objective is quadratic w.r.t. the lower-level variable. Experimental results demonstrate the superiority of our method.

1. Introduction

Bi-Level Optimization (BLO) has recently attracted growing interests due to its wide applications in the field of deep learning, especially for hyperparameter optimization (Pedregosa, 2016; Franceschi et al., 2017; 2018; Okuno et al., 2018; Mackay et al., 2019), meta learning (Franceschi et al., 2018; Zügner & Günnemann, 2018; Rajeswaran et al., 2019; Ji et al., 2020), neural architecture search (Liu et al., 2018; Liang et al., 2019; Chen et al., 2019; Elsken et al., 2020), adversarial learning (Goodfellow et al., 2014; Pfau & Vinyals, 2016), and reinforcement learning (Yang et al., 2019; Hong et al., 2023), etc. BLO tackles nested optimization structures appearing in applications. In the last decade, BLO has emerged as a prevailing optimization technique for modern machine learning tasks with an underlying hierarchy.

BLO is notoriously challenging due to its nested nature. A variety of applications in deep learning have greatly promoted the development of BLO, with a broad collection of methods and algorithms. Among them, due to the effectiveness and simplicity, gradient-based algorithms have become mainstream techniques for BLO in learning and vision fields (Liu et al., 2021a), which can be generally categorized into the iterative differentiation (ITD) based approach (Maclaurin et al., 2015; Franceschi et al., 2017; Shaban et al., 2019; Grazzi et al., 2020; Liu et al., 2020; Li et al., 2020; Liu et al., 2021b; Ji et al., 2021; 2022; Liu et al., 2022a; 2023a;b) and the approximate implicit differentiation (AID) based approach (Pedregosa, 2016; Ghadimi & Wang, 2018; Rajeswaran et al., 2019; Lorraine et al., 2020; Hong et al., 2023; Huang et al., 2022; Ji et al., 2021; 2022; Arbel & Mairal, 2022a). Despite the large literature, most of existing methods for BLO are slow and unsatisfactory in various ways. For example, most of these studies (including both algorithmic design and theoretical investigation) heavily rely on a restrictive Lower-Level Strong Convexity (LLSC) condition, except a few attempts recently (which will be discussed later on). However, it is observed that these conventional gradient-based algorithms may lead to incorrect solutions in the absence of LLSC; see, e.g., Example 1 in (Liu et al., 2020) and the experimental results in (Liu et al., 2020; Ye et al., 2022). Hence, a practical issue lingers: *how to design a fast BLO algorithm, which enjoys both wider applicability and provable convergence guarantee, especially without LLSC?*

The studies beyond LLSC are rather limited. Several recent works partially addressed the above issue by ITD-based approach using a bi-level sequential averaging method (Sabach

[1]International School of Information Science & Engineering, Dalian University of Technology, Dalian, China [2]School of Software Technology, Dalian University of Technology [3]Department of Mathematics and National Center for Applied Mathematics Shenzhen, Southern University of Science and Technology, Shenzhen, China [4]Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia, Canada [5]SUSTech International Center for Mathematics, Southern University of Science and Technology, Shenzhen, China. Correspondence to: Jin Zhang <zhangj9@sustech.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

& Shtern, 2017; Liu et al., 2020; Li et al., 2020) or using a bi-level value function based penalty and barrier methods (Liu et al., 2021b;c). Without using ITD-based approach, recent studies (Ye et al., 2022; Sow et al., 2022) proposed two kinds of first-order BLO algorithms that depend only on first-order gradient information, based on the value function approach for BLO (Outrata, 1990; Ye & Zhu, 1995). However, all of these approaches take a double-loop optimization structure (involving numerical Lower-Level (LL) optimization loops), which heavily relies on solving a series of approximation subproblems with high accuracy. While double-loop algorithms have many applications across deep learning, they suffer from time and memory complexity, and could be difficult to implement in practice as the number of required inner loop iterations are difficult to adjust.

1.1. Main Contributions

The focus of this paper is to develop a provably faster single-loop BLO algorithm beyond LLSC. Such extension of the existing single-loop BLO methods with LLSC is not straightforward, requiring introducing the stationarity measure as stopping criterion in an appropriate way. We start from reformulating BLO to an equality-constrained optimization using the first-order stationarity condition of LL problem. Clearly, when LL problem is unconstrained and convex, this reformulation keeps the equivalence. Our **first insight** is that KKT condition of the resulting equality-constrained optimization can be regarded as an appropriate stationarity measure for BLO in the absence of LLSC. Observe that the equality constraint in the resulting problem comes from a convex or even strongly convex unconstrained LL problem. Our **second insight** is to well exploit the bi-level structure of the resulting equality-constrained problem, keeping the update of LL variable by gradient descent of LL objective and its variants. Notice that this feature is distinctively different from MPEC approach (Kim et al., 2020) and the penalty method for BLO in (Mehra & Hamm, 2021) where the bi-level structure disappeared after reformulation. Our **third insight** is to introduce the dual multiplier associated to the resulting equality-constrained optimization as a new variable in the update of BLO algorithm, which will enable us to develop a provably faster single-loop BLO algorithm in the absence of LLSC. The main contributions of this paper can be summarized as follows:

- We first propose a single loop Bi-level Averaged Method of Multipliers (sl-BAMM) beyond LLSC. The striking feature of si-BAMM is simple yet highly efficient for solving large-scale BLO, and the comparative advantage lies in theoretically guaranteed convergence.
- We use KKT stationarity as the trackable measure of solution quality returned by sl-BAMM. Further, a comprehensive convergence rate analysis for sl-BAMM characterized in terms of KKT stationarity is conducted. Compared to exiting works which are either lack of convergence towards stationarity or convergence rate, our result is evidently stronger and general enough to cover BLO with multiple LL minimizers.
- Finally, we are the first to conduct theoretical analysis for general BLO in the absence of the strong boundedness assumption of gradients of UL objective w.r.t. LL variable, which was prevailing among existing literature. Hence, our result captures a wider range of applications in deep learning, especially where UL objective is quadratic w.r.t. LL variable.

1.2. Related Work

In this subsection we give a brief review of some recent works that are directly related to ours and refer to Appendix A for a detailed comparison of these methods.

BLO Beyond Lower-Level Strong Convexity (LLSC). The studies beyond LLSC are rather limited. To our knowledge, there are three main types of study used to get rid of LLSC. The first one is the sequential averaging method (SAM) used in (Sabach & Shtern, 2017; Liu et al., 2020; Li et al., 2020). In the setting of BLO, for fixed UL variable, SAM iteratively generates a sequence LL points by averaging the gradient descents of UL and LL objectives. The second one is based on the value function approach for BLO, such as (Liu et al., 2021b; Ye et al., 2022; Sow et al., 2022) where they all need numerical LL optimization loops to approximate the value function accurately. The last one is replacing LL problem by its first-order stationarity condition that was used in (Mehra & Hamm, 2021; Liu et al., 2021d) and be called MPEC approach (Kim et al., 2020) when LL problem is constrained. However, all of these approaches take a double-loop optimization structure (involving numerical LL optimization loops), which could be difficult to implement in practice as the number of required inner loop iterations are difficult to adjust. In contrast, our algorithm sl-BAMM has a single loop structure, which admits a much simpler implementation. Moreover, the updates in sl-BAMM evolve simultaneously and then can be performed in parallel instead of sequentially.

Stationarity Measure for BLO Algorithm. The most common stationarity measure for BLO is the gradient norm of the total UL objective (i.e., the norm of hypergradient) (Pedregosa, 2016; Ghadimi & Wang, 2018; Chen et al., 2021; Ji et al., 2021; 2022; Li et al., 2022; Dagréou et al., 2022; Arbel & Mairal, 2022a). Clearly, the norm of hypergradient heavily rely on LLSC. For general BLO, the stationarity measure has been much less studied, from an algorithmic point of view, especially in the context of machine learning. Recently, stationarity measures involving value function or its regularized variant have been proposed in (Ye et al.,

2022) and (Sow et al., 2022) respectively.

Discussion on Gradient Boundedness Assumption. The most of existing analyses in BLO literature reply on a strong assumption on the boundedness of gradients of UL objective w.r.t. LL variable (a.k.a. gradient boundedness assumption, GBA), to guarantee the smoothness of the hypergradient and then to get tight convergence rates (Ghadimi & Wang, 2018; Hong et al., 2023; Ji et al., 2021; Chen et al., 2021; Li et al., 2022; Ye et al., 2022). However, GBA is fairly restrictive for applications. For example, it does not hold if UL objective is quadratic w.r.t. LL variable, which is representative in the machine learning context; see e.g., kernel ridge regression, biased regularization and hyper-representation in (Grazzi et al., 2020). Recently, when LL objective is strongly convex, a weaker boundedness assumption was made in (Ji et al., 2022; Dagréou et al., 2022). Further, for the case where the total UL objective is strongly convex or convex, by using an induction analysis that all iterates are bounded, (Ji & Liang, 2022) provided the first-known analyses of the complexity bound without GBA. However, the argument in (Ji & Liang, 2022) heavily rely on the (strong) convexity assumption on the total UL objective. In comparison, there is no assumption on the total UL objective and also no LLSC assumption in this work.

2. Proposed Algorithm

We consider BLO problem in the following form:

$$\min_{\mathbf{x},\mathbf{y}} F(\mathbf{x},\mathbf{y}) \quad \text{s.t.} \quad \mathbf{y} \in \mathcal{S}(\mathbf{x}) := \arg\min_{\mathbf{y}\in\mathbb{R}^m} f(\mathbf{x},\mathbf{y}), \quad (1)$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ are UL and LL variables respectively. Here UL and LL objectives F, f are smooth functions with Lipschitz continuous first and second order derivatives.

In this section, we propose a single loop Bi-level Averaged Method of Multipliers (sl-BAMM) for BLO without relying on LL strong convexity (LLSC).

2.1. Stationarity Measure

Most of previous works assume that $f(\mathbf{x},\cdot)$ is strongly convex for any $\mathbf{x}$. Thus $\mathcal{S}(\mathbf{x})$ is always a singleton, denoted by $\mathbf{y}^*(\mathbf{x})$. Then the total UL objective $\Phi(\mathbf{x}) := F(\mathbf{x},\mathbf{y}^*(\mathbf{x}))$ is continuously differentiable. Following the standard analysis in nonconvex optimization, since $\Phi(\mathbf{x})$ is nonconvex, algorithms are expected to find a near stationary point of $\Phi(\mathbf{x})$ as measured by the norm of its gradient (i.e., hypergradient). However, without LLSC, this goal is not suitable since $\mathcal{S}(\mathbf{x})$ may be a set-valued mapping and $\Phi(\mathbf{x})$'s variants (in optimistic and pessimistic) may be nonsmooth (even be discontinuous), which makes traditional definitions of stationary point inapplicable. Therefore, one fundamental question arises:

Question: What is a good stationarity measure for BLO algorithm without LLSC?

To answer the above question, we start from formulating BLO to an equality-constrained optimization:

$$\min_{\mathbf{x},\mathbf{y}} F(\mathbf{x},\mathbf{y}) \quad \text{s.t.} \quad \nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) = 0. \quad (2)$$

Clearly, Problems (1) and (2) are equivalent when LL objective $f(\mathbf{x},\mathbf{y})$ is convex w.r.t. $\mathbf{y}$ for any fixed $\mathbf{x}$. This motivates us to use KKT condition of Problem (2) as a measure of stationarity of the solution returned by BLO algorithms. Let $\mathcal{L}(\mathbf{x},\mathbf{y},\mathbf{v}) := F(\mathbf{x},\mathbf{y}) - \mathbf{v}^T\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})$ be the Lagrangian function of Problem (2), where $\mathbf{v}$ is the dual multiplier. We say that $(\mathbf{x}^*,\mathbf{y}^*)$ is a KKT point if there exists a dual multiplier $\mathbf{v}^*$ such that $\nabla\mathcal{L}(\mathbf{x}^*,\mathbf{y}^*,\mathbf{v}^*) = 0$. That is, the following KKT condition holds:

$$\begin{pmatrix} \nabla_{\mathbf{x}} F(\mathbf{x}^*,\mathbf{y}^*) - \nabla^2_{\mathbf{xy}} f(\mathbf{x}^*,\mathbf{y}^*)\mathbf{v}^* \\ \nabla_{\mathbf{y}} F(\mathbf{x}^*,\mathbf{y}^*) - \nabla^2_{\mathbf{yy}} f(\mathbf{x}^*,\mathbf{y}^*)\mathbf{v}^* \\ -\nabla_{\mathbf{y}} f(\mathbf{x}^*,\mathbf{y}^*) \end{pmatrix} = 0. \quad (3)$$

To appropriately characterize the convergence rate of the solution returned by a BLO algorithm, we define KKT residual for $(\mathbf{x},\mathbf{y},\mathbf{v})$ as following

$$\mathrm{KKT}(\mathbf{x},\mathbf{y},\mathbf{v}) := \|\nabla\mathcal{L}(\mathbf{x},\mathbf{y},\mathbf{v})\|^2.$$

Then, $\mathrm{KKT}(\mathbf{x},\mathbf{y},\mathbf{v}) = 0$ if and only if $(\mathbf{x},\mathbf{y})$ is a KKT point and $\mathbf{v}$ is a multiplier. The next proposition provides a strong connection between KKT condition and the hypergradient using in the existing BLO methods.

Proposition 2.1. *Assume that for all $\mathbf{x}$, $f(\mathbf{x},\cdot)$ is strongly convex. Then $\nabla\Phi(\mathbf{x}) = 0$ if and only if $\mathrm{KKT}(\mathbf{x},\mathbf{y},\mathbf{v}) = 0$ for some $\mathbf{y},\mathbf{v} \in \mathbb{R}^m$. In fact, $\mathbf{y} = \mathbf{y}^*(\mathbf{x})$ and*

$$\mathbf{v} = \left[\nabla^2_{\mathbf{yy}} f(\mathbf{x},\mathbf{y}^*(\mathbf{x}))\right]^{-1} \nabla_{\mathbf{y}} F\left(\mathbf{x},\mathbf{y}^*(\mathbf{x})\right).$$

Since the nonsingularity of $\nabla^2_{\mathbf{yy}} f$ is not necessary for KKT condition, KKT residual is more applicable to general BLOs. Further, unlike the computation of hypergradient, which would require knowledge of the unique LL solution to be usable as a stopping criterion, KKT residual is readily available to the algorithm as a stopping criterion. In a novel framework of (Dagréou et al., 2022), assuming LLSC, the newly involved variable $\mathbf{v}$ is seen as the solution of a linear system: $\left[\nabla^2_{\mathbf{yy}} f(\mathbf{x},\mathbf{y}^*(\mathbf{x}))\right]\mathbf{v} = \nabla_{\mathbf{y}} F\left(\mathbf{x},\mathbf{y}^*(\mathbf{x})\right)$. We offer a new perspective on $\mathbf{v}$ that help us to develop a novel single-loop BLO algorithm beyond LLSC.

2.2. Bi-level Averaged Method of Multipliers

As shown in Algorithm 1, by introducing dual multiplier from the viewpoint of KKT condition and adopting the sequencing averaging method in (Sabach & Shtern, 2017;

Liu et al., 2020; Li et al., 2020), we propose a single loop Bi-level Averaged Method of Multipliers (sl-BAMM) in which UL variable and the dual multiplier evolve simultaneously with LL variable, following the directions given by

$$d_{\mathbf{y}}^k = \nabla_{\mathbf{y}} \psi_{\mu_k}(\mathbf{x}_k, \mathbf{y}_k), \tag{4}$$

$$d_{\mathbf{v}}^k = \nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{yy}}^2 \psi_{\mu_k}(\mathbf{x}_k, \mathbf{y}_k)\mathbf{v}_k, \tag{5}$$

$$d_{\mathbf{x}}^k = \nabla_{\mathbf{x}} F(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{xy}}^2 \psi_{\mu_k}(\mathbf{x}_k, \mathbf{y}_k)\mathbf{v}_k, \tag{6}$$

where the aggregation function

$$\psi_\mu(\mathbf{x}, \mathbf{y}) := \mu F(\mathbf{x}, \mathbf{y}) + (1-\mu) f(\mathbf{x}, \mathbf{y}), \tag{7}$$

averaging UL and LL objectives, is inspired by BDA (Liu et al., 2020). Here μ is the averaging (aggregation) parameter that will iteratively goes to zero.

Algorithm 1 single loop Bi-level Averaged Method of Multipliers (sl-BAMM)

Input: initial points $\mathbf{x}_0, \mathbf{y}_0, \mathbf{v}_0$, stepsizes $\alpha_k, \beta_k, \eta_k$, aggregation parameter μ_k;

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: update $\mathbf{y}_{k+1} = \mathbf{y}_k - \beta_k d_{\mathbf{y}}^k$ according to Eq. (4);
3: update $\mathbf{v}_{k+1} = \mathbf{v}_k + \eta_k d_{\mathbf{v}}^k$ according to Eq. (5);
4: update $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k d_{\mathbf{x}}^k$ according to Eq. (6).
5: **end for**

Remark 2.2. The first and most important motivation of sl-BAMM is to leverage KKT conditions to update all variables, but keeping the update of LL variable by gradient descent of LL objective and its variants. This will enable sl-BAMM to be a provably single loop hence faster algorithm, differently from MPEC approach (Kim et al., 2020) and the penalty method for BLO in (Mehra & Hamm, 2021).

Remark 2.3. Secondly, it is worth noting that all the directions in Eq. (4)-(6) are linear in F and f, differently from the existing BLO algorithms without using an extra variable. This kind of linear structure will motivate more stochastic and global variance reduction algorithms as done in (Dagréou et al., 2022) with LL strong convexity.

Note that the dual multiplier in BLO algorithm will result in practical procedures for correcting DARTS (Liu et al., 2018) and other recent algorithms attempt to simplify multi-step iteration with the single gradient, and removing an unavoidable non-vanishing convergence error in ITD-based approach (Ji et al., 2022).

3. Theoretical Investigations

In this section, we provide convergence rates of sl-BAMM with or without LL strong convexity. Note that, unlike existing analyses, *we do not require a strong assumption on the boundedness of* $\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$. Hence, our theoretical analysis can be applied to a wider variety of applications in deep learning. The proofs are deferred in Appendix.

3.1. General Assumptions

We start by stating some standard assumptions on UL and LL objectives.

Assumption 3.1. The objectives F and f satisfy

(a) UL objective F is twice differentiable. The first order derivatives $\nabla_{\mathbf{x}} F(\cdot, \mathbf{y})$, $\nabla_{\mathbf{x}} F(\mathbf{x}, \cdot)$, $\nabla_{\mathbf{y}} F(\cdot, \mathbf{y})$, $\nabla_{\mathbf{y}} F(\mathbf{x}, \cdot)$ are Lipschitz continuous with respective Lipschitz constants $L_{F_{\mathbf{x}1}}, L_{F_{\mathbf{x}2}}, L_{F_{\mathbf{y}1}}, L_{F_{\mathbf{y}2}}$.

(b) LL objective f is twice differentiable. The derivatives $\nabla_{\mathbf{y}} f$ and $\nabla_{\mathbf{xy}}^2 f, \nabla_{\mathbf{yy}}^2 f$ are Lipschitz continuous in $(\mathbf{x}, \mathbf{y})$ with respective Lipschitz constants $L_{f_{\mathbf{y}1}}, L_{f_{\mathbf{y}2}}$ and $L_{f_{\mathbf{xy}1}}, L_{f_{\mathbf{xy}2}}, L_{f_{\mathbf{yy}1}}, L_{f_{\mathbf{yy}2}}$.

The smoothness assumption above is common in BLO literature, see, e.g., (Ghadimi & Wang, 2018; Ji et al., 2021; Khanduri et al., 2021; Ji & Liang, 2022; Chen et al., 2022; Dagréou et al., 2022; Ji et al., 2022). In this paper, we study two classes of BLO: (1) LL objective $f(\mathbf{x}, \cdot)$ is merely convex and UL objective $F(\mathbf{x}, \cdot)$ is strongly convex; (2) LL objective $f(\mathbf{x}, \cdot)$ is strongly convex. This can be unified by the following assumption.

Assumption 3.2. For any $\mathbf{x}$, the aggregation function $\psi_\mu(\mathbf{x}, \cdot) = \mu F(\mathbf{x}, \cdot) + (1-\mu) f(\mathbf{x}, \cdot)$ is strongly convex for $\mu = 0$ or $\mu > 0$.

3.2. General Results for LL Merely Convex Case

To handle BLO with multiple LL minimizers, we take the following standard assumption as in (Liu et al., 2020).

Assumption 3.3. (a) For any fixed $\mathbf{x}$, LL objective $f(\mathbf{x}, \cdot)$ is convex, and UL objective $F(\mathbf{x}, \cdot)$ is σ_F-strongly convex and has a uniform lower bound denoted by F_0.

(b) The derivatives $\nabla_{\mathbf{xy}}^2 F, \nabla_{\mathbf{yy}}^2 F$ are Lipschitz continuous in $(\mathbf{x}, \mathbf{y})$ with respective Lipschitz constants $L_{F_{\mathbf{xy}1}}, L_{F_{\mathbf{xy}2}}, L_{F_{\mathbf{yy}1}}, L_{F_{\mathbf{yy}2}}$.

Under the above assumption, the aggregation function $\psi_\mu(\mathbf{x}, \cdot)$ is σ_{ψ_μ}-strongly convex with $\sigma_{\psi_\mu} = \mu\sigma_F$. Hence $\psi_\mu(\mathbf{x}, \cdot)$ has a unique minimizer, denoted by $\mathbf{y}_\mu^*(\mathbf{x})$. We define the approximate overall UL objective by $\Phi_{\mu_k}(\mathbf{x}_k) := F(\mathbf{x}_k, \mathbf{y}_{\mu_k}^*(\mathbf{x}_k))$, and also denote the correct dual multiplier by $\mathbf{v}_\mu^*(\mathbf{x}) := \left[\nabla_{\mathbf{yy}}^2 \psi_\mu(\mathbf{x}, \mathbf{y}_\mu^*(\mathbf{x}))\right]^{-1} \nabla_{\mathbf{y}} F\left(\mathbf{x}, \mathbf{y}_\mu^*(\mathbf{x})\right)$. Let $(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k)$ be a sequence, we define

$$\begin{aligned}\Pi(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k) =& \|\nabla \Phi_{\mu_k}(\mathbf{x}_k)\|^2 + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &+ \|\mathbf{y}_k - \mathbf{y}_{\mu_k}^*(\mathbf{x}_k)\|^2 + \|\mathbf{v}_k - \mathbf{v}_{\mu_k}^*(\mathbf{x}_k)\|^2.\end{aligned}$$

Now we provide a series of convergence rate analysis for sl-BAMM towards KKT stationary points, using three different simple and handy strategies for choosing the step sizes.

Specially, the first one does **NOT** require any assumption on the boundedness of $\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$.

Theorem 3.4. *Suppose Assumptions 3.1 and 3.3 hold. Choose $\mu_k = \bar{\mu}(k+1)^{-p}$ with $0 < p < 1/10$, and stepsizes $\beta_k \in \left[\bar{\beta}(k+1)^{-\tau/2}, 1/\left(L_{F_{\mathbf{y}2}} + L_{f_{\mathbf{y}2}}\right)\right]$ and*

$$\eta_k = (k+1)^{-\tau/2}\beta_k\mu_k^2, \quad \alpha_k = (k+1)^{-3\tau/2}\beta_k\mu_k^7, \quad (8)$$

with $0 < \tau < 1/30$. Let $(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k)$ be the sequence generated by sl-BAMM. If $(\mathbf{x}_k, \mathbf{y}_k)$ is bounded, then

$$\min_{0\le k\le K}\left\{\Pi(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k)\right\} = O\left(\frac{1}{K^{1-9p-3\tau}}\right).$$

If $\mathbf{v}_k$ is also bounded, then

$$\min_{0\le k\le K}\left\{\mathrm{KKT}(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k)\right\} = O\left(\frac{1}{K^{1-9p-3\tau}} + \frac{1}{K^{2p}}\right).$$

Furthermore, any limit point of $(\mathbf{x}_k, \mathbf{y}_k)$ is a KKT point of Problem (2), and the convergence rate of $\mathbf{x}$ is given by $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = O(k^{-\frac{5p}{2}-\frac{\tau}{4}})$.

Next we establish tighter convergence rates when a weaker gradient boundedness assumption holds.

Theorem 3.5. *Suppose Assumptions 3.1 and 3.3 hold. Under the same setting of Theorem 3.4 except $0 < p < 1/6$, $0 < \tau < 1/18$ and replacing stepsizes (8) by*

$$\eta_k = (k+1)^{-\tau/2}\beta_k\mu_k, \quad \alpha_k = (k+1)^{-3\tau/2}\beta_k\mu_k^5. \quad (9)$$

Let $(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k)$ be the sequence generated by sl-BAMM. If $\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k))$ is bounded, then

$$\min_{0\le k\le K}\left\{\Pi(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k)\right\} = O\left(\frac{1}{K^{1-5p-3\tau}}\right).$$

If $\mathbf{v}_k$ is also bounded, then

$$\min_{0\le k\le K}\left\{\mathrm{KKT}(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k)\right\} = O\left(\frac{1}{K^{1-5p-3\tau}} + \frac{1}{K^{2p}}\right).$$

Furthermore, any limit point of $(\mathbf{x}_k, \mathbf{y}_k)$ is a KKT point of Problem (2), and the convergence rate of $\mathbf{x}$ is given by $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = O(k^{-\frac{5p}{2}-\frac{\tau}{4}})$.

Note that $\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k))$ is bounded if $\mathbf{x}_k$ is bounded, and for any $B > 0$ there are positive constants C and μ_0 such that $\|\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}^*_{\mu}(\mathbf{x}))\| \le C$ for all $\|\mathbf{x}\| \le B$ and $0 < \mu \le \mu_0$. It is satisfied for a broad collection of applications, including BLOs with multiple LL minimizers in our experiment. Specially, assuming LLSC, $\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k))$ is bounded if there exists $C_F > 0$ such that $\|\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \le C_F$ for all $\mathbf{x}$, a weaker boundedness assumption in (Ji et al., 2022; Dagréou et al., 2022).

Finally, if $\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|$ is bounded, the above results can be improved further.

Theorem 3.6. *Suppose Assumptions 3.1 and 3.3 hold. Under the same setting of Theorem 3.4 except $0 < p < 1/4$, $0 < \tau < 1/12$ and replacing stepsizes (8) by*

$$\eta_k = (k+1)^{-\tau/2}\beta_k, \quad \alpha_k = (k+1)^{-3\tau/2}\beta_k\mu_k^3. \quad (10)$$

Let $(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k)$ be the sequence generated by sl-BAMM. If $\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|$ is bounded, then we have

$$\min_{0\le k\le K}\left\{\Pi(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k)\right\} = O\left(\frac{1}{K^{1-3p-3\tau}}\right).$$

If $\mathbf{v}_k$ is also bounded, then

$$\min_{0\le k\le K}\left\{\mathrm{KKT}(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k)\right\} = O\left(\frac{1}{K^{1-3p-3\tau}} + \frac{1}{K^{2p}}\right).$$

Furthermore, any limit point of $(\mathbf{x}_k, \mathbf{y}_k)$ is a KKT point of Problem (2), and the convergence rate of $\mathbf{x}$ is given by $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = O(k^{-\frac{3p}{2}-\frac{\tau}{4}})$.

3.3. Special Discussion for LL Strongly Convex Case

For the special case where LL objective $f(\mathbf{x}, \cdot)$ is strongly convex, Assumption 3.2 holds where $\mu = 0$. Hence Assumption 3.3 can be removed.

Assumption 3.7. For fixed $\mathbf{x}$, $f(\mathbf{x}, \cdot)$ is σ_f-strongly convex.

Under Assumption 3.7, LL solution $\mathbf{y}^*(\mathbf{x})$ is well defined and differentiable. Hence, one can use the gradient of $\Phi(\mathbf{x}) = F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ to characterize the optimality of the generated sequence by the algorithms. Actually, as shown in Proposition 2.1, it is equivalent to KKT residual.

Relinquishing any assumption on the boundedness of $\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$, our next result provides a new convergence rate guarantee for LL strongly convex case.

Theorem 3.8. *Suppose Assumption 3.7 holds. Consider sl-BAMM with $\mu_k = 0$. We choose $\tau > 0$ and stepsizes*

$$\begin{aligned} &\bar{\beta}(k+1)^{-\tau/2} \le \beta_k \le \frac{1}{L_{F_{\mathbf{y}2}} + L_{f_{\mathbf{y}2}}}, \\ &\eta_k = \bar{\eta}(k+1)^{-\tau/2}\beta_k, \ \alpha_k = \bar{\alpha}(k+1)^{-\tau}\beta_k, \end{aligned} \quad (11)$$

where $\bar{\beta}, \bar{\eta}, \bar{\alpha}$ are positive constants satisfying certain conditions. Let $(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k)$ be the sequence generated by sl-BAMM. If $(\mathbf{x}_k, \mathbf{y}_k)$ is bounded, then $\mathbf{x}_k$ satisfies

$$\min_{0\le k\le K}\{\|\nabla\Phi(\mathbf{x}_k)\|^2\} = O\left(\frac{1}{K^{1-\frac{3\tau}{2}}}\right). \quad (12)$$

Remark 3.9. If GBA holds, that is, there is a constant C such that $\|\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})\| \le C$ for all $\mathbf{x}, \mathbf{y}$, then τ in Theorem 3.8 can be taken to be 0 and then $\min_{0\le k\le K}\{\|\nabla\Phi(\mathbf{x}_k)\|^2\} = O(1/K)$, which has been proved recently in (Li et al., 2022) (in the stochastic setting) and in (Ji et al., 2022) (in the

deterministic setting). Since τ can go to 0, Theorem 3.8 shows that sl-BAMM with or without GBA have almost the same convergence rate. It is worth noting that GBA does not hold if UL objective is quadratic in LL variable, which is representative in the machine learning context; referring to kernel ridge regression, biased regularization and hyper-representation in (Grazzi et al., 2020).

To analyze the convergence of sl-BAMM without LLSC and also relinquishing any assumption on the boundedness of $\nabla_{\mathbf{y}} F(\mathbf{x},\mathbf{y})$, according to different step size strategies, we will utilize different intrinsic Lyapunov functions with flexible coefficients given by

$$\begin{aligned} V_k :=& a_k \left[F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)) - F_0\right] \\ &+ b_k\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + c_k\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2. \end{aligned}$$

Here $\{a_k, b_k, c_k\}_{k=1}^{\infty}$ is a sequence of positive constants depending on the step size strategy. In particular, when $f(\mathbf{x},\cdot)$ is strongly convex, we take $\mu_k = 0$ for all k and then $F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ is the total UL objective. Generally, the first term $\Phi_\mu(\mathbf{x}) = F(\mathbf{x}, \mathbf{y}^*_\mu(\mathbf{x}))$ quantifies the (approximate) overall UL objective functions, the second term $\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2$ characterizes LL solution errors, and the third term $\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2$ delineates the multiplier errors.

We will first analyze the descent of the approximate overall UL objective in the next lemma.

Lemma 3.10. *Suppose Assumptions 3.1, and either Assumption 3.3 or Assumption 3.7 holds. Let $\mu_{k+1} \le \mu_k \le \frac{1}{2}$ for all k. Then the sequence of $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ generated by sl-BAMM satisfies*

$$\begin{aligned} &\Phi_{\mu_{k+1}}(\mathbf{x}_{k+1}) - \Phi_{\mu_k}(\mathbf{x}_k) \\ \le& -\frac{\alpha_k}{2}\|\nabla\Phi_{\mu_k}(\mathbf{x}_k)\|^2 \\ &- \frac{1}{2}\left(\frac{1}{\alpha_k} - \frac{C_{\Phi 1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + C_{\Phi 2}}{\sigma^2_{\psi_{\mu_k}}}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &+ \alpha_k\left(L_{\psi_{\mathbf{xy}2}}\left\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\| + L_{F_{\mathbf{x}2}}\right)^2\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\ &+ \alpha_k L^2_{\psi_{\mathbf{y}1}}\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\ &+ \frac{2\left\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2}{\sigma_F}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right), \end{aligned}$$

where $C_{\Phi 1}, C_{\Phi 2}$ are some explicit positive constants given in Lemma D.3.

Remark 3.11. Compared to the previous results using both LLSC and GBA, the coefficients of $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$ and $\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2$ depend on $\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|$ since Lipschitz continuity of the hypergradient and its surrogate could not be guaranteed without any assumption on the boundedness of $\nabla_{\mathbf{y}} F(\mathbf{x},\mathbf{y})$ and its variant. To address this issue, we characterize a weaker smoothness of $\Phi_\mu(\cdot)$ in Lemma D.3. Note also that there is an extra term in the inequality of Lemma 3.10 involving the descent of the averaging parameter. Here we adopt the convention that $0/0 = 0$.

For the convenience of the reader, we give a unified proof sketch of Theorems in Section 3 in Appendix B.

4. Experimental Results

In this section, we conduct experiments to study basic properties of sl-BAMM such as correctness, scalability and practicality. We also compare the performances of sl-BAMM with competitive methods on different tasks. Specially, we first conduct experiments on toy problems to verify the convergence results established in Section 3. Then we test the performance of sl-BAMM on two real-world BLO applications, including data hyper-cleaning and few-shot classification, compared with representative BLO methods. The code is available under `https://github.com/vis-opt-group/sl-BAMM/`.

4.1. Numerical Verification

LL Merely Convex Case. In the following, we first verify the convergence results under LL merely convex cases based on the toy example problem used in BDA (Liu et al., 2020):

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{x} - \mathbf{y}_2\|^2 + \frac{1}{2}\|\mathbf{y}_1 - \mathbf{e}\|^2 \\ &\text{s.t.} \quad \mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2) \in \arg\min_{(\mathbf{y}_1,\mathbf{y}_2)\in\mathbb{R}^{2n}} \frac{1}{2}\|\mathbf{y}_1\|^2 - \mathbf{x}^\top\mathbf{y}_1, \end{aligned} \tag{13}$$

where $\mathbf{e}$ denotes the vector of which the elements are all equal to 1. By simple calculation, we can calculate the optimal solution as $(\mathbf{e}, \mathbf{e}, \mathbf{e})$. Note that we use $\mathbf{d}_{\mathbf{x}}$ to denote the update direction at $\mathbf{x}$ calculated by different tested BLO methods, including sl-BAMM, RHG, Conjugate Gradient (CG) (Pedregosa, 2016), Neumann series (NS) (Rajeswaran et al., 2019) and BDA. In the following, though the updates of $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{v}$ of sl-BAMM are implemented sequentially, to show the potential efficiency of the parallelism, we use the maximum time cost for updating $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{v}$ respectively as the runtime of sl-BAMM at each iteration.

Since other methods don't involve the new introduced variable, for fairness and uniformity, we first compare all of the methods based on several widely accepted indicators in Fig. 1. As it is shown, since only BDA and sl-BAMM are capable to handle BLOs without LLSC assumption, they could converge to the truly global optimal solution, while other methods fail to converge to the correct solution. Moreover, sl-BAMM shows significant advantage over BDA in terms of the convergence speed due to a single-loop update manner. Besides, it can be observed that if μ_k is set as 0 in sl-BAMM (denoted as sl-BAMM (S)), it no longer achieves the convergence property, which also validates that the aggregation step in sl-BAMM is indispensable.

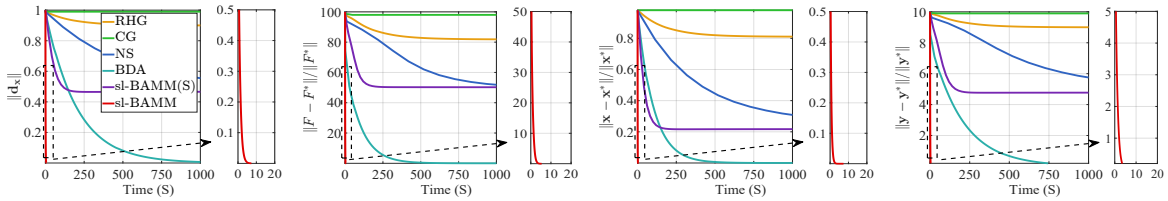
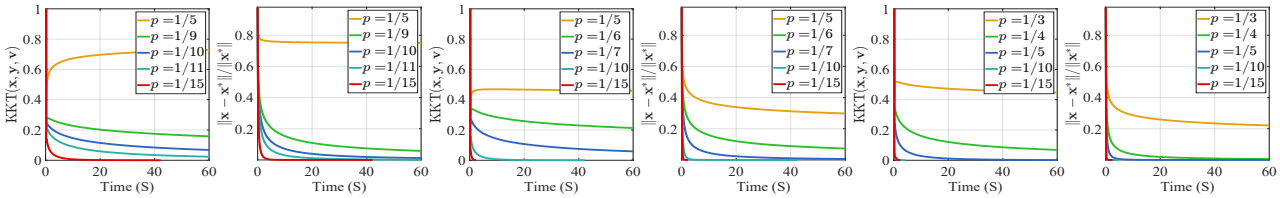


Figure 1. Illustrating the convergence curves of RHG, CG, NS, BDA and sl-BAMM based on different indicators, including $\mathbf{d_x}$ ($\mathbf{d_x}$ denotes the direction of descent of $\mathbf{x}$), F, $\mathbf{x}$, and $\mathbf{y}$. We also zoom in on the right side of each subfigure for better view of the curve at the very beginning. Note that sl-BAMM (S) represents the version of sl-BAMM with $\mu_k = 0$.



(a) Strategy S1 (b) Strategy S2 (c) Strategy S3

Figure 2. The convergence curves of sl-BAMM with three different strategies to update μ_k, α_k, β_k and η_k. In each subfigure, we illustrate the convergence behavior of $\mathbf{x}$ and $\mathrm{KKT}(\mathbf{x}, \mathbf{y}, \mathbf{v})$ by choosing different p to update μ_k.

Here, we use S1, S2 and S3 to denote three different parameter chosen strategies for sl-BAMM proposed in Theorem 3.4, 3.5 and 3.6, respectively. In Fig. 2, we test these three strategies with fixed τ and different values of p. It can be seen that when the value of p violates the bound required in the convergence result in Section 3.2 much, sl-BAMM fails to converge in terms of both UL variable and KKT residual. Besides, in this toy example, which satisfies all the assumptions required by these three strategies, strategy S3 always has the fastest convergence speed.

LL Strongly Convex Case. Next, we verify the performance of sl-BAMM to handle high-dimensional BLOs compared with existing methods based on the following toy example with LLSC assumption:

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{x}-\mathbf{z}_0\|^2 + \frac{1}{2}\mathbf{y}^*(\mathbf{x})^\top \mathbf{A}\mathbf{y}^*(\mathbf{x}) \\ &\text{s.t.} \quad \mathbf{y}^*(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x},\mathbf{y}) = \frac{1}{2}\mathbf{y}^\top\mathbf{A}\mathbf{y} - \mathbf{x}^\top\mathbf{y}, \end{aligned} \tag{14}$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{S}^{n\times n}$ is a positive-definite symmetric matrix and $\mathbf{z}_0 \neq 0$ is a given point in $\mathbb{R}^n$. It can be easily verified that this example satisfies the assumptions required by RHG, CG and NS. Here, we first set $\mathbf{A} = \mathbf{I}$, $\mathbf{z}_0 = \mathbf{e}$, then the unique solution of this problem is $\mathbf{x}^* = \mathbf{y}^* = \mathbf{e}/2$. Under LLSC assumption, function $\Phi(\mathbf{x}) = F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ is differentiable and we use $\nabla\Phi$ to denote its gradient at $\mathbf{x}$.

In Fig.3, we compare the convergence behavior of our proposed sl-BAMM with existing methods with different hyperparameter settings, including the LL iteration number T for RHG, error ϵ for CG and sequence length M for NS.

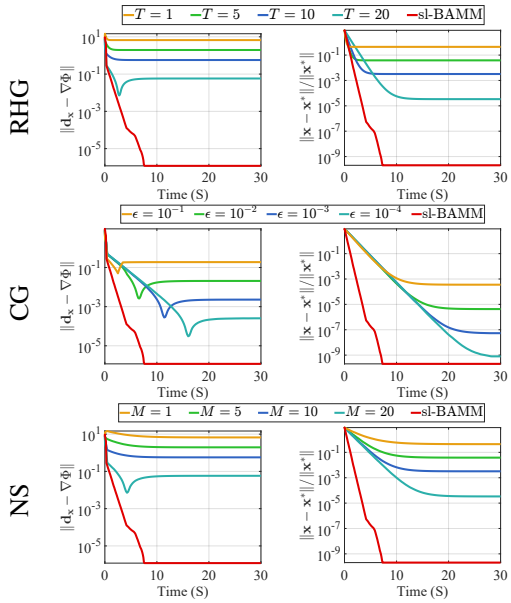
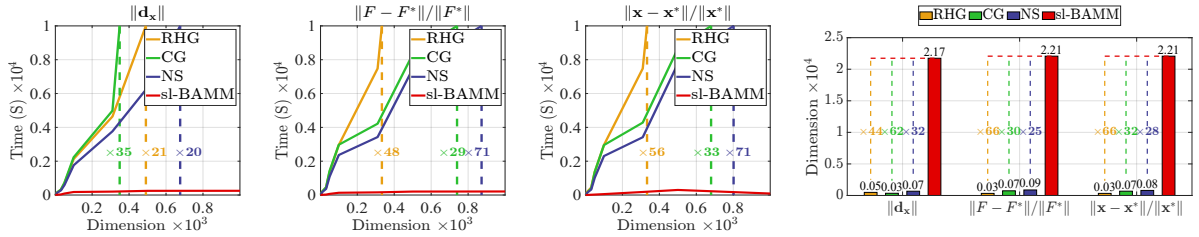


Figure 3. Convergence curves of sl-BAMM, RHG with different LL iteration number T, CG with different error ϵ and NS with different sequence length M.

And we illustrate the error of hypergradient $\|\mathbf{d_x} - \nabla\Phi\|$ and UL variable $\|\mathbf{x} - \mathbf{x}^*\|/\|\mathbf{x}^*\|$ as the calculation time increases. As shown in Fig.3, accelerating RHG, CG and NS by directly simplifying the LL iterations will deteriorate the quality of the iterates generated. Whereas, though sl-BAMM only has one step update for LL problem at each iteration, the update of the introduced multiplier variable v in sl-BAMM serves as a correction helping the generated iterates converge to the solution.



(a) Convergence time (b) Convergence dimension

Figure 4. Computational efficiency comparison of existing BLOs (RHG, CG and NS) and sl-BAMM. (a) The time required for different convergence metrics as the dimension increases. (b) The problem dimension when the runtime reached the upper limit (10^4s).

Table 1. Comparison of the results for hyper-cleaning and few-shot learning. We compared the time and F1 score for hyper-cleaning when achieving almost the same accuracy (81%) and the time for few-shot learning (95% for 5-way and 90% for 20-way).

Method	Hyper-cleaning			5-way		20-way	
	Acc. (%)	F1 score	Time (S)	Acc. (%)	Time (S)	Acc. (%)	Time (S)
RHG	81.01	86.12	95.18	95.01	536.17	90.27	185.28
BDA	81.12	86.37	140.99	95.27	297.35	90.08	425.72
CG	81.08	89.30	35.96	95.01	536.17	90.07	222.72
NS	81.00	87.37	54.78	95.06	318.58	90.78	176.78
sl-BAMM	81.04	89.75	**3.81**	95.11	**178.85**	90.13	**154.06**

Next, we test the toy example problem on high-dimensional case to illustrate the computational efficiency by increasing the dimension of $\mathbf{A}$. Here, we set the time limit as 10^4 s. Fig. 4 (a) shows the convergence time of different indicator $\|\mathbf{d_x}\| \leq 10^{-3}$, $\|F - F^*\|/\|F^*\| \leq 10^{-3}$ and $\|\mathbf{x} - \mathbf{x}^*\|/\|\mathbf{x}^*\| \leq 10^{-4}$. It can be observed that sl-BAMM saves much more time compared to RHG, CG and NS, which need multiple optimization iteration step on LL problem. Fig. 4 (b) illustrates the upper limit of problem dimension that existing methods can solve when reaching the time limit. Since the implementation of sl-BAMM does not contain repeated nested matrix-vector products, it is capable to handle BLO problems with higher dimension.

4.2. Real-world Applications

In this section, we verify the performance of sl-BAMM based on two widely used real-world BLO applications, including data hyper-cleaning and few-shot learning.

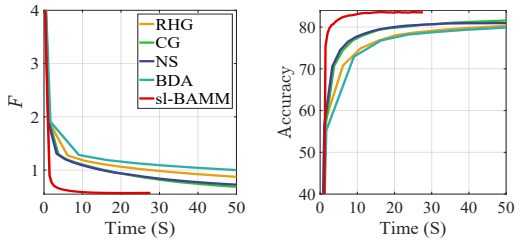


Figure 5. Comparison of the validation loss F and accuracy for hyper-cleaning on FashionMNIST dataset.

Data Hyper-cleaning. Assuming that some of the labels in the dataset are contaminated, data hyper-cleaning aims to reduce the impact of incorrect samples by adding hyper-parameters to label the corrupted data. In the experiment, we follow the general parameter settings in BDA (Liu et al., 2020) and conduct the experiments on FashionMNIST datasets. To demonstrate the significant improvement of sl-BAMM in terms of computational efficiency, in Tab. 1, we report the F1 scores and the time required for different methods to achieve the similar accuracy. It can be seen that sl-BAMM significantly reduces the time needed to achieve desired solution. Fig. 5 presents the validation loss and test accuracy for different methods. It can be seen that our method has the fastest convergence rate and keeps being the fastest at higher accuracies even after quickly achieving 81% accuracy rate.

Few-shot classification. The N-way M-shot classification aims to improve the fast-adaptability of hyper model such that the new task can be solved with comparable performance with quick updates based on few examples from each class. In the experiment, we conduct the 5-way 1-shot and 20-way 1-shot experiments on Omniglot dataset (Finn et al., 2017). In the right part of Tab. 1 , we report the runtime of existing methods to achieve the same accuracy (95% for 5-way and 90% for-20 way). It can be observed that sl-BAMM achieves similar accuracy while significantly reducing the calculation time, which demonstrates the applicability of sl-BAMM for high dimension complicated real-world deep-learning applications.

5. Conclusion

In this paper, a simple yet general single-loop gradient method, named sl-BAMM, is proposed for BLO, without replying on Lower-Level Strong Convexity condition. Our results also provide the theoretical guarantee for various BLOs in the machine learning context where a strong gradient boundedness assumption does not hold. We anticipate that the convergence analysis that we develop will be useful for analyzing other BLOs, e.g., stochastic BLOs, and the proposed algorithm will be useful for other applications such as neural architecture search and reinforcement learning.

Acknowledgements

Authors listed in alphabetical order. This work is partially supported by the National Natural Science Foundation of China (Nos. U22B2052, 12222106), the Shenzhen Science and Technology Program (No. RCYX20200714114700072), the Guangdong Basic and Applied Basic Research Foundation (No. 2022B1515020082), the Pacific Institute for the Mathematical Sciences (PIMS), the LiaoNing Revitalization Talents Program(No. 2022RG04) and the Fundamental Research Funds for the Central Universities.

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*, 2022b.

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

Chen, L., Xu, J., and Zhang, J. On bilevel optimization without lower-level strong convexity. *arXiv preprint arXiv:2301.00712*, 2023.

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

Chen, T., Sun, Y., Xiao, Q., and Yin, W. A single-timescale method for stochastic bilevel optimization. In *AISTATS*, 2022.

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

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

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

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

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

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

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

Gong, C., Liu, X., and Liu, Q. Automatic and harmless regularization with constrained and lexicographic optimization: A dynamic barrier approach. *NeurIPS*, 34: 29630–29642, 2021.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *NeurIPS*, 2014.

Grazzi, R., Franceschi, L., Pontil, M., and Salzo, S. On the iteration complexity of hypergradient computation. In *ICML*, 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. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Huang, F., Li, J., Gao, S., and Huang, H. Enhanced bilevel optimization via bregman distance. *NeurIPS*, 35:28928–28939, 2022.

Ji, K. and Liang, Y. Lower bounds and accelerated algorithms for bilevel optimization. *Journal of Machine Learning Research*, 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. *NeurIPS*, 33:11490–11500, 2020.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *ICML*, 2021.

Ji, K., Liu, M., Liang, Y., and Ying, L. Will bilevel optimizers benefit from loops. In *NeurIPS*, 2022.

Khanduri, P., Zeng, S., Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *NeurIPS*, 2021.

Kim, Y., Leyffer, S., and Munson, T. Mpec methods for bilevel optimization problems. In *Bilevel Optimization*, pp. 335–360. Springer, 2020.

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.

Liang, H., Zhang, S., Sun, J., He, X., Huang, W., Zhuang, K., and Li, Z. Darts+: Improved differentiable architecture search with early stopping. *arXiv preprint arXiv:1909.06035*, 2019.

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*, 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*, 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. PMLR, 2021b.

Liu, R., Liu, X., Zeng, S., Zhang, J., and Zhang, Y. Value-function-based sequential minimization for bi-level optimization. *arXiv preprint arXiv:2110.04974*, 2021c.

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

Liu, R., Liu, X., Zeng, S., Zhang, J., and Zhang, Y. Optimization-derived learning with essential convergence analysis of training and hyper-training. In *ICML*, pp. 13825–13856. PMLR, 2022a.

Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A general descent aggregation framework for gradient-based bi-level optimization. *IEEE TPAMI*, 2022b.

Liu, R., Liu, X., Zeng, S., Zhang, J., and Zhang, Y. Hierarchical optimization-derived learning. *arXiv preprint arXiv:2302.05587*, 2023a.

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

Lorraine, J., Vicol, P., and Duvenaud, D. Optimizing millions of hyperparameters by implicit differentiation. In *AISTATS*, 2020.

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.

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

Mehra, A. and Hamm, J. Penalty method for inversion-free deep bilevel optimization. In *Asian Conference on Machine Learning*, pp. 347–362. PMLR, 2021.

Nesterov, Y. *Lectures on convex optimization*, volume 137. Springer, 2018.

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

Outrata, J. V. On the numerical solution of a class of stackelberg problems. *Zeitschrift für Operations Research*, 34 (4):255–277, 1990.

Pedregosa, F. Hyperparameter optimization with approximate gradient. In *ICML*, 2016.

Pfau, D. and Vinyals, O. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

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

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

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

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*, 2022.

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

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

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

Zhang, J., Lin, H., Jegelka, S., Sra, S., and Jadbabaie, A. Complexity of finding stationary points of nonsmooth nonconvex functions. In *ICML*, pp. 11173–11182, 2020.

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

A. Expanded Related Work

We provide here a detailed review of recent studies that are closely related to ours.

BLO Beyond Lower-Level Strong Convexity (LLSC). The studies beyond LLSC are rather limited. A line of work (including this work) relaxed LLS by considering BLO problem with a merely convex LL objective. For example, by formulating bi-level models from the optimistic viewpoint and aggregating UL and LL objectives (called sequential averaging method in (Sabach & Shtern, 2017)), the authors in (Liu et al., 2020; 2022b) introduced Bilevel Descent Aggregation (BDA) framework with only the asymptotic convergence guarantee; see also a concurrent work in (Li et al., 2020). Another study (Mehra & Hamm, 2021) presented a penalty method for BLO, by replacing LL problem by its first-order stationarity condition, which avoids computing Hessian inverse. Recently, based on the value function approach, by using the gradient descent and ascent method via smooth approximation, Sow et al. (Sow et al., 2022) proposed a primal-dual bilevel optimization (PDBO) algorithm and its proximal version (Proximal-PDBO) without involving second-order Hessian and Jacobian computations. Further, they provided the convergence rate analysis for PDBO and Proximal-PDBO, under certain convexity-type conditions and compactness of the domain.

Another line of work considers BLO problem with a nonconvex LL objective. For example, the authors in (Liu et al., 2021b;c) proposed a bi-level value function based penalty and barrier method. Another study (Liu et al., 2021d) proposed an initialization auxiliary gradient-based algorithm to solve BLO problems with nonconvex LL objectives. Recently, Arbel and Mairal (Arbel & Mairal, 2022b) introduced a bilevel game that resolves the ambiguity in BLO with nonconvex objectives using the notion of selection maps. Further, they extended implicit differentiation to degenerate critical points, provided LL objective satisfies a generalization of the Morse-Bott property. However, only asymptotic result is shown in (Liu et al., 2021b;c;d; Arbel & Mairal, 2022b). Recently, based on the value function approach and a dynamic barrier gradient descent algorithm in (Gong et al., 2021), the authors in (Ye et al., 2022) proposed a simple first-order BLO algorithm that depends only on first-order gradient information and provided non-asymptotic convergence analysis for non-convex objectives under the global or local Polyak-Łojasiewicz conditions.

However, all of these approaches take a double-loop optimization structure (involving numerical LL optimization loops), which could be difficult to implement in practice as the number of required inner loop iterations are difficult to adjust. Differently from the above studies, our sl-BAMM algorithm has a single loop structure, which admits a much simpler implementation. Moreover, the updates in sl-BAMM evolve simultaneously and then can be performed in parallel instead of sequentially.

Stationarity Measure for BLO Algorithm. The most common stationarity measure for BLO is the gradient norm of the total UL objective (i.e., the norm of hypergradient) (Pedregosa, 2016; Ghadimi & Wang, 2018; Chen et al., 2021; Ji et al., 2021; 2022; Li et al., 2022; Dagréou et al., 2022; Arbel & Mairal, 2022a). These studies notoriously rely on LLSC. This was extended in (Huang et al., 2022) to a general BLO problem with a nonsmooth UL objective via Bregman distance, but still assuming LLSC. Further, when the total UL objective is either strongly convex or convex, the suboptimality gap (as measured by the total UL objective values) is used in (Ghadimi & Wang, 2018; Hong et al., 2023; Ji & Liang, 2022) and the optimality gap w.r.t. UL variable is used in (Hong et al., 2023) . Recently, under the weak convexity of the total UL objective, the gradient norm of the standard Moreau envelope is used in (Hong et al., 2023; Chen et al., 2022).

Without LL strongly convex assumption, the stationarity measure for BLO has been much less studied, from an algorithmic point of view. Several recent work (Sow et al., 2022; Ye et al., 2022) reformulate a BLO problem into a constrained problem by using value function approach, and develop non-asymptotic convergence to various notions of approximate KKT points, under stationarity measure involving value function (Ye et al., 2022) or regularized value function (Sow et al., 2022). Very recently, inspired by the recent developments of nonsmooth nonconvex optimization, the authors in (Chen et al., 2023) formalized the local optimality of BLO via the notion of Goldstein stationarity condition (Zhang et al., 2020) of the hyper-objective (i.e., the value function of a parameterized simple BLO problem), under some growth conditions. When LL problem is unconstrained and be convex, BLO problem is equivalent to an equality-constrained optimization problem by replacing LL problem by its first-order stationarity condition. This motivates us to use KKT condition of the resulting equality-constrained problem as a measure of stationarity of the solution returned by BLO algorithms. Note that this kind of reformulation is called MPEC formulation in the optimization theory for general BLO problems with LL inequality and equality constraints (Kim et al., 2020).

B. A Unified Proof Sketch of Theorems in Section 3

Theorems 3.4-3.8 in Section 3 provide various practical step size strategies for sl-BAMM, which enjoy favorably both extremely simpler implementation and theoretical convergence guarantee. In this section, we highlight the key steps of the proofs towards Theorems 3.4-3.8 **without the gradient boundedness assumption (GBA) and beyond lower-level strong convexity (LLSC)**, and highlight the differences between our new convergence analysis and the existing ones.

To analyze the convergence of sl-BAMM with or without LLSC, according to different step size strategies, we will identify different intrinsic Lyapunov (potential) functions, whose general form is given by

$$V_k := a_k \left[F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)) - F_0\right] + b_k\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + c_k\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2, \tag{15}$$

where $\mathbf{y}^*_\mu(\mathbf{x})$ is the unique minimizer of the aggregation function $\psi_\mu(\mathbf{x},\cdot) = \mu F(\mathbf{x},\cdot) + (1-\mu)f(\mathbf{x},\cdot)$, and $\mathbf{v}^*_\mu(\mathbf{x}) = \left[\nabla^2_{\mathbf{yy}}\psi_\mu(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}))\right]^{-1}\nabla_{\mathbf{y}}F\left(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x})\right)$ is the correct dual multiplier. Here $\{a_k,b_k,c_k\}_{k=1}^\infty$ is a sequence of positive constants chosen later depending on the step size strategy. In particular, when $f(\mathbf{x},\cdot)$ is strongly convex, we take $\mu_k = 0$ for all k and then $F(\mathbf{x},\mathbf{y}^*(\mathbf{x}))$ is the total UL objective. Generally, the first term $\Phi_\mu(\mathbf{x}) = F(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}))$ in the expression of V_k quantifies the (approximate) overall UL objective functions, the second term $\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2$ characterizes LL solution errors, and the third term $\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2$ delineates the multiplier errors.

The proofs of Theorem 3.4-3.8 contain three major steps, which include: (1) upper-bounding the descent of the total UL objective; (2) controlling both LL solution and multiplier errors in the descent of the total UL objective; and (3) combining all results in the previous steps and choosing suitable coefficients $\{a_k,b_k,c_k\}_{k=1}^\infty$ of Lyapunov function to prove the convergence guarantee. More detailed steps can be found as follows.

Step 1: upper-bounding the descent of the total UL objective.

We first analyze the descent of the approximate total UL objective as below, which does not require either LLSC or any assumption on the boundedness of $\nabla_{\mathbf{y}}F(\mathbf{x},\mathbf{y})$.

$$\begin{aligned}
&\Phi_{\mu_{k+1}}(\mathbf{x}_{k+1}) - \Phi_{\mu_k}(\mathbf{x}_k)\\
\leq& -\frac{\alpha_k}{2}\|\nabla\Phi_{\mu_k}(\mathbf{x}_k)\|^2 - \frac{1}{2}\left(\frac{1}{\alpha_k} - \frac{C_{\Phi 1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + C_{\Phi 2}}{\sigma^2_{\psi_{\mu_k}}}\right)\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 + \alpha_k L^2_{\psi_{\mathbf{y}1}}\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2\\
&+ \alpha_k\left(L_{\psi_{\mathbf{xy}2}}\left\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\| + L_{F_{\mathbf{x}2}}\right)^2\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{2\left\|\nabla_{\mathbf{y}}F(\mathbf{x}_{k+1},\mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2}{\sigma_F}\left(\frac{\mu_k-\mu_{k+1}}{\mu_k}\right),
\end{aligned} \tag{16}$$

where both LL solution and multiplier errors appear on the right hand side. Here $C_{\Phi 1}$ and $C_{\Phi 2}$ are some explicit positive constants given in Lemma D.3. Importantly, compared to the previous results using both LLSC and GBA, the coefficients of $\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2$ and $\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2$ depend on $\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|$ since Lipschitz continuity of the hypergradient and its surrogate could not be guaranteed without any assumption on the boundedness of $\nabla_{\mathbf{y}}F(\mathbf{x},\mathbf{y})$ and its variant. To address this issue, we characterize a weaker smoothness of $\Phi_\mu(\cdot)$ in Lemma D.3. Note also that there is an extra term in Inequality (16) involving the descent of the averaging parameter.

Step 2: controlling both LL solution and multiplier errors.

Noticing the above key points, we then estimate carefully LL solution error as below,

$$\begin{aligned}
\|\mathbf{y}_{k+1} - \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2 - \|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 \leq& -\frac{1}{2}\beta_k\sigma_{\psi_{\mu_k}}\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{6L^2_{\psi_{\mathbf{y}1}}}{\beta_k\sigma^3_{\psi_{\mu_k}}}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2\\
&+ \frac{24\|\nabla_{\mathbf{y}}F(\mathbf{x}_{k+1},\mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2}{\sigma^2_F\beta_k\sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k-\mu_{k+1}}{\mu_k}\right)^2,
\end{aligned} \tag{17}$$

and the multiplier error is bound by

$$\begin{aligned}
&\|\mathbf{v}_{k+1} - \mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2 - \|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\
\leq& -\frac{1}{2}\eta_k \sigma_{\psi_{\mu_k}} \|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{6\left(L_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{\mathbf{v}2}\right)^2}{\eta_k \sigma^5_{\psi_{\mu_k}}} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \\
&+ \frac{3\eta_k}{\sigma_{\psi_{\mu_k}}} \left(L_{\psi_{\mathbf{yy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{y}2}}\right)^2 \|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\
&+ \frac{6\left(C_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| + C_{\mathbf{v}2}\right)^2 \|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2}{\eta_k \sigma_{\psi_{\mu_k}}} \left(\frac{\mu_k - \mu_{k+1}}{\mu_k^2}\right)^2.
\end{aligned} \tag{18}$$

Here $C_{\mathbf{v}1}$ and $C_{\mathbf{v}2}$ are some explicit positive constants given in Lemma D.4.

Step 3: choosing suitable coefficients such that the descent of Lyapunov function is well controlled.

Combining the estimates in Steps 1 and 2, we can find suitable decreasing coefficients $\{a_k, b_k, c_k\}_{k=1}^{\infty}$ and the averaging parameter μ_k such that the descent of Lyapunov function is well controlled by a summable series as follows:

$$\begin{aligned}
V_{k+1} - V_k \leq& -\frac{a_{k+1}\alpha_k}{2}\|\nabla \Phi_{\mu_k}(\mathbf{x}_k)\|^2 - \frac{\widehat{\alpha}_k}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 - \frac{\widehat{\beta}_k}{2}\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 - \frac{\widehat{\eta}_k}{2}\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\
&+ \frac{2\left\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2 a_{k+1}}{\sigma_F}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right) \\
&+ \frac{24\left\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2 b_{k+1}}{\sigma_F^2 \beta_k \sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right)^2 \\
&+ \frac{6\left(C_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| + C_{\mathbf{v}2}\right)^2 \|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2 c_{k+1}}{\eta_k \sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k^2}\right)^2,
\end{aligned} \tag{19}$$

where the coefficients $\{\widehat{\alpha}_k, \widehat{\beta}_k, \widehat{\eta}_k\}$ explicitly given in (24) will be positive after choosing proper decreasing coefficients $\{a_k, b_k, c_k\}_{k=1}^{\infty}$ and the averaging parameter μ_k. Then telescoping leads to $V_K = O(1)$ as $K \to \infty$ and then

$$\sum_{k=K}^{2K+1} \frac{a_{k+1}\alpha_k}{2}\|\nabla \Phi_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{\widehat{\alpha}_k}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \frac{\widehat{\beta}_k}{2}\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{\widehat{\eta}_k}{2}\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 = O(1).$$

The desired convergence guarantee follows by analyzing the decay or grow rates of $a_{k+1}\alpha_k$, $\widehat{\alpha}_k, \widehat{\beta}_k, \widehat{\eta}_k$. We refer to next section for detailed proofs.

Note that Lyapunov function argument for BLO algorithms has been used in establishing the convergence rates of ALSET (Chen et al., 2021), STABLE (Chen et al., 2022), FSLA (Li et al., 2022), SOBA and SABA (Dagréou et al., 2022), and so on. In contrast to the above mentioned works, without GBA and LLSC, Lyapunov function with flexible coefficients must be considered.

C. Detailed Proofs

This section is devoted to proofs of Theorems 3.4-3.8 in Section 3.

C.1. Fundamental descent lemmas

We first analyze the descent of the approximate overall UL objective $\Phi_{\mu_k}(\mathbf{x}_k) = F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k))$ in the next lemma, which is the main result in Step 1 of proof sketch.

Lemma C.1. *Suppose Assumptions 3.1, and either Assumption 3.3 or Assumption 3.7 holds. Let $\mu_{k+1} \leq \mu_k \leq \frac{1}{2}$ for all k.*

Then the sequence of $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ *generated by sl-BAMM satisfies*

$$\begin{aligned}
&\Phi_{\mu_{k+1}}(\mathbf{x}_{k+1}) - \Phi_{\mu_k}(\mathbf{x}_k) \\
\leq& -\frac{\alpha_k}{2}\|\nabla\Phi_{\mu_k}(\mathbf{x}_k)\|^2 - \frac{1}{2}\left(\frac{1}{\alpha_k} - \frac{C_{\Phi 1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + C_{\Phi 2}}{\sigma^2_{\psi_{\mu_k}}}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \alpha_k L^2_{\psi_{\mathbf{y}1}}\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\
&+ \alpha_k\left(L_{\psi_{\mathbf{xy}2}}\left\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\| + L_{F_{\mathbf{x}2}}\right)^2\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{2\left\|\nabla_{\mathbf{y}}F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2}{\sigma_F}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right),
\end{aligned} \tag{20}$$

where both $C_{\Phi 1}$ *and* $C_{\Phi 2}$ *are constants given by*

$$\begin{aligned}
C_{\Phi 1} :=& L^2_{\psi_{\mathbf{y}1}}L_{\psi_{\mathbf{yy}2}} + L_{\psi_{\mathbf{y}1}}\left(L_{\psi_{\mathbf{yy}1}} + L_{\psi_{\mathbf{xy}2}}\right)\sigma_{\psi_\mu} + L_{\psi_{\mathbf{xy}1}}\sigma^2_{\psi_\mu}, \\
C_{\Phi 2} :=& L^2_{\psi_{\mathbf{y}1}}L_{F_{\mathbf{y}2}} + L_{\psi_{\mathbf{y}1}}\left(L_{F_{\mathbf{y}1}} + L_{F_{\mathbf{x}2}}\right)\sigma_{\psi_\mu} + L_{F_{\mathbf{x}1}}\sigma^2_{\psi_\mu}.
\end{aligned}$$

In particular, if LL objective $f(\mathbf{x}, \cdot)$ *is* σ_f*-strongly convex, i.e., Assumption 3.7 holds, we take* $\mu_k = 0$ *for all* k *and then the last term in Estimate (20) is redundant.*

Since the descent of the approximate overall UL objective depends on both LL solution and multiplier errors, we next analyze these errors, yielding the main result in Step 2 of proof sketch.

Lemma C.2. *Suppose Assumptions in Lemma C.1 hold. If we choose*

$$\beta_k \leq \frac{2}{\sigma_{\psi_{\mu_k}} + L_{\psi_{\mu_k}}}, \quad \eta_k \leq \frac{1}{L_{\psi_{\mu_k}}}.$$

Then the sequence of $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ *generated by sl-BAMM satisfies*

$$\begin{aligned}
\|\mathbf{y}_{k+1} - \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2 - \|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 \leq& -\frac{1}{2}\beta_k\sigma_{\psi_{\mu_k}}\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{6L^2_{\psi_{\mathbf{y}1}}}{\beta_k\sigma^3_{\psi_{\mu_k}}}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \\
&+ \frac{24\|\nabla_{\mathbf{y}}F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2}{\sigma^2_F\beta_k\sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right)^2,
\end{aligned} \tag{21}$$

and

$$\begin{aligned}
&\|\mathbf{v}_{k+1} - \mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2 - \|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\
\leq& -\frac{1}{2}\eta_k\sigma_{\psi_{\mu_k}}\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{6\left(L_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{\mathbf{v}2}\right)^2}{\eta_k\sigma^5_{\psi_{\mu_k}}}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \\
&+ \frac{3\eta_k}{\sigma_{\psi_{\mu_k}}}\left(L_{\psi_{\mathbf{yy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{y}2}}\right)^2\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\
&+ \frac{6\left(C_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| + C_{\mathbf{v}2}\right)^2\|\nabla_{\mathbf{y}}F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2}{\eta_k\sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k^2}\right)^2,
\end{aligned} \tag{22}$$

where both $C_{\mathbf{v}1}$ *and* $C_{\mathbf{v}2}$ *are constants given by*

$$C_{\mathbf{v}1} := 2\left(L_{F_{\mathbf{yy}2}} + L_{f_{\mathbf{yy}2}}\right)/\sigma^2_F, \quad C_{\mathbf{v}2} := 2\left(2L_{F_{\mathbf{y}2}} + L_{f_{\mathbf{y}2}}\right)/\sigma^2_F.$$

In particular, if LL objective $f(\mathbf{x}, \cdot)$ *is* σ_f*-strongly convex, we take* $\mu_k = 0$ *for all* k *and then the last terms in Estimates (21)-(22) are both redundant.*

Combining the estimates in Steps 1 and 2, for any decreasing positive coefficients $\{a_k, b_k, c_k\}_{k=1}^{\infty}$, we get the following descent of Lyapunov function, which is the key step to improving the existing result.

Lemma C.3. *Suppose Assumptions in Lemma C.2 hold. Let $\{a_k, b_k, c_k\}_{k=1}^{\infty}$ be a sequence of decreasing positive constants, then the following descent of Lyapunov function holds:*

$$\begin{aligned} V_{k+1} - V_k \leq & -\frac{a_{k+1}\alpha_k}{2}\|\nabla\Phi_{\mu_k}(\mathbf{x}_k)\|^2 - \frac{\widehat{\alpha}_k}{2}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 - \frac{\widehat{\beta}_k}{2}\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 - \frac{\widehat{\eta}_k}{2}\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\ & + \frac{2\left\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2 a_{k+1}}{\sigma_F}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right) \\ & + \frac{24\left\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2 b_{k+1}}{\sigma_F^2\beta_k\sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right)^2 \\ & + \frac{6\left(C_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| + C_{\mathbf{v}2}\right)^2\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2 c_{k+1}}{\eta_k\sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k^2}\right)^2, \end{aligned} \tag{23}$$

where the coefficients $\{\widehat{\alpha}_k, \widehat{\beta}_k, \widehat{\eta}_k\}$ are given as below.

$$\begin{aligned} \widehat{\alpha}_k :=& \frac{a_{k+1}}{\alpha_k} - \frac{a_{k+1}\left(C_{\Phi 1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + C_{\Phi 2}\right)}{\sigma^2_{\psi_{\mu_k}}} - \frac{12L^2_{\psi_{\mathbf{y}1}} b_{k+1}}{\beta_k\sigma^3_{\psi_{\mu_k}}} - \frac{12\left(L_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{\mathbf{v}2}\right)^2 c_{k+1}}{\eta_k\sigma^5_{\psi_{\mu_k}}}, \\ \widehat{\beta}_k :=& b_{k+1}\beta_k\sigma_{\psi_{\mu_k}} - 2a_{k+1}\alpha_k\left(L_{\psi_{\mathbf{xy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{x}2}}\right)^2 - 6\left(L_{\psi_{\mathbf{yy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{y}2}}\right)^2\frac{c_{k+1}\eta_k}{\sigma_{\psi_{\mu_k}}}, \\ \widehat{\eta}_k :=& c_{k+1}\eta_k\sigma_{\psi_{\mu_k}} - 2a_{k+1}\alpha_k L^2_{\psi_{\mathbf{y}1}}. \end{aligned} \tag{24}$$

Here $L_{\mathbf{v}1} := L_{\psi_{\mathbf{yy}2}}L_{\psi_{\mathbf{y}1}} + L_{\psi_{\mathbf{yy}1}}\sigma_{\psi_\mu}$ and $L_{\mathbf{v}2} := L_{F_{\mathbf{y}2}}L_{\psi_{\mathbf{y}1}} + L_{F_{\mathbf{y}1}}\sigma_{\psi_\mu}$. In particular, if LL objective $f(\mathbf{x}, \cdot)$ is σ_f-strongly convex, we take $\mu_k = 0$ for all k and then the terms involving $\mu_k - \mu_{k+1}$ in Estimate (23) are all redundant.

C.2. A general setting of step sizes

To simplify and unify the proofs of Theorems 3.4-3.8, we introduce our framework in which all of the step size strategies in Theorems 3.4-3.8 have a unified form as below,

$$\beta_k \in \left[\bar{\beta}(k+1)^{-\tau/2}, 1/\left(L_{F_{\mathbf{y}2}} + L_{f_{\mathbf{y}2}}\right)\right] =: I_\beta, \quad \eta_k = c_\eta\frac{b_{k+1}}{c_{k+1}}\beta_k\sigma^2_{\psi_{\mu_k}}, \quad \alpha_k = c_\alpha\frac{a_{k+1}}{c_{k+1}}\eta_k\sigma^5_{\psi_{\mu_k}}, \tag{25}$$

where $a_{k+1}, b_{k+1}, c_{k+1}$ are the coefficients of Lyapunov function (15), and the positive constants c_η, c_α may depend on k. Here and in what follows, we drop their dependency in k for clarity. Actually, for any fixed step size strategy $(\beta_k, \eta_k, \alpha_k)$ with $\beta_k \in I_\beta$, after defining

$$c_\eta = \left(\frac{c_{k+1}}{b_{k+1}}\right)\left(\frac{\eta_k}{\beta_k}\right)\sigma^{-2}_{\psi_{\mu_k}}, \quad c_\alpha = \left(\frac{c_{k+1}}{a_{k+1}}\right)\left(\frac{\alpha_k}{\eta_k}\right)\sigma^{-5}_{\psi_{\mu_k}}, \tag{26}$$

the step size strategy $(\beta_k, \eta_k, \alpha_k)$ has the form (25).

Thanks to the setting (25), the key coefficients $\{\widehat{\alpha}_k, \widehat{\beta}_k, \widehat{\eta}_k\}$ in Estimate (23) have the following simplified form:

$$\widehat{\alpha}_k = \left(1 - c_\alpha\left(12\left(L_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{\mathbf{v}2}\right)^2 + \frac{12L^2_{\psi_{\mathbf{y}1}} b_{k+1}\alpha_k}{c_\alpha a_{k+1}\beta_k\sigma^3_{\psi_{\mu_k}}} + \frac{\alpha_k}{c_\alpha\sigma^2_{\psi_{\mu_k}}}\left(C_{\Phi 1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + C_{\Phi 2}\right)\right)\right)\frac{a_{k+1}}{\alpha_k}, \tag{27}$$

$$\widehat{\beta}_k = \left(1 - c_\eta\left(6\left(L_{\psi_{\mathbf{yy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{y}2}}\right)^2 + \frac{2a_{k+1}\alpha_k}{c_\eta b_{k+1}\beta_k\sigma_{\psi_{\mu_k}}}\left(L^2_{\psi_{\mathbf{xy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{x}2}}\right)^2\right)\right) b_{k+1}\beta_k\sigma_{\psi_{\mu_k}}, \tag{28}$$

$$\widehat{\eta}_k = \left(1 - 2L^2_{\psi_{\mathbf{y}1}}\frac{a_{k+1}\alpha_k}{c_{k+1}\eta_k\sigma_{\psi_{\mu_k}}}\right)c_{k+1}\eta_k\sigma_{\psi_{\mu_k}}. \tag{29}$$

For any fixed step size strategy $(\beta_k, \eta_k, \alpha_k)$, to control the descent of Lyapunov function very well, it is helpful to find suitable coefficients $a_{k+1}, b_{k+1}, c_{k+1}$ such that

$$\widehat{\alpha}_k \geq \frac{a_{k+1}}{2\alpha_k}, \quad \widehat{\beta}_k \geq \frac{1}{2}b_{k+1}\beta_k\sigma_{\psi_{\mu_k}}, \quad \widehat{\eta}_k \geq \frac{1}{2}c_{k+1}\eta_k\sigma_{\psi_{\mu_k}}, \tag{30}$$

and all of the sequences involving $\mu_k - \mu_{k+1}$ in Estimate (23) (denoted by $A_\mu^k, B_\mu^k, C_\mu^k$ respectively) are summable. If this is well done, from Estimate (23), one first get

$$V_{k+1} - V_k \leq A_\mu^k + B_\mu^k + C_\mu^k. \tag{31}$$

By the definition (15), we get $V_k \geq 0$ for all k and then

$$V_K \leq V_{K_0} + \sum_{k=1}^{\infty} \left(A_\mu^k + B_\mu^k + C_\mu^k\right) \implies V_K = O(1) \text{ as } K \to \infty. \tag{32}$$

Further, by using Estimate (23) again, one can get

$$\sum_{k=K}^{2K+1} \left(\frac{a_{k+1}\alpha_k}{2} \|\nabla \Phi_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{\widehat{\alpha}_k}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \frac{\widehat{\beta}_k}{2} \|\mathbf{y}_k - \mathbf{y}_{\mu_k}^*(\mathbf{x}_k)\|^2 + \frac{\widehat{\eta}_k}{2} \|\mathbf{v}_k - \mathbf{v}_{\mu_k}^*(\mathbf{x}_k)\|^2 \right) = O(1). \tag{33}$$

Then Estimate (30) yields

$$\begin{aligned} \sum_{k=K}^{2K+1} \Big(& 2a_{k+1}\alpha_k \|\nabla \Phi_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{a_{k+1}}{\alpha_k} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ & + b_{k+1}\beta_k \sigma_{\psi_{\mu_k}} \|\mathbf{y}_k - \mathbf{y}_{\mu_k}^*(\mathbf{x}_k)\|^2 + c_{k+1}\eta_k \sigma_{\psi_{\mu_k}} \|\mathbf{v}_k - \mathbf{v}_{\mu_k}^*(\mathbf{x}_k)\|^2 \Big) = O(1). \end{aligned} \tag{34}$$

Now, by lower-bounding the coefficients in Estimate (34), according to different step size strategies and the coefficients of Lyapunov function, we can get the estimates on

$$\min_{0 \leq k \leq K} \|\nabla \Phi_{\mu_k}(\mathbf{x}_k)\|^2 + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \|\mathbf{y}_k - \mathbf{y}_{\mu_k}^*(\mathbf{x}_k)\|^2 + \|\mathbf{v}_k - \mathbf{v}_{\mu_k}^*(\mathbf{x}_k)\|^2,$$

by choosing different upper bounds on p and τ. Finally, by the optimality of $\mathbf{y}_{\mu_k}^*(\mathbf{x}_k)$ and the definition of $\mathbf{v}_{\mu_k}^*(\mathbf{x}_k)$, there exists a positive constant C independent of k such that

$$\mathrm{KKT}(\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k) \leq C \left(\|\nabla \Phi_{\mu_k}(x_k)\|^2 + \|\mathbf{y}_k - \mathbf{y}_{\mu_k}^*(\mathbf{x}_k)\|^2 + \|\mathbf{v}_k - \mathbf{v}_{\mu_k}^*(\mathbf{x}_k)\|^2 + \mu_k^2 \right), \tag{35}$$

which yields the convergence rates measured by KKT residual.

C.3. Proofs of Theorems 3.4-3.8

For the convenience of the reader, we first give an overview of the step sizes strategies in the main theorems and of the well-chosen coefficients of Lyapunov function in Table 2.

Table 2. Summary of the step size strategies and the well-chosen Lyapunov coefficients.

Strategy	Step sizes			Coefficients of Lyapunov function		
	β_k	η_k	α_k	a_{k+1}	b_{k+1}	c_{k+1}
Unified form	β_k	$c_\eta \frac{b_{k+1}}{c_{k+1}} \beta_k \sigma_{\psi_{\mu_k}}^2$	$c_\alpha \frac{a_{k+1}}{c_{k+1}} \eta_k \sigma_{\psi_{\mu_k}}^5$	a_{k+1}	b_{k+1}	c_{k+1}
S1 (Thm 3.4)	$\beta_k \in I_\beta$	$(k+1)^{-\tau/2} \beta_k \mu_k^2$	$(k+1)^{-3\tau/2} \beta_k \mu_k^7$	$(k+1)^{-\tau} \sigma_{\psi_{\mu_k}}^2$	$(k+1)^{-\tau/2} \sigma_{\psi_{\mu_k}}^2$	$(k+1)^{-\tau} \sigma_{\psi_{\mu_k}}^6$
S2 (Thm 3.5)	$\beta_k \in I_\beta$	$(k+1)^{-\tau/2} \beta_k \mu_k$	$(k+1)^{-3\tau/2} \beta_k \mu_k^5$	$(k+1)^{-\tau}$	$(k+1)^{-\tau/2}$	$(k+1)^{-\tau} \sigma_{\psi_{\mu_k}}^3$
S3 (Thm 3.6)	$\beta_k \in I_\beta$	$(k+1)^{-\tau/2} \beta_k$	$(k+1)^{-3\tau/2} \beta_k \mu_k^3$	$(k+1)^{-\tau}$	$(k+1)^{-\tau/2}$	$(k+1)^{-\tau} \sigma_{\psi_{\mu_k}}^2$
Thm 3.10	$\beta_k \in I_\beta$	$\bar{\eta}(k+1)^{-\tau/2} \beta_k$	$\bar{\alpha}(k+1)^{-\tau} \beta_k$	1	1	1

Using the simplified form (27) of $\widehat{\alpha}_k, \widehat{\beta}_k, \widehat{\eta}_k$, note that $\frac{a_{k+1}\alpha_k}{c_{k+1}\eta_k \sigma_{\psi_{\mu_k}}} = \frac{a_{k+1}\alpha_k}{c_\eta b_{k+1} \beta_k \sigma_{\psi_{\mu_k}}^3} = c_\alpha \left(\frac{a_{k+1}}{c_{k+1}}\right)^2 \sigma_{\psi_{\mu_k}}^4$ and

$$\frac{b_{k+1}\alpha_k}{c_\alpha a_{k+1} \beta_k \sigma_{\psi_{\mu_k}}^3} = c_\eta \left(\frac{b_{k+1}}{c_{k+1}}\right)^2 \sigma_{\psi_{\mu_k}}^4, \quad \frac{\alpha_k}{c_\alpha \sigma_{\psi_{\mu_k}}^2} = \left(\frac{a_{k+1}}{c_{k+1}}\right) \eta_k \sigma_{\psi_{\mu_k}}^3, \quad \frac{a_{k+1}\alpha_k}{c_\eta b_{k+1} \beta_k \sigma_{\psi_{\mu_k}}} = c_\alpha \left(\frac{a_{k+1}}{c_{k+1}}\right)^2 \sigma_{\psi_{\mu_k}}^6,$$

Table 3. Estimates for the proof of Estimate (30).

Strategy	c_α	c_η	$\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|$	$\frac{b_{k+1}\alpha_k}{c_\alpha a_{k+1}\beta_k\sigma^3_{\psi_{\mu_k}}}$	$\frac{\alpha_k}{c_\alpha\sigma^2_{\psi_{\mu_k}}}$	$\frac{a_{k+1}\alpha_k}{c_{k+1}\eta_k\sigma_{\psi_{\mu_k}}}$
S1 (Thm 3.4)	$(k+1)^{-\tau}\mu_k^4$	$(k+1)^{-\tau}\mu_k^4$	μ_k^{-2}	1	$(k+1)^{-\tau/2}\beta_k\mu_k$	$(k+1)^{-\tau}$
S2 (Thm 3.5)	$(k+1)^{-\tau}\mu_k^2$	$(k+1)^{-\tau}\mu_k^2$	μ_k^{-1}	1	$(k+1)^{-\tau/2}\beta_k\mu_k$	$(k+1)^{-\tau}$
S3 (Thm 3.6)	$(k+1)^{-\tau}$	$(k+1)^{-\tau}$	1	1	$(k+1)^{-\tau/2}\beta_k\mu_k$	$(k+1)^{-\tau/2}$
Thm 3.10	$(k+1)^{-\tau/2}$	$(k+1)^{-\tau/2}$	1	$(k+1)^{-\tau/2}$	$(k+1)^{-\tau/2}$	$(k+1)^{-\tau/2}$

we can prove that Estimate (30) holds in the setting of Theorems 3.4-3.8, with the help of the estimates in Table 3.

Next we show that all of the sequences involving $\mu_k - \mu_{k+1}$ in Estimate (23) (denoted by $A^k_\mu, B^k_\mu, C^k_\mu$ respectively) are summable. Recall that

$$A^k_\mu := \frac{2\left\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2 a_{k+1}}{\sigma_F}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right),$$

$$B^k_\mu := \frac{24\left\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2 b_{k+1}}{\sigma_F^2\beta_k\sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right)^2,$$

$$C^k_\mu := \frac{6\left(C_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| + C_{\mathbf{v}2}\right)^2\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2 c_{k+1}}{\eta_k\sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k^2}\right)^2.$$

Using the step sizes strategies and the well-chosen coefficients of Lyapunov function in Table 2, we can get the explicit estimates of $A^k_\mu, B^k_\mu, C^k_\mu$ and the ones for the coefficients in Estimate (34), which is summarized in Table 4. Finally, by

Table 4. Important estimates used in the proofs of Theorems 3.4-3.8.

Strategy	A^k_μ	B^k_μ	C^k_μ	$a_{k+1}\alpha_k$	$\frac{1}{\alpha_k^2}$	$\frac{b_{k+1}\beta_k\sigma_{\psi_{\mu_k}}}{a_{k+1}\alpha_k}$	$\frac{c_{k+1}\eta_k\sigma_{\psi_{\mu_k}}}{a_{k+1}\alpha_k}$	$\frac{a_{k+1}}{\alpha_k}$
S1 (Thm 3.4)	$k^{-1-\tau}$	k^{-2+p}	k^{-2+5p}	$k^{-9p-3\tau}$	$k^{14p+3\tau}$	$k^{6p+2\tau}$	k^τ	$k^{5p+\tau/2}$
S2 (Thm 3.5)	$k^{-1-\tau}$	k^{-2+p}	k^{-2+3p}	$k^{-5p-3\tau}$	$k^{10p+3\tau}$	$k^{4p+2\tau}$	k^τ	$k^{5p+\tau/2}$
S3 (Thm 3.6)	$k^{-1-\tau}$	$k^{-2-\tau}$	$k^{-2-\tau}$	$k^{-3p-3\tau}$	$k^{4p+2\tau}$	k^τ	k^τ	$k^{3p+\tau/2}$
Thm 3.10	0	0	0	$k^{-3\tau/2}$	$k^{2\tau}$	k^τ	$k^{\tau/2}$	k^τ

using the following inequality

$$\sum_{k=K}^{2K+1}(k+1)^{-s} \geq \int_{K+1}^{2K+2} t^{-s}dt = \frac{2^{1-s}-1}{1-s}(K+1)^{1-s}, \tag{36}$$

we can get the estimates on

$$\min_{K\leq k\leq 2K+1}\|\nabla\Phi_{\mu_k}(\mathbf{x}_k)\|^2 + \|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 + \|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2,$$

and then the convergence rate results in Theorems 3.4-3.8 hold.

D. Proofs of fundamental descent lemmas

This section is devoted to the proofs of fundamental descent lemmas in Section C.1. Note that, unlike existing analyses, we do not require a strong assumption on the boundedness of $\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$. Hence, our theoretical analysis can be applied to a wider variety of applications in deep learning. To handle general BLOs, e.g., BLOs with multiple LL minimizers, we have taken the following standard assumption as in (Liu et al., 2020).

Assumption D.1 (Assumption 3.2). For any $\mathbf{x}$, the aggregation function $\psi_\mu(\mathbf{x},\cdot) = \mu F(\mathbf{x},\cdot) + (1-\mu) f(\mathbf{x},\cdot)$ is strongly convex for $\mu = 0$ or $\mu > 0$.

Under this assumption, the aggregation function $\psi_\mu(\mathbf{x},\mathbf{y}) = \mu F(\mathbf{x},\mathbf{y}) + (1-\mu) f(\mathbf{x},\mathbf{y})$ is σ_{ψ_μ}-strongly convex in $\mathbf{y}$ with $\sigma_{\psi_\mu} = \mu\sigma_F + (1-\mu)\sigma_f$. Hence $\psi_\mu(\mathbf{x},\mathbf{y})$ has a unique minimal point for every $\mathbf{x}$, denoted by $\mathbf{y}^*_\mu(\mathbf{x})$. Furthermore, under the smoothness assumption of f and F, $\psi_\mu(\mathbf{x},\mathbf{y})$ is also L_{ψ_μ}-smooth in $\mathbf{y}$ for any fixed $\mathbf{x}$.

D.1. Auxiliary Lemmas

Lemma D.2. *Suppose Assumptions 3.1, and either Assumption 3.3 or Assumption 3.7 holds, the following statements hold.*

(i) The function $\mathbf{y}^*_\mu(\mathbf{x})$ is Lipschitz continuous in $\mathbf{x}$, i.e., for all $\mathbf{x},\mathbf{x}'$,

$$\left\|\mathbf{y}^*_\mu(\mathbf{x}) - \mathbf{y}^*_\mu(\mathbf{x}')\right\| \le \frac{L_{\psi_{\mathbf{y}1}}}{\sigma_{\psi_\mu}}\|\mathbf{x}-\mathbf{x}'\|, \tag{37}$$

where both $L_{\psi_{\mathbf{y}1}}$ and σ_{ψ_μ} are constants given by

$$\sigma_{\psi_\mu} = \mu\sigma_F + (1-\mu)\sigma_f, \quad L_{\psi_{\mathbf{y}1}} = \mu L_{F_{\mathbf{y}1}} + (1-\mu)L_{f_{\mathbf{y}1}}.$$

(ii) The function $\mathbf{v}^*_\mu(\mathbf{x})$ satisfies

$$\|\mathbf{v}^*_\mu(\mathbf{x})\| = \left\|\left[\nabla^2_{\mathbf{yy}}\psi_\mu(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}))\right]^{-1}\nabla_{\mathbf{y}}F(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}))\right\| \le \frac{\|\nabla_{\mathbf{y}}F(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}))\|}{\sigma_{\psi_\mu}}, \tag{38}$$

and for all $\mathbf{x},\mathbf{x}'$,

$$\left\|\mathbf{v}^*_\mu(\mathbf{x}) - \mathbf{v}^*_\mu(\mathbf{x}')\right\| \le \frac{\left(L_{\mathbf{v}1}\|\mathbf{v}^*_\mu(\mathbf{x})\| + L_{\mathbf{v}2}\right)}{\sigma^2_{\psi_\mu}}\|\mathbf{x}-\mathbf{x}'\|, \tag{39}$$

where both $L_{\mathbf{v}1} := L_{\psi_{\mathbf{yy}2}}L_{\psi_{\mathbf{y}1}} + L_{\psi_{\mathbf{yy}1}}\sigma_{\psi_\mu}$ and $L_{\mathbf{v}2} := L_{F_{\mathbf{y}2}}L_{\psi_{\mathbf{y}1}} + L_{F_{\mathbf{y}1}}\sigma_{\psi_\mu}$ are constants.

Proof. (i) By the optimality of $\mathbf{y}^*_\mu$ to LL problem, we have $\nabla_{\mathbf{y}}\psi_\mu(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x})) = 0$. Thus

$$\left(\nabla_{\mathbf{y}}\psi_\mu(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x})) - \nabla_{\mathbf{y}}\psi_\mu(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}'))\right) + \left(\nabla_{\mathbf{y}}\psi_\mu(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}')) - \nabla_{\mathbf{y}}\psi_\mu(\mathbf{x}',\mathbf{y}^*_\mu(\mathbf{x}'))\right) = 0.$$

Multiplying the above equation by $\mathbf{y}^*_\mu(\mathbf{x}) - \mathbf{y}^*_\mu(\mathbf{x}')$, by the σ_{ψ_μ}-strong convexity of $\psi_\mu(\mathbf{x},\cdot)$ and $L_{\psi_{y1}}$-smoothness of $\psi_\mu(\cdot,\mathbf{y})$, we have

$$\begin{aligned}\sigma_{\psi_\mu}\|\mathbf{y}^*_\mu(\mathbf{x}) - \mathbf{y}^*_\mu(\mathbf{x}')\|^2 \le& \|\nabla_{\mathbf{y}}\psi_\mu(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}')) - \nabla_{\mathbf{y}}\psi_\mu(\mathbf{x}',\mathbf{y}^*_\mu(\mathbf{x}'))\| \cdot \|\mathbf{y}^*_\mu(\mathbf{x}) - \mathbf{y}^*_\mu(\mathbf{x}')\| \\ \le& L_{\psi_{\mathbf{y}1}}\|\mathbf{x}-\mathbf{x}'\| \cdot \|\mathbf{y}^*_\mu(\mathbf{x}) - \mathbf{y}^*_\mu(\mathbf{x}')\|.\end{aligned}$$

Then the conclusion follows immediately from the above inequality.

(ii) First, by the σ_{ψ_μ}-strong convexity of $\psi_\mu(\mathbf{x},\cdot)$ and Cauchy-Schwarz inequality, we have

$$\sigma_{\psi_\mu}\|\mathbf{v}^*_\mu(\mathbf{x})\|^2 \le \left\langle \mathbf{v}^*_\mu(\mathbf{x}), \nabla^2_{\mathbf{yy}}\psi_\mu(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}))\mathbf{v}^*_\mu(\mathbf{x})\right\rangle = \left\langle \mathbf{v}^*_\mu(\mathbf{x}), \nabla_{\mathbf{y}}F(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}))\right\rangle \le \|\mathbf{v}^*_\mu(\mathbf{x})\| \cdot \|\nabla_{\mathbf{y}}F(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}))\|,$$

which implies Inequality (38).

Second, the definition of $\mathbf{v}^*_\mu(\mathbf{x})$ says that $\left[\nabla^2_{\mathbf{yy}}\psi_\mu(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}))\right]\mathbf{v}^*_\mu(\mathbf{x}) = \nabla_{\mathbf{y}}F(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}))$. Thus

$$\begin{aligned}&\left[\nabla^2_{\mathbf{yy}}\psi_\mu(\mathbf{x}',\mathbf{y}^*_\mu(\mathbf{x}'))\right]\left[\mathbf{v}^*_\mu(\mathbf{x}') - \mathbf{v}^*_\mu(\mathbf{x})\right] \\ =& \left[\nabla_{\mathbf{y}}F(\mathbf{x}',\mathbf{y}^*_\mu(\mathbf{x}')) - \nabla_{\mathbf{y}}F(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x}))\right] + \left[\nabla^2_{\mathbf{yy}}\psi_\mu(\mathbf{x},\mathbf{y}^*_\mu(\mathbf{x})) - \nabla^2_{\mathbf{yy}}\psi_\mu(\mathbf{x}',\mathbf{y}^*_\mu(\mathbf{x}'))\right]\mathbf{v}^*_\mu(\mathbf{x}).\end{aligned}$$

By (i), the σ_{ψ_μ}-strong convexity of $\psi_\mu(\mathbf{x}', \cdot)$ and Lipschitz continuity of $\nabla_{\mathbf{y}}\psi_\mu$ and $\nabla_{\mathbf{y}}F$, we have

$$\begin{aligned}\sigma_{\psi_\mu}\|\mathbf{v}_\mu^*(\mathbf{x}) - \mathbf{v}_\mu^*(\mathbf{x}')\| \leq& \left(L_{F_{\mathbf{y}1}}\|\mathbf{x}-\mathbf{x}'\| + L_{F_{\mathbf{y}2}}\|\mathbf{y}_\mu^*(\mathbf{x}) - \mathbf{y}_\mu^*(\mathbf{x}')\|\right)\\ &+ \|\mathbf{v}_\mu^*(\mathbf{x})\|\left(L_{\psi_{\mathbf{yy}1}}\|\mathbf{x}-\mathbf{x}'\| + L_{\psi_{\mathbf{yy}2}}\|\mathbf{y}_\mu^*(\mathbf{x}) - \mathbf{y}_\mu^*(\mathbf{x}')\|\right)\\ \leq& \frac{\|\mathbf{x}-\mathbf{x}'\|}{\sigma_{\psi_\mu}}\Big(\|\mathbf{v}_\mu^*(\mathbf{x})\|\left(L_{\psi_{\mathbf{yy}2}}L_{\psi_{\mathbf{y}1}} + \sigma_{\psi_\mu}L_{\psi_{\mathbf{yy}1}}\right) + L_{F_{\mathbf{y}2}}L_{\psi_{\mathbf{y}1}} + \sigma_{\psi_\mu}L_{F_{\mathbf{y}1}}\Big),\end{aligned}$$

which implies the desired result.

□

Next, we prove a lemma that characterizes the smoothness of $\Phi_\mu(\mathbf{x})$ in $\mathbf{x}$, without any boundedness assumption of $\nabla_{\mathbf{y}}F(\mathbf{x},\mathbf{y})$.

Lemma D.3. *Suppose Assumptions 3.1, and either Assumption 3.3 or Assumption 3.7 holds, we have*

$$\|\nabla\Phi_\mu(\mathbf{x}) - \nabla\Phi_\mu(\mathbf{x}')\| \leq \frac{\left(C_{\Phi1}\|\mathbf{v}_\mu^*(\mathbf{x})\| + C_{\Phi2}\right)}{\sigma_{\psi_\mu}^2}\|\mathbf{x}-\mathbf{x}'\|, \tag{40}$$

where both $C_{\Phi1}$ and $C_{\Phi2}$ are constants given by

$$\begin{aligned}C_{\Phi1} :=& L_{\psi_{\mathbf{y}1}}^2 L_{\psi_{\mathbf{yy}2}} + L_{\psi_{\mathbf{y}1}}\left(L_{\psi_{\mathbf{yy}1}} + L_{\psi_{\mathbf{xy}2}}\right)\sigma_{\psi_\mu} + L_{\psi_{\mathbf{xy}1}}\sigma_{\psi_\mu}^2,\\ C_{\Phi2} :=& L_{\psi_{\mathbf{y}1}}^2 L_{F_{\mathbf{y}2}} + L_{\psi_{\mathbf{y}1}}\left(L_{F_{\mathbf{y}1}} + L_{F_{\mathbf{x}2}}\right)\sigma_{\psi_\mu} + L_{F_{\mathbf{x}1}}\sigma_{\psi_\mu}^2.\end{aligned}$$

Proof. Recall that $\nabla\Phi_\mu(\mathbf{x}) = \nabla_{\mathbf{x}}F(\mathbf{x},\mathbf{y}_\mu^*(\mathbf{x})) - \left[\nabla_{\mathbf{xy}}^2\psi_\mu(\mathbf{x},\mathbf{y}_\mu^*(\mathbf{x}))\right]\mathbf{v}_\mu^*(\mathbf{x})$. Thus

$$\begin{aligned}\nabla\Phi_\mu(\mathbf{x}) - \nabla\Phi_\mu(\mathbf{x}') =& \nabla_{\mathbf{x}}F(\mathbf{x},\mathbf{y}_\mu^*(\mathbf{x})) - \nabla_{\mathbf{x}}F(\mathbf{x}',\mathbf{y}_\mu^*(\mathbf{x}'))\\ &+ \left[\nabla_{\mathbf{xy}}^2\psi_\mu(\mathbf{x}',\mathbf{y}_\mu^*(\mathbf{x}'))\right]\mathbf{v}_\mu^*(\mathbf{x}') - \left[\nabla_{\mathbf{xy}}^2\psi_\mu(\mathbf{x},\mathbf{y}_\mu^*(\mathbf{x}))\right]\mathbf{v}_\mu^*(\mathbf{x}).\end{aligned}$$

First, by Lipschitz continuity of $\nabla_{\mathbf{x}}F$, we have

$$\|\nabla_{\mathbf{x}}F(\mathbf{x},\mathbf{y}_\mu^*(\mathbf{x})) - \nabla_{\mathbf{x}}F(\mathbf{x}',\mathbf{y}_\mu^*(\mathbf{x}'))\| \leq L_{F_{\mathbf{x}1}}\|\mathbf{x}-\mathbf{x}'\| + L_{F_{\mathbf{x}2}}\|\mathbf{y}_\mu^*(\mathbf{x}) - \mathbf{y}_\mu^*(\mathbf{x}')\|.$$

Second, note that

$$\begin{aligned}&\left[\nabla_{\mathbf{xy}}^2\psi_\mu(\mathbf{x}',\mathbf{y}_\mu^*(\mathbf{x}'))\right]\mathbf{v}_\mu^*(\mathbf{x}') - \left[\nabla_{\mathbf{xy}}^2\psi_\mu(\mathbf{x},\mathbf{y}_\mu^*(\mathbf{x}))\right]\mathbf{v}_\mu^*(\mathbf{x})\\ =&\left[\nabla_{\mathbf{xy}}^2\psi_\mu(\mathbf{x}',\mathbf{y}_\mu^*(\mathbf{x}'))\right]\left[\mathbf{v}_\mu^*(\mathbf{x}') - \mathbf{v}_\mu^*(\mathbf{x})\right] + \left[\nabla_{\mathbf{xy}}^2\psi_\mu(\mathbf{x}',\mathbf{y}_\mu^*(\mathbf{x}')) - \nabla_{\mathbf{xy}}^2\psi_\mu(\mathbf{x},\mathbf{y}_\mu^*(\mathbf{x}))\right]\mathbf{v}_\mu^*(\mathbf{x}).\end{aligned}$$

Thus, by Lipschitz continuity of $\nabla_{\mathbf{yy}}^2\psi_\mu$ and Lemma D.2, we have

$$\begin{aligned}&\left\|\left[\nabla_{\mathbf{xy}}^2\psi_\mu(\mathbf{x}',\mathbf{y}_\mu^*(\mathbf{x}'))\right]\mathbf{v}_\mu^*(\mathbf{x}') - \left[\nabla_{\mathbf{xy}}^2\psi_\mu(\mathbf{x},\mathbf{y}_\mu^*(\mathbf{x}))\right]\mathbf{v}_\mu^*(\mathbf{x})\right\|\\ \leq& L_{\psi_{\mathbf{y}1}}\left\|\mathbf{v}_\mu^*(\mathbf{x}') - \mathbf{v}_\mu^*(\mathbf{x})\right\| + \|\mathbf{v}_\mu^*(\mathbf{x})\|\left(L_{\psi_{\mathbf{xy}1}}\|\mathbf{x}'-\mathbf{x}\| + L_{\psi_{\mathbf{xy}2}}\|\mathbf{y}_\mu^*(\mathbf{x}) - \mathbf{y}_\mu^*(\mathbf{x}')\|\right)\\ \leq& \frac{L_{\psi_{\mathbf{y}1}}\left(L_{\mathbf{v}1}\|\mathbf{v}_\mu^*(\mathbf{x})\| + L_{\mathbf{v}2}\right)}{\sigma_{\psi_\mu}^2}\|\mathbf{x}-\mathbf{x}'\| + \|\mathbf{v}_\mu^*(\mathbf{x})\|\left(L_{\psi_{\mathbf{xy}1}}\|\mathbf{x}'-\mathbf{x}\| + L_{\psi_{\mathbf{xy}2}}\|\mathbf{y}_\mu^*(\mathbf{x}) - \mathbf{y}_\mu^*(\mathbf{x}')\|\right).\end{aligned}$$

Hence, by using Lemma D.2, we have

$$\begin{aligned}&\|\nabla\Phi_\mu(\mathbf{x}) - \nabla\Phi_\mu(\mathbf{x}')\|\\ \leq& \left(L_{F_{\mathbf{x}1}} + L_{\psi_{\mathbf{xy}1}}\|\mathbf{v}_\mu^*(\mathbf{x})\| + \frac{L_{\psi_{\mathbf{y}1}}\left(L_{\mathbf{v}1}\|\mathbf{v}_\mu^*(\mathbf{x})\| + L_{\mathbf{v}2}\right)}{\sigma_{\psi_\mu}^2}\right)\|\mathbf{x}-\mathbf{x}'\| + \left(L_{F_{\mathbf{x}2}} + L_{\psi_{\mathbf{xy}2}}\|\mathbf{v}_\mu^*(\mathbf{x})\|\right)\|\mathbf{y}_\mu^*(\mathbf{x}) - \mathbf{y}_\mu^*(\mathbf{x}')\|\\ \leq& \left(L_{F_{\mathbf{x}1}} + L_{\psi_{\mathbf{xy}1}}\|\mathbf{v}_\mu^*(\mathbf{x})\| + \frac{L_{\psi_{\mathbf{y}1}}\left(L_{\mathbf{v}1}\|\mathbf{v}_\mu^*(\mathbf{x})\| + L_{\mathbf{v}2}\right)}{\sigma_{\psi_\mu}^2}\right)\|\mathbf{x}-\mathbf{x}'\| + \left(L_{F_{\mathbf{x}2}} + L_{\psi_{\mathbf{xy}2}}\|\mathbf{v}_\mu^*(\mathbf{x})\|\right)\frac{L_{\psi_{\mathbf{y}1}}}{\sigma_{\psi_\mu}}\|\mathbf{x}-\mathbf{x}'\|,\end{aligned}$$

which implies the desired result since

$$L_{\psi_{\mathbf{y}1}} L_{\mathbf{v}1} + L_{\psi_{\mathbf{xy}2}} L_{\psi_{\mathbf{y}1}} \sigma_{\psi_\mu} + L_{\psi_{\mathbf{xy}1}} \sigma_{\psi_\mu}^2 = L_{\psi_{\mathbf{y}1}}^2 L_{\psi_{\mathbf{yy}2}} + L_{\psi_{\mathbf{y}1}} \left(L_{\psi_{\mathbf{yy}1}} + L_{\psi_{\mathbf{xy}2}} \right) \sigma_{\psi_\mu} + L_{\psi_{\mathbf{xy}1}} \sigma_{\psi_\mu}^2$$

and

$$L_{\psi_{\mathbf{y}1}} L_{\mathbf{v}2} + L_{F_{\mathbf{x}2}} L_{\psi_{\mathbf{y}1}} \sigma_{\psi_\mu} + L_{F_{\mathbf{x}1}} \sigma_{\psi_\mu}^2 = L_{\psi_{\mathbf{y}1}}^2 L_{F_{\mathbf{y}2}} + L_{\psi_{\mathbf{y}1}} \left(L_{F_{\mathbf{y}1}} + L_{F_{\mathbf{x}2}} \right) \sigma_{\psi_\mu} + L_{F_{\mathbf{x}1}} \sigma_{\psi_\mu}^2.$$

□

Next lemma characterizes the smoothness of $\mathbf{y}_\mu^*(\mathbf{x})$ and $\mathbf{v}_\mu^*(\mathbf{x})$ in μ. It will plays an important role in the non-asymptotic analysis for sl-BAMM when BLO has multiple LL minimizers.

Lemma D.4. *Suppose Assumptions 3.1, and either Assumption 3.3 or Assumption 3.7 holds, let $0 < \mu \leq 1/2$ and $\mu' \leq 2\mu$,*

(i)

$$\|\mathbf{y}_\mu^*(\mathbf{x}) - \mathbf{y}_{\mu'}^*(\mathbf{x})\| \leq \frac{2\|\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*(\mathbf{x}))\|}{\sigma_F} \cdot \frac{|\mu - \mu'|}{\mu'}. \tag{41}$$

(ii)

$$\|\mathbf{v}_\mu^*(\mathbf{x}) - \mathbf{v}_{\mu'}^*(\mathbf{x})\| \leq \left(C_{\mathbf{v}1} \|\mathbf{v}_\mu^*(\mathbf{x})\| + C_{\mathbf{v}2} \right) \|\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*(\mathbf{x}))\| \cdot \frac{|\mu - \mu'|}{\mu'^2}, \tag{42}$$

where both $C_{\mathbf{v}1}$ and $C_{\mathbf{v}2}$ are constants given by

$$C_{\mathbf{v}1} := 2 \left(L_{F_{\mathbf{yy}2}} + L_{f_{\mathbf{yy}2}} \right) / \sigma_F^2, \quad C_{\mathbf{v}2} := 2 \left(2L_{F_{\mathbf{y}2}} + L_{f_{\mathbf{y}2}} \right) / \sigma_F^2.$$

Proof. (i) Recall that $\mathbf{y}_\mu^* = \mathbf{y}_\mu^*(\mathbf{x})$ satisfies $\mu \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*) + (1-\mu) \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_\mu^*) = 0$. Thus

$$\begin{aligned} &(\mu - \mu') \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*) + \mu' \left[\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_{\mu'}^*) \right] \\ &+ (\mu' - \mu) \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_\mu^*) + (1 - \mu') \left[\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_{\mu'}^*) \right] = 0, \end{aligned}$$

which implies that

$$\begin{aligned} &\mu' \left[\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_{\mu'}^*) \right] + (1 - \mu') \left[\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_{\mu'}^*) \right] \\ =&(\mu' - \mu) \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*) + (\mu' - \mu) \frac{\mu}{1-\mu} \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*) = (\mu' - \mu) \frac{\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*)}{1-\mu}. \end{aligned}$$

Since $F(\mathbf{x}, \cdot)$ is σ_F-strongly convex, we have

$$\left\langle \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_{\mu'}^*),\ \mathbf{y}_\mu^* - \mathbf{y}_{\mu'}^* \right\rangle \geq \sigma_F \|\mathbf{y}_\mu^* - \mathbf{y}_{\mu'}^*\|^2.$$

Similarly, since $f(\mathbf{x}, \cdot)$ is convex, we get

$$\left\langle \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_{\mu'}^*),\ \mathbf{y}_\mu^* - \mathbf{y}_{\mu'}^* \right\rangle \geq 0.$$

Combining the above inequalities, we have

$$0 \leq \mu' \sigma_F \|\mathbf{y}_\mu^* - \mathbf{y}_{\mu'}^*\|^2 \leq \left(\frac{\mu' - \mu}{1-\mu} \right) \left\langle \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*), \mathbf{y}_\mu^* - \mathbf{y}_{\mu'}^* \right\rangle.$$

Since $0 < \mu \leq 1/2$, we get $1/(1-\mu) \leq 2$ and then

$$\mu' \sigma_F \|\mathbf{y}_\mu^* - \mathbf{y}_{\mu'}^*\|^2 \leq 2 \|\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*)\| \cdot |\mu - \mu'| \cdot \|\mathbf{y}_\mu^* - \mathbf{y}_{\mu'}^*\|,$$

which implies the desired result.

(ii) The definition of $\mathbf{v}_\mu^* = \mathbf{v}_\mu^*(\mathbf{x})$ says that $\left[\nabla_{\mathbf{yy}}^2 \psi_\mu(\mathbf{x}, \mathbf{y}_\mu^*)\right] \mathbf{v}_\mu^* = \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*)$. Thus we have

$$[\nabla_{\mathbf{yy}}^2 \psi_{\mu'}(\mathbf{x}, \mathbf{y}_{\mu'}^*)](\mathbf{v}_\mu^* - \mathbf{v}_{\mu'}^*) + \left[\nabla_{\mathbf{yy}}^2 \psi_\mu(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{yy}}^2 \psi_{\mu'}(\mathbf{x}, \mathbf{y}_{\mu'}^*)\right] \mathbf{v}_\mu^* = \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_{\mu'}^*).$$

Multiplying the above equation by $\mathbf{v}_\mu^* - \mathbf{v}_{\mu'}^*$, by the σ_{ψ_μ}-strong convexity of $\psi_\mu(\mathbf{x}, \cdot)$, we get

$$\begin{aligned}\sigma_{\psi_{\mu'}} \|\mathbf{v}_\mu^* - \mathbf{v}_{\mu'}^*\|^2 \leq & \|\nabla_{\mathbf{yy}}^2 \psi_\mu(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{yy}}^2 \psi_{\mu'}(\mathbf{x}, \mathbf{y}_{\mu'}^*)\| \cdot \|\mathbf{v}_\mu^*\| \cdot \|\mathbf{v}_\mu^* - \mathbf{v}_{\mu'}^*\| \\ & + \|\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_{\mu'}^*)\| \cdot \|\mathbf{v}_\mu^* - \mathbf{v}_{\mu'}^*\|.\end{aligned}$$

Note that by the definition of ψ_μ we have

$$\begin{aligned}\nabla_{\mathbf{yy}}^2 \psi_\mu(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{yy}}^2 \psi_{\mu'}(\mathbf{x}, \mathbf{y}_{\mu'}^*) = & (\mu - \mu') \nabla_{\mathbf{yy}}^2 F(\mathbf{x}, \mathbf{y}_\mu^*) + \mu' \left[\nabla_{\mathbf{yy}}^2 F(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{yy}}^2 F(\mathbf{x}, \mathbf{y}_{\mu'}^*)\right] \\ & + (\mu' - \mu) \nabla_{\mathbf{yy}}^2 f(\mathbf{x}, \mathbf{y}_\mu^*) + (1 - \mu') \left[\nabla_{\mathbf{yy}}^2 f(\mathbf{x}, \mathbf{y}_\mu^*) - \nabla_{\mathbf{yy}}^2 f(\mathbf{x}, \mathbf{y}_{\mu'}^*)\right].\end{aligned}$$

Since $\mu' \leq 2\mu \leq 1$, we get

$$\begin{aligned}\sigma_{\psi_{\mu'}} \|\mathbf{v}_\mu^* - \mathbf{v}_{\mu'}^*\| \leq & \left[(L_{F_{\mathbf{y}2}} + L_{f_{\mathbf{y}2}})|\mu - \mu'| + (L_{F_{\mathbf{yy}2}} + L_{f_{\mathbf{yy}2}})\|\mathbf{y}_\mu^* - \mathbf{y}_{\mu'}^*\|\right] \|\mathbf{v}_\mu^*\| + L_{F_{\mathbf{y}2}} \|\mathbf{y}_\mu^* - \mathbf{y}_{\mu'}^*\| \\ \leq & \left[\left(L_{F_{\mathbf{yy}2}} + L_{f_{\mathbf{yy}2}}\right) \|\mathbf{v}_\mu^*\| + L_{F_{\mathbf{y}2}}\right] \left\|\mathbf{y}_\mu^* - \mathbf{y}_{\mu'}^*\right\| + (L_{F_{\mathbf{y}2}} + L_{f_{\mathbf{y}2}}) \|\mathbf{v}_\mu^*\| \cdot |\mu - \mu'|.\end{aligned}$$

By (i), since $\sigma_{\psi_{\mu'}} = \mu' \sigma_F$ when $\sigma_f = 0$, we have

$$\begin{aligned}\|\mathbf{v}_\mu^* - \mathbf{v}_{\mu'}^*\| \leq & \left[\left(L_{F_{\mathbf{yy}2}} + L_{f_{\mathbf{yy}2}}\right) \|\mathbf{v}_\mu^*\| + L_{F_{\mathbf{y}2}}\right] \frac{2\|\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*(\mathbf{x}))\|}{\sigma_F^2} \cdot \frac{|\mu - \mu'|}{\mu'^2} \\ & + \frac{(L_{F_{\mathbf{y}2}} + L_{f_{\mathbf{y}2}})}{\sigma_F} \|\mathbf{v}_\mu^*\| \cdot \frac{|\mu - \mu'|}{\mu'}.\end{aligned}$$

Since $\mu' \leq 2\mu$, Lemma D.2 implies that

$$\mu' \|\mathbf{v}_\mu^*(\mathbf{x})\| \leq \mu' \frac{\|\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*(\mathbf{x}))\|}{\mu \sigma_F} \leq \frac{2\|\nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_\mu^*(\mathbf{x}))\|}{\sigma_F}.$$

Hence we get the desired result.

□

Note that some of the coefficients of terms in Inequality (23) depends on $\|\mathbf{v}_{\mu_k}^*(\mathbf{x}_k)\|$ and $\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}_{\mu_{k+1}}^*(\mathbf{x}_{k+1}))\|$. We next upper-bound such two types of terms.

Lemma D.5. *Assume that $\nabla_{\mathbf{y}} F$ and $\nabla_{\mathbf{y}} f$ are Lipschitz continuous in $(\mathbf{x}, \mathbf{y})$. If $(\mathbf{x}_k, \mathbf{y}_k)$ is bounded, that is, $\|\mathbf{x}_k\| + \|\mathbf{y}_k\| \leq D$. Then*

$$\|\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}_{\mu_k}^*(\mathbf{x}_k))\| = O(1/\sigma_{\psi_{\mu_k}}), \quad \|\mathbf{v}_{\mu_k}^*(\mathbf{x}_k)\| = O\left(1/\sigma_{\psi_{\mu_k}}^2\right). \tag{43}$$

In detail,

$$\|\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}_{\mu_k}^*(\mathbf{x}_k))\| \leq (C_{d1} D + C_{d2}) / \sigma_{\psi_{\mu_k}}, \quad \|\mathbf{v}_{\mu_k}^*(\mathbf{x}_k)\| \leq (C_{d1} D + C_{d2}) / \sigma_{\psi_{\mu_k}}^2, \tag{44}$$

where

$$\begin{aligned}C_{d1} := & L_{F_{\mathbf{y}2}} \left(L_{f_{\mathbf{y}1}} + L_{f_{\mathbf{y}2}}\right) + \left(\sigma_{\psi_{\mu_k}} + \mu_k L_{F_{\mathbf{y}2}}\right) \left(L_{F_{\mathbf{y}1}} + L_{F_{\mathbf{y}2}}\right), \\ C_{d2} := & L_{F_{\mathbf{y}2}} \|\nabla_{\mathbf{y}} f(0, 0)\| + \left(\sigma_{\psi_{\mu_k}} + \mu_k L_{F_{\mathbf{y}2}}\right) \|\nabla_{\mathbf{y}} F(0, 0)\|,\end{aligned}$$

are constants independent of D.

Proof. First, since $\|\mathbf{x}_k\| + \|\mathbf{y}_k\| \leq D$, by Lipschitz continuity of $\nabla_{\mathbf{y}} F$ and $\nabla_{\mathbf{y}} f$, we have

$$\begin{aligned}\|\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}_k)\| \leq & \|\nabla_{\mathbf{y}} F(0, 0)\| + \|\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{y}} F(0, 0)\| \\ \leq & \|\nabla_{\mathbf{y}} F(0, 0)\| + L_{F_{\mathbf{y}1}} \|\mathbf{x}_k\| + L_{F_{\mathbf{y}2}} \|\mathbf{y}_k\| \\ \leq & \|\nabla_{\mathbf{y}} F(0, 0)\| + \left(L_{F_{\mathbf{y}1}} + L_{F_{\mathbf{y}2}}\right) D.\end{aligned}$$

Similarly, we also have $\|\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k)\| \leq \|\nabla_{\mathbf{y}} f(0,0)\| + \left(L_{f_{\mathbf{y}1}} + L_{f_{\mathbf{y}2}}\right) D$.

Next, since $\nabla_{\mathbf{y}} \psi_{\mu_k}(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)) = 0$, by the $\sigma_{\psi_{\mu_k}}$-strong convexity of $\psi_{\mu_k}(\mathbf{x}, \cdot)$, we get

$$\sigma_{\psi_{\mu_k}} \left\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\| \leq \|\nabla_{\mathbf{y}} \psi_{\mu_k}(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{y}} \psi_{\mu_k}(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k))\| = \|\nabla_{\mathbf{y}} \psi_{\mu_k}(\mathbf{x}_k, \mathbf{y}_k)\|.$$

This implies that $\left\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\| \leq \|\nabla_{\mathbf{y}} \psi_{\mu_k}(\mathbf{x}_k, \mathbf{y}_k)\| / \sigma_{\psi_{\mu_k}}$ and then

$$\begin{aligned}
\left\|\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k))\right\| \leq & \|\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}_k)\| + \|\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)) - \nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}_k)\| \\
\leq & \|\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}_k)\| + L_{F_{\mathbf{y}2}} \left\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\| \\
\leq & \|\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}_k)\| + \frac{L_{F_{\mathbf{y}2}} \|\nabla_{\mathbf{y}} \psi_{\mu_k}(\mathbf{x}_k, \mathbf{y}_k)\|}{\sigma_{\psi_{\mu_k}}} \\
\leq & \left(1 + \frac{\mu_k L_{F_{\mathbf{y}2}}}{\sigma_{\psi_{\mu_k}}}\right) \|\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}_k)\| + \frac{L_{F_{\mathbf{y}2}}}{\sigma_{\psi_{\mu_k}}} \|\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k)\|,
\end{aligned}$$

since $\nabla_{\mathbf{y}} \psi_{\mu_k} = \mu_k \nabla_{\mathbf{y}} F + (1 - \mu_k) \nabla_{\mathbf{y}} f$. Thus, by Lemma D.2, we have

$$\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| \leq \left(\frac{1}{\sigma_{\psi_{\mu_k}}} + \frac{\mu_k L_{F_{\mathbf{y}2}}}{\sigma^2_{\psi_{\mu_k}}}\right) \|\nabla_{\mathbf{y}} F(\mathbf{x}_k, \mathbf{y}_k)\| + \frac{L_{F_{\mathbf{y}2}}}{\sigma^2_{\psi_{\mu_k}}} \|\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k)\| \leq \frac{(C_{d1} D + C_{d2})}{\sigma^2_{\psi_{\mu_k}}}.$$

□

D.2. Proof of Lemma C.1

By the above lemmas, we can analyze the descent of the approximate overall UL objective.

Lemma D.6 (**Lemma C.1**). *Suppose Assumptions 3.1, and either Assumption 3.3 or Assumption 3.7 holds. Let $\mu_{k+1} \leq \mu_k \leq \frac{1}{2}$ for all k. Then the sequence of $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ generated by sl-BAMM satisfies*

$$\begin{aligned}
& \Phi_{\mu_{k+1}}(\mathbf{x}_{k+1}) - \Phi_{\mu_k}(\mathbf{x}_k) \\
\leq & -\frac{\alpha_k}{2} \|\nabla \Phi_{\mu_k}(\mathbf{x}_k)\|^2 - \frac{1}{2}\left(\frac{1}{\alpha_k} - \frac{C_{\Phi 1} \|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + C_{\Phi 2}}{\sigma^2_{\psi_{\mu_k}}}\right) \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \alpha_k L^2_{\psi_{\mathbf{y}1}} \|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\
& + \alpha_k \left(L_{\psi_{\mathbf{xy}2}} \left\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\| + L_{F_{\mathbf{x}2}}\right)^2 \|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{2\left\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2}{\sigma_F} \left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right),
\end{aligned} \tag{45}$$

where both $C_{\Phi 1}$ and $C_{\Phi 2}$ are constants given by

$$\begin{aligned}
C_{\Phi 1} := & L^2_{\psi_{\mathbf{y}1}} L_{\psi_{\mathbf{yy}2}} + L_{\psi_{\mathbf{y}1}} \left(L_{\psi_{\mathbf{yy}1}} + L_{\psi_{\mathbf{xy}2}}\right) \sigma_{\psi_\mu} + L_{\psi_{\mathbf{xy}1}} \sigma^2_{\psi_\mu}, \\
C_{\Phi 2} := & L^2_{\psi_{\mathbf{y}1}} L_{F_{\mathbf{y}2}} + L_{\psi_{\mathbf{y}1}} \left(L_{F_{\mathbf{y}1}} + L_{F_{\mathbf{x}2}}\right) \sigma_{\psi_\mu} + L_{F_{\mathbf{x}1}} \sigma^2_{\psi_\mu}.
\end{aligned}$$

In particular, if LL objective $f(\mathbf{x}, \cdot)$ is σ_f-strongly convex, i.e., Assumption 3.7 holds, we take $\mu_k = 0$ for all k and then the last term in Estimate (20) is redundant.

Proof. First, note that

$$\begin{aligned}
F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})) - F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)) = & F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})) - F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_k}(\mathbf{x}_{k+1})) \\
& + F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_k}(\mathbf{x}_{k+1})) - F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)).
\end{aligned}$$

By the convexity of $F(\mathbf{x}, \cdot)$ and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})) - F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_k}(\mathbf{x}_{k+1})) \leq & \left\langle \nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})), \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}) - \mathbf{y}^*_{\mu_k}(\mathbf{x}_{k+1}) \right\rangle \\
\leq & \left\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\| \cdot \|\mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}) - \mathbf{y}^*_{\mu_k}(\mathbf{x}_{k+1})\| \\
\leq & \frac{2\left\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2}{\sigma_F} \left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right),
\end{aligned}$$

where the last inequality follows from (41) with $\mu' = \mu_k$ and $\mu = \mu_{k+1}$.

Next, denote $L_{\Phi\mu_k} := C_{\Phi 1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + C_{\Phi 2}$. Since $L_{\Phi\mu}$ is independent of $\mathbf{x}'$ in Lemma D.3, similar to the proof of the classical descent lemma (cf. Lemma 5.7 in (Beck, 2017)), one can get

$$\begin{aligned}
&F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_k}(\mathbf{x}_{k+1})) - F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)) = \Phi_{\mu_k}(\mathbf{x}_{k+1}) - \Phi_{\mu_k}(\mathbf{x}_k)\\
\le&\left\langle \nabla\Phi_{\mu_k}(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\right\rangle + \frac{L_{\Phi\mu_k}}{2\sigma^2_{\psi_{\mu_k}}}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2\\
\le& -\alpha_k\left\langle \nabla\Phi_{\mu_k}(\mathbf{x}_k), \nabla_{\mathbf{x}}F(\mathbf{x}_k,\mathbf{y}_k) - \nabla^2_{\mathbf{xy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)\mathbf{v}_k\right\rangle + \frac{L_{\Phi\mu_k}}{2\sigma^2_{\psi_{\mu_k}}}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2\\
\le& -\frac{\alpha_k}{2}\|\nabla\Phi_{\mu_k}(\mathbf{x}_k)\|^2 - \frac{\alpha_k}{2}\left\|\nabla_{\mathbf{x}}F(\mathbf{x}_k,\mathbf{y}_k) - \nabla^2_{\mathbf{xy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)\mathbf{v}_k\right\|^2\\
&+\frac{\alpha_k}{2}\left\|\nabla\Phi_{\mu_k}(\mathbf{x}_k) - \nabla_{\mathbf{x}}F(\mathbf{x}_k,\mathbf{y}_k) + \nabla^2_{\mathbf{xy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)\mathbf{v}_k\right\|^2 + \frac{L_{\Phi\mu_k}}{2\sigma^2_{\psi_{\mu_k}}}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2,
\end{aligned}$$

where the second inequality used the definition of $\mathbf{x}_{k+1}$ and the third one follows from $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$. Note that the update of $\mathbf{x}_{k+1}$ implies that

$$\left\|\nabla_{\mathbf{x}}F(\mathbf{x}_k,\mathbf{y}_k) - \nabla^2_{\mathbf{xy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)\mathbf{v}_k\right\|^2 = \frac{1}{\alpha_k^2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2,$$

and by the definition of $\Phi_{\mu_k}(\mathbf{x})$ we get

$$\nabla\Phi_{\mu_k}(\mathbf{x}_k) = \nabla_{\mathbf{x}}F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)) - \nabla^2_{\mathbf{xy}}\psi_{\mu_k}(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k))\mathbf{v}^*_{\mu_k}(\mathbf{x}_k),$$

which implies that

$$\begin{aligned}
&\left\|\nabla\Phi_{\mu_k}(\mathbf{x}_k) - \nabla_{\mathbf{x}}F(\mathbf{x}_k,\mathbf{y}_k) + \nabla^2_{\mathbf{xy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)\mathbf{v}_k\right\|\\
\le&\|\nabla_{\mathbf{x}}F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)) - \nabla_{\mathbf{x}}F(\mathbf{x}_k,\mathbf{y}_k)\| + \left\|\left[\nabla^2_{\mathbf{xy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k) - \nabla^2_{\mathbf{xy}}\psi_{\mu_k}(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k))\right]\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\|\\
&+\left\|\nabla^2_{\mathbf{xy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)[\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)]\right\|\\
\le&\left(L_{F_{\mathbf{x}2}} + L_{\psi_{\mathbf{xy}2}}\left\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\|\right)\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\| + L_{\psi_{\mathbf{y}1}}\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|.
\end{aligned}$$

The desired result follows from the inequality $\left(\sum_{i=1}^r a_i\right)^2 \le r\sum_{i=1}^r a_i^2$. □

D.3. Proof of Lemma C.2

Lemma D.6 implies that the descent of the approximate overall UL objective depends on the errors of LL and multiplier variables. We next analyze the errors of LL and multiplier variables.

Lemma D.7. *Suppose Assumptions 3.1, and either Assumption 3.3 or Assumption 3.7 holds. If we choose*

$$\beta_k \le \frac{2}{\sigma_{\psi_{\mu_k}} + L_{\psi_{\mu_k}}}, \quad \eta_k \le \frac{1}{L_{\psi_{\mu_k}}}. \tag{46}$$

Then the sequence of $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ generated by sl-BAMM satisfies

$$\left\|\mathbf{y}_{k+1} - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\|^2 \le \left(1 - \beta_k\sigma_{\psi_{\mu_k}}\right)\left\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\|^2, \tag{47}$$

and

$$\left\|\mathbf{v}_{k+1} - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\|^2 \le \left(1 - \eta_k\sigma_{\psi_{\mu_k}}\right)\left\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\|^2 + \frac{2\eta_k}{\sigma_{\psi_{\mu_k}}}\left(L_{\psi_{\mathbf{yy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{y}2}}\right)^2\left\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\|^2.$$

Proof. First, by the update of $\mathbf{y}_{k+1}$, we have

$$\begin{aligned}
\left\|\mathbf{y}_{k+1} - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\|^2 &= \left\|\mathbf{y}_k - \beta_k\nabla_{\mathbf{y}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k) - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\|^2\\
&= \left\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\|^2 - 2\beta_k\left\langle \mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k), \nabla_{\mathbf{y}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)\right\rangle + \beta_k^2\left\|\nabla_{\mathbf{y}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)\right\|^2.
\end{aligned}$$

By the $\sigma_{\psi_{\mu_k}}$-strong convexity and $L_{\psi_{\mu_k}}$-smoothness of $\psi_{\mu_k}(\mathbf{x},\cdot)$, since $\nabla_{\mathbf{y}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}^*_{\mu_k}(\mathbf{x}_k))=0$, Theorem 2.1.12 in (Nesterov, 2018) implies that

$$\begin{aligned}\left\langle \mathbf{y}_k-\mathbf{y}^*_{\mu_k}(\mathbf{x}_k),\nabla_{\mathbf{y}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)\right\rangle =&\left\langle \mathbf{y}_k-\mathbf{y}^*_{\mu_k}(\mathbf{x}_k),\nabla_{\mathbf{y}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)-\nabla_{\mathbf{y}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}^*_{\mu_k}(\mathbf{x}_k))\right\rangle\\ \geq&\frac{\sigma_{\psi_{\mu_k}}L_{\psi_{\mu_k}}}{\sigma_{\psi_{\mu_k}}+L_{\psi_{\mu_k}}}\left\|\mathbf{y}_k-\mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\|^2+\frac{1}{\sigma_{\psi_{\mu_k}}+L_{\psi_{\mu_k}}}\left\|\nabla_{\mathbf{y}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)\right\|^2.\end{aligned}$$

Hence, when $0<\beta_k\leq\frac{2}{\sigma_{\psi_{\mu_k}}+L_{\psi_{\mu_k}}}$, we have

$$\left\|\mathbf{y}_{k+1}-\mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\|^2\leq\left(1-\frac{2\sigma_{\psi_{\mu_k}}L_{\psi_{\mu_k}}}{\sigma_{\psi_{\mu_k}}+L_{\psi_{\mu_k}}}\beta_k\right)\left\|\mathbf{y}_k-\mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\|^2,$$

which implies the desired result since $\frac{1}{2}\leq\frac{L_{\psi_{\mu_k}}}{\sigma_{\psi_{\mu_k}}+L_{\psi_{\mu_k}}}\leq 1$.

Second, by the update of $\mathbf{v}_{k+1}$, using $\left[\nabla^2_{\mathbf{yy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}^*_{\mu_k}(\mathbf{x}_k))\right]\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)=\nabla_{\mathbf{y}}F(\mathbf{x}_k,\mathbf{y}^*_{\mu_k}(\mathbf{x}_k))$,

$$\begin{aligned}\mathbf{v}_{k+1}-\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)=&\mathbf{v}_k-\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)-\eta_k\left([\nabla^2_{\mathbf{yy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)]\mathbf{v}_k-\nabla_{\mathbf{y}}F(\mathbf{x}_k,\mathbf{y}_k)\right)\\ =&\mathbf{v}_k-\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)-\eta_k\left[\nabla^2_{\mathbf{yy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)\right]\left[\mathbf{v}_k-\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right]\\ &-\eta_k\left[\nabla^2_{\mathbf{yy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)-\nabla^2_{\mathbf{yy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}^*_{\mu_k}(\mathbf{x}_k))\right]\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\\ &-\eta_k\left[\nabla_{\mathbf{y}}F(\mathbf{x}_k,\mathbf{y}^*_{\mu_k}(\mathbf{x}_k))-\nabla_{\mathbf{y}}F(\mathbf{x}_k,\mathbf{y}_k)\right].\end{aligned}$$

Hence, by Cauchy-Schwarz inequality, for all $\varepsilon>0$, we have

$$\begin{aligned}&\left\|\mathbf{v}_{k+1}-\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\|^2\\ \leq&(1+\varepsilon)\left\|\left[I-\eta_k\nabla^2_{\mathbf{yy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)\right]\left[\mathbf{v}_k-\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right]\right\|^2\\ &+\left(1+\frac{1}{\varepsilon}\right)\eta_k^2\left\|\left[\nabla^2_{\mathbf{yy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)-\nabla^2_{\mathbf{yy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}^*_{\mu_k}(\mathbf{x}_k))\right]\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)+\nabla_{\mathbf{y}}F(\mathbf{x}_k,\mathbf{y}^*_{\mu_k}(\mathbf{x}_k))-\nabla_{\mathbf{y}}F(\mathbf{x}_k,\mathbf{y}_k)\right\|^2.\end{aligned}$$

When $\eta_k\leq 1/L_{\psi_{\mu_k}}$, by the $\sigma_{\psi_{\mu_k}}$-strong convexity of $\psi_{\mu_k}(\mathbf{x},\cdot)$, since the spectral norm is monotone, we have

$$\left\|\left[I-\eta_k\nabla^2_{\mathbf{yy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_{k+1})\right]\left[\mathbf{v}_k-\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right]\right\|\leq\left(1-\eta_k\sigma_{\psi_{\mu_k}}\right)\left\|\mathbf{v}_k-\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\|.$$

Next, by Lipschitz continuity of $\nabla^2_{\mathbf{yy}}\psi_{\mu_k}(\mathbf{x},\cdot)$ and $\nabla_{\mathbf{y}}F(\mathbf{x},\cdot)$, we get

$$\begin{aligned}&\left\|\left[\nabla^2_{\mathbf{yy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}_k)-\nabla^2_{\mathbf{yy}}\psi_{\mu_k}(\mathbf{x}_k,\mathbf{y}^*_{\mu_k}(\mathbf{x}_k))\right]\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)+\nabla_{\mathbf{y}}F(\mathbf{x}_k,\mathbf{y}^*_{\mu_k}(\mathbf{x}_k))-\nabla_{\mathbf{y}}F(\mathbf{x}_k,\mathbf{y}_k)\right\|\\ \leq&\left(L_{\psi_{\mathbf{yy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|+L_{F_{\mathbf{y}2}}\right)\left\|\mathbf{y}_k-\mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\|.\end{aligned}$$

Taking $\varepsilon=\eta_k\sigma_{\psi_{\mu_k}}$, we have

$$\begin{aligned}\left\|\mathbf{v}_{k+1}-\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\|^2\leq&\left(1+\eta_k\sigma_{\psi_{\mu_k}}\right)\left(1-\eta_k\sigma_{\psi_{\mu_k}}\right)^2\left\|\mathbf{v}_k-\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\|^2\\ &+\left(1+\frac{1}{\eta_k\sigma_{\psi_{\mu_k}}}\right)\eta_k^2\left(L_{\psi_{\mathbf{yy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|+L_{F_{\mathbf{y}2}}\right)^2\left\|\mathbf{y}_k-\mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\|^2,\end{aligned}$$

which implies the desired result since $\eta_k^2\leq\eta_k/L_{\psi_{\mu_k}}\leq\eta_k/\sigma_{\psi_{\mu_k}}$. □

Lemma D.8. *Suppose Assumptions in Lemma D.4 hold. Then the sequence of $\mathbf{x}_k,\mathbf{y}_k,\mathbf{v}_k$ generated by sl-BAMM satisfies*

$$\begin{aligned}\|\mathbf{y}^*_{\mu_k}(\mathbf{x}_k)-\mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2\leq&\frac{2L^2_{\psi_{\mathbf{y}1}}}{\sigma^2_{\psi_{\mu_k}}}\|\mathbf{x}_k-\mathbf{x}_{k+1}\|^2+\frac{8\|\nabla_{\mathbf{y}}F(\mathbf{x}_{k+1},\mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2}{\sigma_F^2}\left(\frac{\mu_k-\mu_{k+1}}{\mu_k}\right)^2,\\ \left\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)-\mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\right\|^2\leq&\frac{2\left(L_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|+L_{\mathbf{v}2}\right)^2}{\sigma^4_{\psi_{\mu_k}}}\|\mathbf{x}_k-\mathbf{x}_{k+1}\|^2\\ &+2\left(C_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|+C_{\mathbf{v}2}\right)^2\|\nabla_{\mathbf{y}}F(\mathbf{x}_{k+1},\mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2\left(\frac{\mu_k-\mu_{k+1}}{\mu_k^2}\right)^2.\end{aligned}$$

Proof. By the triangle inequality, and Lemma D.2(i) and D.4(i) with $\mu' = \mu_k$ and $\mu = \mu_{k+1}$,

$$\begin{aligned}\|\mathbf{y}^*_{\mu_k}(\mathbf{x}_k) - \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| \leq & \|\mathbf{y}^*_{\mu_k}(\mathbf{x}_k) - \mathbf{y}^*_{\mu_k}(\mathbf{x}_{k+1})\| + \|\mathbf{y}^*_{\mu_k}(\mathbf{x}_{k+1}) - \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| \\ \leq & \frac{L_{\psi_{\mathbf{y}1}}}{\sigma_{\psi_{\mu_k}}}\|\mathbf{x}_k - \mathbf{x}_{k+1}\| + \frac{2\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|}{\sigma_F}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right),\end{aligned}$$

and by Lemma D.2(iii) and D.4(ii) with $\mu' = \mu_k$ and $\mu = \mu_{k+1}$,

$$\begin{aligned}&\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k) - \mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| \leq \|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k) - \mathbf{v}^*_{\mu_k}(\mathbf{x}_{k+1})\| + \|\mathbf{v}^*_{\mu_k}(\mathbf{x}_{k+1}) - \mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| \\ \leq & \frac{\left(L_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{\mathbf{v}2}\right)}{\sigma^2_{\psi_{\mu_k}}}\|\mathbf{x}_k - \mathbf{x}_{k+1}\| + \left(C_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| + C_{\mathbf{v}2}\right)\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|\left(\frac{\mu_k - \mu_{k+1}}{\mu_k^2}\right).\end{aligned}$$

Then the desired result follows from the inequality $\left(\sum_{i=1}^2 a_i\right)^2 \leq 2\sum_{i=1}^2 a_i^2$. □

By the above lemmas, We can next analyze the error of LL and multiplier variables.

Lemma D.9. *Suppose Assumptions in Lemma C.1 hold. If we choose*

$$\beta_k \leq \frac{2}{\sigma_{\psi_{\mu_k}} + L_{\psi_{\mu_k}}}, \quad \eta_k \leq \frac{1}{L_{\psi_{\mu_k}}}.$$

Then the sequence of $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ generated by sl-BAMM satisfies

$$\begin{aligned}\|\mathbf{y}_{k+1} - \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2 - \|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 \leq & -\frac{1}{2}\beta_k\sigma_{\psi_{\mu_k}}\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{6L^2_{\psi_{\mathbf{y}1}}}{\beta_k\sigma^3_{\psi_{\mu_k}}}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \\ & + \frac{24\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2}{\sigma_F^2\beta_k\sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right)^2,\end{aligned} \tag{48}$$

and

$$\begin{aligned}&\|\mathbf{v}_{k+1} - \mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2 - \|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\ \leq & -\frac{1}{2}\eta_k\sigma_{\psi_{\mu_k}}\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{6\left(L_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{\mathbf{v}2}\right)^2}{\eta_k\sigma^5_{\psi_{\mu_k}}}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \\ & + \frac{3\eta_k}{\sigma_{\psi_{\mu_k}}}\left(L_{\psi_{\mathbf{y}\mathbf{y}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{y}2}}\right)^2\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\ & + \frac{6\left(C_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| + C_{\mathbf{v}2}\right)^2\|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2}{\eta_k\sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k^2}\right)^2,\end{aligned} \tag{49}$$

where both $C_{\mathbf{v}1}$ and $C_{\mathbf{v}2}$ are constants given by

$$C_{\mathbf{v}1} := 2\left(L_{F_{\mathbf{y}\mathbf{y}2}} + L_{f_{\mathbf{y}\mathbf{y}2}}\right)/\sigma_F^2, \quad C_{\mathbf{v}2} := 2\left(2L_{F_{\mathbf{y}2}} + L_{f_{\mathbf{y}2}}\right)/\sigma_F^2.$$

In particular, if LL objective $f(\mathbf{x}, \cdot)$ is σ_f-strongly convex, we take $\mu_k = 0$ for all k and then the last terms in Estimates (48)-(49) are both redundant.

Proof. By Cauchy-Schwarz inequality, it is easy to check that for any $\varepsilon > 0$ we have

$$\begin{aligned}\|\mathbf{y}_{k+1} - \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2 = & \|\mathbf{y}_{k+1} - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k) + \mathbf{y}^*_{\mu_k}(\mathbf{x}_k) - \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2 \\ \leq & (1+\varepsilon)\|\mathbf{y}_{k+1} - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \left(1 + \frac{1}{\varepsilon}\right)\|\mathbf{y}^*_{\mu_k}(\mathbf{x}_k) - \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2.\end{aligned}$$

Taking $\varepsilon = \frac{1}{2}\beta_k\sigma_{\psi_{\mu_k}}$, by Lemmas D.7 and D.8, we have

$$\|\mathbf{y}_{k+1} - \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2 \le \left(1 - \frac{1}{2}\beta_k\sigma_{\psi_{\mu_k}}\right)\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{2L^2_{\psi_{\mathbf{y}1}}}{\sigma^2_{\psi_{\mu_k}}}\left(1 + \frac{2}{\beta_k\sigma_{\psi_{\mu_k}}}\right)\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 + \frac{8\|\nabla_{\mathbf{y}}F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2}{\sigma^2_F}\left(1 + \frac{2}{\beta_k\sigma_{\psi_{\mu_k}}}\right)\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right)^2,$$

which implies the desired result since $\beta_k\sigma_{\psi_{\mu_k}} \le 1$ when $\beta_k \le 2/(\sigma_{\psi_{\mu_k}} + L_{\psi_{\mu_k}})$.

Similarly, for any $\delta > 0$, by Cauchy-Schwarz inequality,

$$\begin{aligned}\left\|\mathbf{v}_{k+1} - \mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\right\|^2 =& \left\|\mathbf{v}_{k+1} - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k) + \mathbf{v}^*_{\mu_k}(\mathbf{x}_k) - \mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\right\|^2 \\ \le& (1+\delta)\left\|\mathbf{v}_{k+1} - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\|^2 + \left(1 + \frac{1}{\delta}\right)\left\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k) - \mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\right\|^2.\end{aligned}$$

Taking $\delta = \frac{1}{2}\eta_k\sigma_{\psi_{\mu_k}}$, Lemmas D.7 and D.8 imply that

$$\begin{aligned}&\left\|\mathbf{v}_{k+1} - \mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\right\|^2 \\ \le& \left(1 - \frac{1}{2}\eta_k\sigma_{\psi_{\mu_k}}\right)\left\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\right\|^2 + \frac{2\eta_k}{\sigma_{\psi_{\mu_k}}}\left(1 + \frac{1}{2}\eta_k\sigma_{\psi_{\mu_k}}\right)\left(L_{\psi_{\mathbf{yy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{y}2}}\right)^2\left\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\right\|^2 \\ &+ \frac{2\left(L_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{\mathbf{v}2}\right)^2}{\sigma^4_{\psi_{\mu_k}}}\left(1 + \frac{2}{\eta_k\sigma_{\psi_{\mu_k}}}\right)\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \\ &+ 2\left(C_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| + C_{\mathbf{v}2}\right)^2\|\nabla_{\mathbf{y}}F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2\left(1 + \frac{2}{\eta_k\sigma_{\psi_{\mu_k}}}\right)\left(\frac{\mu_k - \mu_{k+1}}{\mu_k^2}\right)^2.\end{aligned}$$

The desired result follows since $\eta_k\sigma_{\psi_{\mu_k}} \le 1$ when $\eta_k \le 1/L_{\psi_{\mu_k}}$. □

D.4. Proof of Lemma C.3

Lemma D.10. *Suppose Assumptions in Lemma C.2 hold. Let $\{a_k, b_k, c_k\}_{k=1}^{\infty}$ be a sequence of decreasing positive constants, then the following descent of Lyapunov function holds:*

$$\begin{aligned}V_{k+1} - V_k \le& -\frac{a_{k+1}\alpha_k}{2}\|\nabla\Phi_{\mu_k}(\mathbf{x}_k)\|^2 - \frac{\widehat{\alpha}_k}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 - \frac{\widehat{\beta}_k}{2}\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 - \frac{\widehat{\eta}_k}{2}\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\ &+ \frac{2\left\|\nabla_{\mathbf{y}}F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2 a_{k+1}}{\sigma_F}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right) \\ &+ \frac{24\left\|\nabla_{\mathbf{y}}F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\right\|^2 b_{k+1}}{\sigma^2_F\beta_k\sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right)^2 \\ &+ \frac{6\left(C_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| + C_{\mathbf{v}2}\right)^2\|\nabla_{\mathbf{y}}F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2 c_{k+1}}{\eta_k\sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k^2}\right)^2,\end{aligned} \tag{50}$$

where the coefficients $\{\widehat{\alpha}_k, \widehat{\beta}_k, \widehat{\eta}_k\}$ are given as below.

$$\begin{aligned}\widehat{\alpha}_k :=& \frac{a_{k+1}}{\alpha_k} - \frac{a_{k+1}\left(C_{\Phi 1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + C_{\Phi 2}\right)}{\sigma^2_{\psi_{\mu_k}}} - \frac{12L^2_{\psi_{\mathbf{y}1}}b_{k+1}}{\beta_k\sigma^3_{\psi_{\mu_k}}} - \frac{12\left(L_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{\mathbf{v}2}\right)^2 c_{k+1}}{\eta_k\sigma^5_{\psi_{\mu_k}}}, \\ \widehat{\beta}_k :=& b_{k+1}\beta_k\sigma_{\psi_{\mu_k}} - 2a_{k+1}\alpha_k\left(L_{\psi_{\mathbf{xy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{x}2}}\right)^2 - 6\left(L_{\psi_{\mathbf{yy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{y}2}}\right)^2\frac{c_{k+1}\eta_k}{\sigma_{\psi_{\mu_k}}}, \\ \widehat{\eta}_k :=& c_{k+1}\eta_k\sigma_{\psi_{\mu_k}} - 2a_{k+1}\alpha_k L^2_{\psi_{\mathbf{y}1}}.\end{aligned} \tag{51}$$

Here $L_{\mathbf{v}1} := L_{\psi_{\mathbf{yy}2}} L_{\psi_{\mathbf{y}1}} + L_{\psi_{\mathbf{yy}1}} \sigma_{\psi_\mu}$ *and* $L_{\mathbf{v}2} := L_{F_{\mathbf{y}2}} L_{\psi_{\mathbf{y}1}} + L_{F_{\mathbf{y}1}} \sigma_{\psi_\mu}$. *In particular, if LL objective* $f(\mathbf{x}, \cdot)$ *is* σ_f*-strongly convex, we take* $\mu_k = 0$ *for all* k *and then the terms involving* $\mu_k - \mu_{k+1}$ *in Estimate* (50) *are all redundant.*

Proof. Denote $L_{\Phi\mu_k} := C_{\Phi 1} \|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + C_{\Phi 2}$, $\nabla_{\mathbf{y}} F_{k+1} := \nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))$, $B_{\mathbf{vx}} := 6\left(L_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{\mathbf{v}2}\right)^2$, $B_{\mathbf{vy}} := 3\left(L_{\psi_{\mathbf{yy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{y}2}}\right)^2$, and $B_{\mathbf{v}\mu_{k+1}} := 6\left(C_{\mathbf{v}1}\|\mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\| + C_{\mathbf{v}2}\right)^2 \|\nabla_{\mathbf{y}} F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1}))\|^2$. Then, by Lemmas D.6-D.9, we have

$$
\begin{aligned}
V_{k+1} - V_k =& a_{k+1}\left[F(\mathbf{x}_{k+1}, \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})) - F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k))\right] + (a_{k+1} - a_k) F(\mathbf{x}_k, \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)) \\
&+ b_{k+1}\left[\|\mathbf{y}_{k+1} - \mathbf{y}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2 - \|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2\right] + (b_{k+1} - b_k)\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\
&+ c_{k+1}\left[\|\mathbf{v}_{k+1} - \mathbf{v}^*_{\mu_{k+1}}(\mathbf{x}_{k+1})\|^2 - \|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2\right] + (c_{k+1} - c_k)\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\
\leq& -\frac{a_{k+1}\alpha_k}{2}\|\nabla \Phi_{\mu_k}(\mathbf{x}_k)\|^2 - \frac{a_{k+1}}{2}\left(\frac{1}{\alpha_k} - \frac{L_{\Phi\mu_k}}{\sigma^2_{\psi_{\mu_k}}}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + a_{k+1}\alpha_k L^2_{\psi_{\mathbf{y}1}}\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\
&+ a_{k+1}\alpha_k\left(L_{\psi_{\mathbf{xy}2}}\|\mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\| + L_{F_{\mathbf{x}2}}\right)^2\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{2a_{k+1}\left\|\nabla_{\mathbf{y}} F_{k+1}\right\|^2}{\sigma_F}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right) \\
&- \frac{b_{k+1}}{2}\beta_k \sigma_{\psi_{\mu_k}}\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{6b_{k+1}L^2_{\psi_{\mathbf{y}1}}}{\beta_k \sigma^3_{\psi_{\mu_k}}}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 + \frac{24 b_{k+1}\|\nabla_{\mathbf{y}} F_{k+1}\|^2}{\sigma^2_F \beta_k \sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k}\right)^2 \\
&- \frac{c_{k+1}}{2}\eta_k \sigma_{\psi_{\mu_k}}\|\mathbf{v}_k - \mathbf{v}^*_{\mu_k}(\mathbf{x}_k)\|^2 + \frac{c_{k+1}B_{\mathbf{vy}}\eta_k}{\sigma_{\psi_{\mu_k}}}\|\mathbf{y}_k - \mathbf{y}^*_{\mu_k}(\mathbf{x}_k)\|^2 \\
&+ \frac{c_{k+1}B_{\mathbf{vx}}}{\eta_k \sigma^5_{\psi_{\mu_k}}}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 + \frac{c_{k+1}B_{\mathbf{v}\mu_{k+1}}}{\eta_k \sigma_{\psi_{\mu_k}}}\left(\frac{\mu_k - \mu_{k+1}}{\mu_k^2}\right)^2,
\end{aligned}
$$

which implies the desired result. □