Journal of Machine Learning Research 25 (2024) 1-51 Submitted 10/23; Revised 04/24; Published 05/24

Optimal Algorithms for Stochastic Bilevel Optimization under Relaxed Smoothness Conditions *

Xuxing Chen XUXCHEN@UCDAVIS.EDU
Department of Mathematics
University of California, Davis

Tesi Xiao † TESIXIAO.STATS@GMAIL.COM
Amazon Inc.

Krishnakumar Balasubramanian KBALA@UCDAVIS.EDU
Department of Statistics
University of California, Davis

Editor: Mladen Kolar

Abstract

We consider stochastic bilevel optimization problems involving minimizing an upper-level (`UL`) function that is dependent on the arg-min of a strongly-convex lower-level (`LL`) function. Several algorithms utilize Neumann series to approximate certain matrix inverses involved in estimating the implicit gradient of the `UL` function (hypergradient). The state-of-the-art StOchastic Bilevel Algorithm (`SOBA`) (Dagréou et al., 2022) instead uses stochastic gradient descent steps to solve the linear system associated with the explicit matrix inversion. This modification enables `SOBA` to obtain a sample complexity of $\mathcal{O}(1/\epsilon^2)$ for finding an ϵ-stationary point. Unfortunately, the current analysis of `SOBA` relies on the assumption of higher-order smoothness for the `UL` and `LL` functions to achieve optimality. In this paper, we introduce a novel fully single-loop and Hessian-inversion-free algorithmic framework for stochastic bilevel optimization and present a tighter analysis under standard smoothness assumptions (first-order Lipschitzness of the `UL` function and second-order Lipschitzness of the `LL` function). Furthermore, we show that a slight modification of our algorithm can handle a more general multi-objective robust bilevel optimization problem. For this case, we obtain the state-of-the-art oracle complexity results demonstrating the generality of both the proposed algorithmic and analytic frameworks. Numerical experiments demonstrate the performance gain of the proposed algorithms over existing ones.

Keywords: bilevel optimization, optimal complexity, relaxed smoothness, stochastic optimization, non-convex optimization

1. Introduction

Bilevel optimization is gaining increasing popularity within the machine learning community due to its extensive range of applications, including meta-learning (Bertinetto et al., 2019; Franceschi et al., 2018; Rajeswaran et al., 2019; Ji et al., 2020), hyperparameter optimization (Bengio, 2000; Franceschi et al., 2018; Bertrand et al., 2020), data augmentation (Cubuk et al., 2019; Rommel et al., 2022), and neural architecture search (Liu et al., 2019; Zhang

*The first two authors contributed equally to this work.
†This work was done prior to joining Amazon.

et al., 2022). The objective of bilevel optimization is to minimize a function that is dependent on the solution of another optimization problem. Formally, we have

$$\min_{x\in\mathcal{X}\subseteq\mathbb{R}^{d_x}} \Phi(x) := f(x, y^*(x)) \qquad \text{s.t.} \quad y^*(x) = \arg\min_{y\in\mathbb{R}^{d_y}} g(x, y) \tag{1}$$

where the upper-level (UL) function f (a.k.a. *outer function*) and the lower-level (LL) function g (a.k.a. *inner function*) are two real-valued functions defined on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. The set $\mathcal{X}$ is either $\mathbb{R}^{d_x}$ or a closed convex set in $\mathbb{R}^{d_x}$, and the LL function g is strongly convex. We call x the *outer variable* and y the *inner variable*. The objective function $\Phi(x)$ is called the *value function*. In this paper, we consider the stochastic setting in which the f and g are expressed in the form of expectations, i.e., $f(x, y) = \mathbb{E}_{\xi\sim\mathcal{D}_f}\left[F(x, y; \xi)\right]$, $g(x, y) = \mathbb{E}_{\phi\sim\mathcal{D}_g}\left[G(x, y; \phi)\right]$. Stochastic bilevel optimization can be considered as an extension of bilevel empirical risk minimization (Dagréou et al., 2023), allowing to effectively handle streaming data (ξ, ϕ).

In many instances, the analytical expression of $y^*(x)$ is unknown and can only be approximated using an optimization algorithm. This adds to the complexity of problem (1) compared to its single-level counterpart. Under regular conditions such that Φ is differentiable, the *hypergradient* $\nabla\Phi(x)$ derived by chain rule and implicit function theorem is given by

$$\nabla\Phi(x) = \nabla_1 f(x, y^*(x)) - \nabla^2_{12} g(x, y^*(x)) z^*(x), \tag{2}$$

where $z^*(x) \in \mathbb{R}^{d_y}$ is the solution of a linear system:

$$z^*(x) = \left[\nabla^2_{22} g(x, y^*(x))\right]^{-1} \nabla_2 f(x, y^*(x)). \tag{3}$$

Solving (1) using only stochastic oracles poses significant challenges since there is no direct unbiased estimator available for $\left[\nabla^2_{22} g(x, y^*(x))\right]^{-1}$ and also for $\nabla\Phi(x)$ as a consequence.

To mitigate the estimation bias, many existing methods (Ghadimi and Wang, 2018; Ji et al., 2021; Yang et al., 2021; Hong et al., 2023; Guo et al., 2021b; Khanduri et al., 2021; Chen et al., 2021a; Akhtar et al., 2022) employ a Hessian Inverse Approximation (HIA) subroutine, which involves drawing a mini-batch of stochastic Hessian matrices and computing a truncated Neumann series (Stewart, 1998). However, this subroutine comes with an increased computational burden and introduces an additional factor of $\log(\epsilon^{-1})$ in the sample complexity. Alternative methods proposed by Chen et al. (2022) and Guo et al. (2021a) calculate the explicit inverse of the stochastic Hessian matrix with momentum updates. To circumvent the need for explicit Hessian inversion and the HIA subroutine, Arbel and Mairal (2022); Dagréou et al. (2022) propose running Stochastic Gradient Descent (SGD) steps to approximate the solution $z^*(x)$ of the linear system (3). In particular, the state-of-the-art Stochastic Bilevel Algorithm (SOBA) *only* utilizes SGD steps to simultaneously update three variables: the inner variable y, the outer variable x, and the auxiliary variable z. It was claimed that SOBA achieves the same complexity lower bound of its single-level counterpart ($\Phi \in \mathcal{C}_L^{1,1}$ [‡‡]) in the non-convex setting (Arjevani et al., 2023).

Despite the superior computational and sample efficiency of SOBA, there is crucial shortcoming as the current theoretical framework assumes high-order smoothness for the UL function f and the LL function g such that $z^*(x)$ has Lipschitz gradient. Specifically, unlike

[‡‡] $\mathcal{C}_L^{p,p}$ denotes p-times differentiability with Lipschitz k-th order derivatives for $0 < k \leq p$.

the typical assumptions in stochastic bilevel optimization that state $f \in \mathcal{C}_L^{1,1}$ and $g \in \mathcal{C}_L^{2,2}$ (A1), the current theory of SOBA requires $f \in \mathcal{C}_L^{2,2}$ and $g \in \mathcal{C}_L^{3,3}$ (A2). The necessity of (A2) is counter-intuitive as the partial gradients of x, y, z utilized in constructing SGD steps are already Lipschitz continuous under (A1). Furthermore, assuming g is strongly convex and the partial gradient of the UL function with respect to the inner variable y is bounded for all pairs of $(x, y^*(x))$, (i.e., $\|\nabla_2 f(x, y^*(x))\| \leq L_f$ for all $x \in \mathcal{X}$), there exists a subset relation among three function classes as indicated by Ghadimi and Wang (2018, Lemma 2.2) that

$$\text{(A2)}\ \{f \in \mathcal{C}_L^{2,2}, g \in \mathcal{C}_L^{3,3}\} \subset\ \text{(A1)}\ \{f \in \mathcal{C}_L^{1,1}, g \in \mathcal{C}_L^{2,2}\} \subset \{\Phi \in \mathcal{C}_L^{1,1}\}.$$

In light of this, it can be concluded that (A1) is sufficient to ensure the first-order Lipschitzness of Φ, which is the standard assumption in the single-level setting. On the other hand, it is worth noting that under (A2) it can be shown that $\Phi \in \mathcal{C}_L^{2,2}$, i.e., $\nabla\Phi(x)$ and $\nabla^2\Phi(x)$ are both Lipschitz continuous. It is known that higher order smoothness (e.g., Lipschitz continuity of $\nabla^2\Phi(x)$) will lead to better sample complexity (Carmon et al., 2017; Arjevani et al., 2020). This indicates that the sample complexity $\mathcal{O}(\epsilon^{-2})$ obtained in Dagréou et al. (2022) may not be optimal under the set of assumptions made in their work.

Therefore, a natural question follows: *Is it possible to develop a* fully single-loop *and* Hessian-inversion-free *algorithm for solving stochastic bilevel optimization problems that achieves an* optimal *sample complexity of* $\mathcal{O}(\epsilon^{-2})$ *under* standard smoothness assumptions $\{f \in \mathcal{C}_L^{1,1}, g \in \mathcal{C}_L^{2,2}\}$[‡‡]*?* In this paper, we provide an affirmative answer to the aforementioned question. Our **contributions** can be summarized as follows.

- We propose a class of fully single-loop and Hession-inversion-free algorithm, named Moving-Average SOBA (MA-SOBA), which builds upon the SOBA algorithm by incorporating an additional sequence of average hypergradients. Unlike SOBA, MA-SOBA achieves an optimal sample complexity of $\mathcal{O}(\epsilon^{-2})$ under standard smoothness assumptions, without relying on high-order smoothness. In particular, in Section 7.1.1 we explain how the introduced MA updates help reduce the order of bias in hypergradient estimation, and avoid higher order Taylor expansion (which requires higher-order smoothness of f and g) used in Dagréou et al. (2022). Moreover, the introduced sequence of average hypergradients converges to $\nabla\Phi(x)$, thus offering a reliable termination criterion in the stochastic setting.

- We expand the scope of MA-SOBA to tackle a broader class of problems, specifically the min-max multi-objective bilevel optimization problem with significant applications in robust machine learning. We introduce MORMA-SOBA, an algorithm that can find an ϵ-first-order stationary point of the μ_λ-strongly-concave regularized formulation while achieving a sample complexity of $\mathcal{O}(n^5\mu_\lambda^{-4}\epsilon^{-2})$, which fills a gap (in terms of the order of ϵ-dependency) in the existing literature.

- We conduct experiments on several machine learning problems. Our numerical results show the efficiency and superiority of our algorithms.

[‡] All methods also assume $\|\nabla_2 f(x, y^*(x))\| \leq L_f < \infty$ for all $x \in \mathcal{X}$.

[¶] ALSET can achieve convergence without the need for double loops, but it comes at the cost of a worse dependence on κ in sample complexity. The mechanisms of single-loop ALSET and TTSA are essentially the same, except that ALSET employs single time-scale stepsizes while TTSA employs two time-scales.

Method (double-loop)	Sample Complexity	(UL) f ‡	(LL) g	Hession Inversion	Inner Loop	Batch Size
BSA (Ghadimi and Wang, 2018)	$\tilde{\mathcal{O}}(\epsilon^{-3})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	Neumann approx.	SGD on inner	$\tilde{\mathcal{O}}(1)$
stocBiO (Ji et al., 2021)	$\tilde{\mathcal{O}}(\epsilon^{-2})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	Neumann approx.	SGD on inner	$\tilde{\mathcal{O}}(\epsilon^{-1})$
¶ALSET (Chen et al., 2021a)	$\tilde{\mathcal{O}}(\epsilon^{-2})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	Neumann approx.	SGD on inner	$\tilde{\mathcal{O}}(1)$
AmIGO (Arbel and Mairal, 2022)	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	SGD	SGD on inner	$\mathcal{O}(\epsilon^{-1})$
Method (single-loop)	Sample Complexity	(UL) f‡	(LL) g	Hession Inversion	Inner Step	Batch Size
¶TTSA (Hong et al., 2023)	$\tilde{\mathcal{O}}(\epsilon^{-2.5})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	Neumann approx.	SGD	$\tilde{\mathcal{O}}(1)$
STABLE (Chen et al., 2022)	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	Direct	SGD	$\mathcal{O}(1)$
SOBA (Dagréou et al., 2022)	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{3,3}$	SGD	SGD	$\mathcal{O}(1)$
MA-SOBA (Alg. 1)	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	SGD	SGD	$\mathcal{O}(1)$

Table 1: Comparison of the stochastic bilevel optimization solvers in the nonconvex-strongly-convex setting under smoothness assumptions ‡‡ on f and g. We omit the comparison with variance reduction-based methods (VRBO, MRBO (Yang et al., 2021); SUSTAIN (Khanduri et al., 2021); SABA (Yang et al., 2021); SRBA (Dagréou et al., 2023); SVRB (Guo et al., 2021a); FLSA (Li et al., 2022); SBFW (Akhtar et al., 2022)) that may achieve $\mathcal{O}(\epsilon^{-1.5})$ sample complexity and under mean-squared smoothness assumptions on stochastic functions F_ξ and G_ϕ, and SBMA (Guo et al., 2021b) that achieves $\mathcal{O}(\epsilon^{-4})$ sample complexity. The sample complexity corresponds to the number of calls to stochastic gradients and Hessian(Jacobian)-vector products to get an ϵ-stationary point. The $\tilde{\mathcal{O}}$ notation hides a factor of $\log(\epsilon^{-1})$. "SC" means "strongly-convex".

Related Work. The concept of bilevel optimization was initially introduced in the work of Bracken and McGill (1973). Since then, numerous gradient-based bilevel optimization algorithms have been proposed, broadly categorized into two groups: ITerative Differentiation (ITD) based methods (Domke, 2012; Maclaurin et al., 2015; Franceschi et al., 2018; Grazzi et al., 2020; Ji et al., 2021) and Approximate Implicit Differentiation (AID) based methods (Domke, 2012; Pedregosa, 2016; Gould et al., 2016; Ghadimi and Wang, 2018; Grazzi et al., 2020; Ji et al., 2021; Arbel and Mairal, 2022; Grazzi et al., 2023). The ITD-based algorithms typically involve approximating the solution of the inner problem using an iterative algorithm and then computing an approximate hypergradient through automatic differentiation. However, a major drawback of this approach is the necessity of storing each iterate of the inner optimization algorithm in memory. The AID-based algorithms leverage the implicit gradient given by (2), which requires the solution of a linear system characterized by (3). Extensive research has been conducted on designing and analyzing deterministic bilevel optimization algorithms with strongly-convex LL functions; see Ji et al. (2021) and references therein.

In recent years, there has been a growing interest in stochastic bilevel optimization, especially in the setting of a non-convex UL function and a strongly-convex LL function. To address estimation bias, one set of methods uses SGD iterations for the inner problem and employs truncated stochastic Neumann series to approximate the inverse of the Hessian

matrix in $z^*(x)$ (Ghadimi and Wang, 2018; Ji et al., 2021; Yang et al., 2021; Hong et al., 2023; Guo et al., 2021b; Khanduri et al., 2021; Chen et al., 2021a; Akhtar et al., 2022). The analysis of such methods was refined by (Chen et al., 2021a) to achieve convergence rates similar to those of SGD. However, Neumann approximation subroutine introduces an additional factor of $\log(\epsilon^{-1})$ in the sample complexity. Some alternative approaches Arbel and Mairal (2022); Chen et al. (2022); Guo et al. (2021a) calculate the explicit inverse of the stochastic Hessian matrix with momentum updates. Nevertheless, these methods encounter challenges related to computational complexity in matrix inversion and numerical stability.

To avoid the need for explicit Hessian inversion and Neumann approximation, recent algorithms (Arbel and Mairal, 2022; Dagréou et al., 2022) propose running SGD steps to approximate the solution $z^*(x)$ of the linear system (3). One such algorithm called `AmIGO` (Arbel and Mairal, 2022) employs a double-loop approach with warm-start strategy and achieves an optimal sample complexity of $\mathcal{O}(\epsilon^{-2})$ under regular assumptions. However, `AmIGO` requires a growing batch size inversely proportional to ϵ. Following `AmIGO`, Grazzi et al. (2023) proposes `BSGM`, which avoids using large batch size in the `LL` problem and warm-start strategy, but still requires double-loop framework and large batch sizes in the `UL` problem. On the other hand, the single-loop algorithm `SOBA` (Dagréou et al., 2022) achieves the same complexity lower bound but with constant batch size. Unfortunately, the current analysis of `SOBA` relies on the assumption of higher-order smoothness for the `UL` and `LL` functions. In this work, we introduce a novel algorithm framework that differs slightly from `SOBA` but can achieve optimal sample complexity in theory without higher-order smoothness assumptions. A summary of our results and comparison to prior work is provided in Table 1.

In addition, there exist several variance reduction-based methods following the line of research by Yang et al. (2021); Khanduri et al. (2021); Yang et al. (2021); Dagréou et al. (2023); Guo et al. (2021a); Li et al. (2022). Some of these methods achieve an improved sample complexity of $\mathcal{O}(\epsilon^{-1.5})$ and match the lower bounds of their single-level counterparts when stochastic functions F_ξ and G_ϕ satisfy mean-squared smoothness assumptions and the algorithm is allowed simultaneous queries at the same random seed (Arjevani et al., 2023). However, since we are specifically considering smoothness assumptions on f and g, we will not delve into the comparison with these methods.

The most recent advancements in (stochastic) bilevel optimization focus on several new ideas: (i) addressing constrained lower-level problems (Shen and Chen, 2023; Xiao et al., 2023; Tsaknakis et al., 2022; Giovannelli et al., 2021), (ii) handling lower-level problems that lack strong convexity (Chen et al., 2023a; Huang, 2023; Liu et al., 2023, 2021; Sow et al., 2022a; Jiang et al., 2023), (iii) developing fully first-order (Hessian-free) algorithms (Liu et al., 2022; Kwon et al., 2023; Sow et al., 2022b), (iv) establishing convergence to the second-order stationary point (Huang et al., 2023), and (v) expanding the framework to encompass multi-objective optimization problems (Giovannelli et al., 2023; Gu et al., 2023; Hu et al., 2022). It is promising to apply some of these advancements to our specific framework. Moreover, in this work, we also contribute to multi-objective bilevel problems with a slight modification of our approach. Other directions are left as future work.

Notation. We use $\|\cdot\|$ for ℓ^2 norm. $\mathbf{1}_n$ denotes the all-one vector in $\mathbb{R}^n$. $\Delta_n = \{\lambda \mid \lambda_i \geq 0, \sum_{i=1}^n \lambda_i = 1\}$ denotes the probability simplex. $\Pi_{\mathcal{X}}(\cdot)$ denotes the projection onto $\mathcal{X}$.

2. Proposed Framework: the MA-SOBA Algorithm

Similar to Dagréou et al. (2022); Arbel and Mairal (2022), our algorithm initiates with inexact hypergradient descent techniques and seeks to offer an alternative in the stochastic setting. To provide a clear illustration, let us initially consider the deterministic setting. The `SOBA` framework keeps track of three sequences, denoted as $\{x^k, y^k, z^k\}$, and updates them using D_x, D_y, D_z as follows:

$$\textbf{(inner)} \quad y^{k+1} = y^k - \beta_k \boxed{\nabla_2 g(x^k, y^k)} = y^k - \beta_k D_y(x^k, y^k, z^k) \tag{4}$$

$$\textbf{(aux)} \quad z^{k+1} = z^k - \gamma_k \left\{\nabla^2_{22} g(x^k, y^*(x^k)) z^k - \nabla_2 f(x^k, y^*(x^k))\right\}$$

$$\textbf{bias} \to \approx z^k - \gamma_k \boxed{\left\{\nabla^2_{22} g(x^k, y^k) z^k - \nabla_2 f(x^k, y^k)\right\}} = z^k - \gamma_k D_z(x^k, y^k, z^k) \tag{5}$$

$$\textbf{(outer)} \quad x^{k+1} = x^k - \alpha_k \left\{\nabla_1 f(x^k, y^*(x^k)) - \nabla^2_{12} g(x^k, y^*(x^k)) z^*(x^k)\right\} = x^k - \alpha_k \nabla \Phi(x^k)$$

$$\textbf{bias} \to \approx x^k - \alpha_k \boxed{\left\{\nabla_1 f(x^k, y^k) - \nabla^2_{12} g(x^k, y^k) z^k\right\}} = x^k - \alpha_k D_x(x^k, y^k, z^k) \tag{6}$$

where (4) is the GD step to minimize $g(x^k, \cdot)$, (6) is the inexact hyper gradient descent step, and (5) is the GD step to minimize a quadratic function with $z^*(x^k)$ being the solution, i.e.,

$$z^*(x^k) = \arg\min_z \ \frac{1}{2} \left\langle \nabla^2_{22} g(x^k, y^*(x^k)) z, z \right\rangle - \left\langle \nabla_2 f(x^k, y^*(x^k)), z \right\rangle.$$

Given that the above update rule, highlighted in blue, does not involve the Hessian matrix inversion, `SOBA` can directly utilize the stochastic oracles of $\nabla_1 f, \nabla_2 f, \nabla_2 g, \nabla^2_{22} g, \nabla^2_{12} g$ to obtain unbiased estimators of D_x, D_y, D_z in Eq.(4), (5), (6). This approach circumvents the requirement for a Neumann approximation subroutine or a direct matrix inversion. However, due to the update rule for y, which only utilizes one-step SGD at each iteration k, the value of y^k does not coincide with $y^*(x^k)$. As a result, a certain bias is introduced in the partial gradient of z in Eq.(5). Similarly, when estimating the hypergradient $\nabla \Phi(x)$, another bias term arises in Eq.(6). Although the bias decreases to zero as $y^k \to y^*(x^k)$ and $z^k \to z^*(x^k)$ under standard smoothness assumptions as indicated by Lemma 3.4 in Dagréou et al. (2022), the current analysis of `SOBA` requires more regularity on f and g to carefully handle the bias; it assume that f has Lipschitz Hessian and g has Lipschitz third-order derivative.

The inability to obtain an unbiased gradient estimator is a common characteristic in stochastic optimization involving nested structures; see, for example, stochastic compositional optimization (Wang et al., 2017; Yang et al., 2019; Ghadimi et al., 2020; Balasubramanian et al., 2022; Chen et al., 2021b) as a specific case of (1). One popular approach is to introduce a sequence of dual variables that approximates the true gradient by aggregating all past biased stochastic gradients using a moving averaging technique (Ghadimi et al., 2020; Balasubramanian et al., 2022; Xiao et al., 2022). Motivated by this approach, we introduce another sequence of variables, denoted as $\{h^k\}$, and update it at k-th iteration given the past iterates $\mathscr{F}_k$ as $h^{k+1} = (1-\theta_k)h^k + \theta_k w^{k+1}$, where $\mathbb{E}[w^{k+1}|\mathscr{F}_k] = D_x(x^k, y^k, z^k), \theta_k \in (0, 1]$. Following the update rule in the constrained setting ($\mathcal{X} \subset \mathbb{R}^{d_x}$) (Ghadimi et al., 2020), the outer variable is updated as $x^{k+1} = x^k + \alpha_k \left(\Pi_{\mathcal{X}}(x^k - \tau h^k) - x^k\right)$, which is reduced to the GD step when $\mathcal{X} \equiv \mathbb{R}^{d_x}$. Denote the stochastic oracles of $\nabla_1 f(x^k, y^k), \nabla_2 f(x^k, y^k), \nabla_2 g(x^k, y^k)$,

Algorithm 1: `Moving-Average SOBA`

Input: $x^0, y^0, z^0, h^0 = 0, \{\alpha_k\}, \{\beta_k\}, \{\gamma_k\}, \{\theta_k\}$

for $k = 0, 1, \ldots, K-1$ **do**

- $x^{k+1} = x^k + \alpha_k \left(\Pi_{\mathcal{X}}(x^k - \tau h^k) - x^k\right)$ # update x^k via average hypergradient h^k
- $y^{k+1} = y^k - \beta_k v^{k+1}$ # update y^k by one-step SGD based on (4)
- $z^{k+1} = z^k - \gamma_k(H^{k+1} z^k - u_y^{k+1})$ # update z^k by one-step SGD based on (5)
- $h^{k+1} = (1-\theta_k)h^k + \theta_k \left(u_x^{k+1} - J^{k+1} z^k\right)$ # update average hypergradient h^k

end

$\nabla^2_{22} g(x^k, y^k), \nabla^2_{12} g(x^k, y^k)$ at k-th iteration as $u_x^{k+1}, u_y^{k+1}, v^{k+1}, H^{k+1}, J^{k+1}$ respectively. We present our method, referred to as **M**oving-**A**verage `SOBA` (`MA-SOBA`), in Algorithm 1.

3. Theoretical Analysis

In this section, we provide convergence rates of `MA-SOBA` under *standard* smoothness conditions on f, g and *regular* assumptions on stochastic oracles. We also present a proof sketch and have detailed discussions about assumptions made in the literature. The complete proofs are deferred in Section 7.

3.1 Preliminaries and Assumptions

As we consider the general setting in which $\mathcal{X}$ can be either $\mathbb{R}^{d_x}$ or a closed convex set in $\mathbb{R}^{d_x}$, we use the notion of gradient mapping to characterize the first-order stationarity, which is a classical measure widely used in the literature as a convergence criterion when solving nonconvex constrained problems (Nesterov, 2018). For $\tau > 0$, we define the gradient mapping of at point $\bar{x} \in \mathcal{X}$ as $\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla\Phi(\bar{x}), \tau) := \frac{1}{\tau}(\bar{x} - \Pi_{\mathcal{X}}(\bar{x} - \tau\nabla\Phi(\bar{x})))$. When $\mathcal{X} \equiv \mathbb{R}^d$, the gradient mapping simplifies to $\nabla\Phi(\bar{x})$. Our main goal in this work is to find an ϵ-stationary solution to (1), in the sense of $\mathbb{E}[\|\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla\Phi(\bar{x}), \tau)\|^2] \leq \epsilon$.

We first state some regularity assumptions on the functions f and g.

Assumption 1 *The functions f and g satisfy: (a) $f \in \mathcal{C}_L^{1,1}$, $g \in \mathcal{C}_L^{2,2}$ and $\nabla f, \nabla g, \nabla^2 g$ are $L_{\nabla f}, L_{\nabla g}, L_{\nabla^2 g}$ Lipschitz continuous respectively, (b) g is μ_g-strongly convex, and (c) $\|\nabla_2 f(x, y^*(x))\| \leq L_f < \infty$ for all $x \in \mathcal{X}$.*

Remark 1 *The above assumption serves as a sufficient condition for the Lipschitz continuity of $\nabla\Phi$, $y^*(x)$, and $z^*(x)$, as well as D_x, D_y, and D_z in Eq. (4), (5), (6). The inclusion of high-order smoothness assumptions ($f \in \mathcal{C}_L^{2,2}$ and $g \in \mathcal{C}_L^{3,3}$) in the current analysis of `SOBA` (Dagréou et al., 2022) is primarily intended to ensure the Lipschitzness of $\nabla z^*(x)$. However, the necessity of such assumptions is subject to doubt, given that $\nabla z^*(x)$ is not involved in designing the algorithm. Furthermore, the Lipschitzness of f or uniformly boundedness of $\nabla_2 f$ made in several previous works is unnecessary. Instead, the boundedness assumption on $\nabla_2 f$ is only required for all pairs of $(x, y^*(x))$ as demonstrated by Assumption 1(c).*

Next, we discuss assumptions made on the stochastic oracles.

Assumption 2 *For any $k \geq 0$, denote by $\mathscr{F}_k$ the sigma algebra generated by all iterates with superscripts not greater than k : $\mathscr{F}_k = \sigma\left\{h^1, \ldots, h^k, x^1, \ldots, x^k, y^1, \ldots, y^k, z^1, \ldots, z^k\right\}$. The **stochastic oracles** of $\nabla_1 f(x^k, y^k), \nabla_2 f(x^k, y^k), \nabla_2 g(x^k, y^k), \nabla^2_{22} g(x^k, y^k), \nabla^2_{12} g(x^k, y^k)$, denoted as $u_x^{k+1}, u_y^{k+1}, v^{k+1}, H^{k+1}, J^{k+1}$ respectively, used in Algorithm 2 at k-th iteration are **unbiased** with **bounded variance** given $\mathscr{F}_k$, i.e., there exist positive constants $\sigma_{f,1}, \sigma_{f,2}, \sigma_{g,1}, \sigma_{g,2}$ such that*

$$\begin{aligned}
&\mathbb{E}[u_x^{k+1}|\mathscr{F}_k] = \nabla_1 f(x^k, y^k),\ \mathbb{E}\left[\|u_x^{k+1} - \nabla_1 f(x^k, y^k)\|^2|\mathscr{F}_k\right] \leq \sigma_{f,1}^2,\\
&\mathbb{E}\left[u_y^{k+1}|\mathscr{F}_k\right] = \nabla_2 f(x^k, y^k),\ \mathbb{E}\left[\|u_y^{k+1} - \nabla_2 f(x^k, y^k)\|^2|\mathscr{F}_k\right] \leq \sigma_{f,2}^2,\\
&\mathbb{E}\left[v^{k+1}|\mathscr{F}_k\right] = \nabla_2 g(x^k, y^k),\ \mathbb{E}\left[\|v^{k+1} - \nabla_2 g(x^k, y^k)\|^2|\mathscr{F}_k\right] \leq \sigma_{g,1}^2,\\
&\mathbb{E}\left[H^{k+1}|\mathscr{F}_k\right] = \nabla^2_{22} g(x^k, y^k),\ \mathbb{E}\left[\|H^{k+1} - \nabla^2_{22} g(x^k, y^k)\|^2|\mathscr{F}_k\right] \leq \sigma_{g,2}^2,\\
&\mathbb{E}\left[J^{k+1}|\mathscr{F}_k\right] = \nabla^2_{12} g(x^k, y^k),\ \mathbb{E}\left[\|J^{k+1} - \nabla^2_{12} g(x^k, y^k)\|^2|\mathscr{F}_k\right] \leq \sigma_{g,2}^2.
\end{aligned}$$

*In addition, they are conditionally **independent** conditioned on $\mathscr{F}_k$.*

Remark 2 *The unbiasedness and bounded variance assumptions on stochastic oracles are standard and typically satisfied in several practical stochastic optimization problems (Lan, 2020). It is important to highlight that we explicitly impose these assumptions on the stochastic oracles, unlike Assumption 3.6 in Dagréou et al. (2022), which assumes $\mathbb{E}[\|v^{k+1}\|^2|\mathscr{F}_k] \leq B_y^2(1 + \|D_y(x^k, y^k, z^k)\|^2)$ and $\mathbb{E}[\|H^{k+1}z^k - u_y^{k+1}\|^2|\mathscr{F}_k] \leq B_z^2(1 + \|D_z(x^k, y^k, z^k)\|^2)$. In this case, B_y and B_z represent constants in terms of the Lipschitz constants (L) and variance bounds (σ^2). Moreover, Assumption 3.7 in Dagréou et al. (2022) assumes $\mathbb{E}[\|w^{k+1}\|^2|\mathscr{F}_k] \leq B_x^2$ holds for a constant B_x, which is considerably stronger than our assumptions and may not hold for a broad class of problems.*

3.2 Convergence Results

We have the following theorem characterizing the convergence results of `MA-SOBA`.

Theorem 3 *Define $x_+^k = \Pi_{\mathcal{X}}(x^k - \tau h^k)$. Suppose Assumptions 1 and 2 hold. Then there exist positive constants $c_1, c_2, c_3, \tau > 0$ such that if $\alpha_k \equiv \Theta(1/\sqrt{K})$, $\beta_k = c_1\alpha_k, \gamma_k = c_2\alpha_k, \theta_k = c_3\alpha_k$, in Algorithm 1, then the iterates in Algorithm 1 satisfy*

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau^2}\mathbb{E}[\|x_+^k - x^k\|^2] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \quad \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|h^k - \nabla\Phi(x^k)\|^2] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \tag{7}$$

which imply

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau^2}\mathbb{E}[\|\left(x^k - \Pi_{\mathcal{X}}(x^k - \tau\nabla\Phi(x^k))\right)\|^2] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

That is to say, when uniformly randomly selecting a solution x^R from $\{x^1, \ldots, x^K\}$, the sample complexity of Algorithm 1 for finding an ϵ-stationary point is $\mathcal{O}(\epsilon^{-2})$.

Remark 4 *In contrast to most existing methods, in* `MA-SOBA`*, the introduced sequence of dual variables* $\{h^k\}$ *converges to the exact hypergradient* $\nabla\Phi(x)$*, even in the presence of estimation bias. This attribute provides reliable terminating criteria in practice. In addition, similar results with an extra factor of* $\log(K)$ *in the convergence rate can be established under decreasing* α_k *(Dagréou et al., 2022). We also note that Algorithm 1 only requires stochastic gradient and Hessian(Jacobian)-vector product oracles, whose computational complexity are typically* $\mathcal{O}(\max(d_x, d_y))$ *with the help of automatic differentiation techniques (Pearlmutter, 1994; ?). Moreover, the sample complexity of fully first-order methods for bilevel optimization usually have worse dependency on* ϵ *(Kwon et al., 2023).*

3.3 Proof Sketch of Theorem 3

Define $V_k = \frac{1}{\tau^2}\|x_+^k - x^k\|^2 + \|h^k - \nabla\Phi(x^k)\|^2$. To obtain (7), we consider the merit function:

$$W_k = \Phi(x^k) - \eta_{\mathcal{X}}(x^k, h^k, \tau) + \|y^k - y_*^k\|^2 + \|z^k - z_*^k\|^2,$$

where $\eta_{\mathcal{X}}(x, h, \tau) = \langle h, x_+ - x\rangle + \frac{1}{2\tau}\|x_+ - x\|^2$. By leveraging the moving-average updates of x^k (line 2 of Algorithm 1), we can obtain

$$\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[V_k\right] = \mathcal{O}\bigg(\sum_{k=0}^{K}\left(\alpha_k\mathbb{E}[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Phi(x^k)\|^2] + \alpha_k^2\right)\bigg),$$

which reduces the error analysis to controlling the hypergradient estimation bias, i.e., $\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Phi(x^k)\|^2$. This term, by the construction of w^{k+1}, satisfies

$$\textstyle\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Phi(x^k)\|^2\right] = \mathcal{O}\left(\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\|x_+^k - x^k\|^2 + \|y^k - y_*^k\|^2 + \|z^k - z_*^k\|^2\right]\right).$$

It is worth noting that Dagréou et al. (2022) requires the existence and Lipschitzness of $\nabla^2 f$ and $\nabla^3 g$ to ensure the Lipschitzness of $\nabla z^*(x)$ (see (3)) which is used in proving the sufficient decrease of $\|z^k - z_*^k\|^2$. In contrast, based on the moving-average updates of x^k and h^k, our refined analysis does not necessitate such assumptions to obtain that

$$\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|y^k - y_*^k\|^2 + \|z^k - z_*^k\|^2] = \mathcal{O}\bigg(\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|x_+^k - x^k\|^2]\bigg).$$

The proof of Theorem 3 can then be completed by choosing appropriate $\alpha_k, c_1, c_2, c_3, \tau > 0$.

4. Min-Max Bilevel Optimization

To incorporate robustness in the multi-objective setting where each objective can be expressed as a bilevel optimization problem in (1), the following mini-max bilevel problem formulation was proposed in Gu et al. (2023):

$$\min_{x\in\mathcal{X}}\max_{1\le i\le n}\ \Phi_i(x) \coloneqq f_i(x, y_i^*(x)) \qquad \text{s.t.}\quad y_i^*(x) = \operatorname*{arg\,min}_{y_i\in\mathbb{R}^{d_{y_i}}} g_i(x, y_i), 1 \le i \le n. \tag{8}$$

Note that (8) can be reformulated as a general nonconvex-concave min-max optimization problem (with a bilevel substructure):

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\Delta_n} \Phi(x,\lambda) \coloneqq \sum_{i=1}^{n} \lambda_i \Phi_i(x). \tag{9}$$

Instead of solving (9) directly, in this work, we focus on solving the regularized version,

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\Delta_n} \Phi_{\mu_\lambda}(x,\lambda) \coloneqq \Phi(x,\lambda) - \frac{\mu_\lambda}{2}\|\lambda - \frac{1}{n}\mathbf{1}_n\|^2. \tag{10}$$

Note that in (10), we include an ℓ^2 regularization term that penalizes the discrepancy between λ and $\frac{\mathbf{1}_n}{n}$. When $\mu_\lambda = 0$, it corresponds to equation (8), and as $\mu_\lambda \to +\infty$, it enforces $\lambda = \frac{\mathbf{1}_n}{n}$, leading to direct minimizing of the average loss. It is important to note that minimizing the worst-case loss (i.e., $\max_{1\le i\le n} f_i(x, y_i^*(x))$) does not necessarily imply the minimization of the average loss (i.e., $\frac{1}{n}\sum_{i=1}^n f_i(x, y_i^*(x))$). Therefore, in practice, it may be preferable to select an appropriate $\mu_\lambda > 0$ (Qian et al., 2019; Wang et al., 2021) to strike a balance between these two types of losses. Hu et al. (2022) considered solving a similar problem under stronger assumptions. We defer a detailed discussion to Section A.2.

4.1 Proposed Framework: the MORMA-SOBA Algorithm

The proposed algorithm, which we refer as to **M**ulti-**O**bjective **R**obust `MA-SOBA` (`MORMA-SOBA`), for solving (10) is presented in Algorithm 2. In addition to the basic framework of Algorithm 1, we also maintain a moving-average step in the updates of λ^k for solving the max part of problem 6. It is worth noting that in its single-level counterpart without the inner variable y, the proposed `MORMA-SOBA` algorithm is fundamentally similar to the single-timescale averaged `SGDA` algorithm proposed in the general nonconvex-strongly-concave setting (Qiu et al., 2020). Moreover, our algorithmic framework can be leveraged to solve the distributionally robust compositional optimization problem as discussed in Gao et al. (2021).

Remark 5 (Comparison with `MORBiT` (Gu et al., 2023)) *In contrast to our approach in* (10)*, the work of Gu et al. (2023), for the min-max bilevel problem, attempted to combine* ***`TTSA`*** *(Hong et al., 2023) and* ***`SGDA`*** *(Lin et al., 2020a) to solve the nonconvex-concave problem as* (9)*. However, we identified an issue in Gu et al. (2023) related to the ambiguity and inconsistency in the expectation and filtration, which may not be easily resolved within their current proof framework. As a consequence, their current proof is unable to demonstrate* $\mathbb{E}[\max_{i\in[n]} \|y_i^k - y_i^*(x^{(k-1)})\|^2] \le \tilde{\mathcal{O}}(\sqrt{n}K^{-2/5})$ *as claimed in Theorem 1 (10b) of Gu et al. (2023). Thus, the subsequent arguments made regarding the convergence analysis of* x *and* λ *are incorrect (at least in its current form); see Section A for further discussions. Moreover, the practical implementation of* ***`MORBiT`*** *incorporates momentum and weight decay techniques to optimize the simplex variable* λ*. This approach can be seen as a means of solving the regularized formulation in* (10)*.*

4.2 Convergence Results

We first present additional assumptions required in the analysis of `MORMA-SOBA`.

Algorithm 2: `Multi-Objective Robust Moving-Average SOBA`

Input: $x^0, \lambda^0, \{y_i^0\}, \{z_i^0\}, h_x^0 = 0, h_\lambda^0 = 0, \{\alpha_k\}, \{\beta_k\}, \{\gamma_k\}, \{\theta_k\}$

for $k = 0, 1, \ldots, K-1$ **do**
- $x^{k+1} = x^k + \alpha_k \left(\Pi_{\mathcal{X}}(x^k - \tau_x h_x^k) - x^k\right)$ # update x^k via average hypergradient h_x^k
- $\lambda^{k+1} = \lambda^k + \alpha_k \left(\Pi_{\Delta_n}(\lambda^k + \tau_\lambda h_\lambda^k) - \lambda^k\right)$ # update λ^k via average gradient h_λ^k
- **for** $i = 1, \ldots, n$ *(in parallel)* **do**
  - $y_i^{k+1} = y_i^k - \beta_k v_i^{k+1}$ # update y_i^k by one-step SGD based on (4)
  - $z_i^{k+1} = z_i^k - \gamma_k\left(H_i^{k+1} z_i^k - u_{y,i}^{k+1}\right)$ # update z_i^k by one-step SGD based on (5)
- **end**
- $h_x^{k+1} = (1-\theta_k) h_x^k + \theta_k \sum_{i=1}^n \lambda_i^k \left(u_{x,i}^{k+1} - J_i^{k+1} z_i^k\right)$ # update average hypergradient h_x^k
- $h_\lambda^{k+1} = (1-\theta_k) h_\lambda^k + \theta_k \left(s^{k+1} - \mu_\lambda\left(\lambda^k - \frac{\mathbf{1}_n}{n}\right)\right)$ # update average gradient h_λ^k

end

Assumption 3 *For any $k \geq 0$, functions $\Phi(x), \nabla\Phi_i(x)$ are bounded, functions f_i are L_f-Lipschitze continuous in the second input, and their stochastic versions are unbiased with bounded variance, i.e., there exists $L_\Phi, L_f, \sigma_{f,0} \geq 0$ such that*

$$|\Phi_i(x)| \leq b_\Phi,\ \|\nabla\Phi_i(x)\| \leq L_\Phi,\ |f_i(x,y) - f_i(x,\tilde{y})| \leq L_f\|y-\tilde{y}\|,\ \text{ for all } x, y, \tilde{y}, 1 \leq i \leq n,$$
$$s^{k+1} = \left(s_1^{k+1}, ..., s_n^{k+1}\right)^\top,\ \mathbb{E}\left[s_i^{k+1}|\mathscr{F}_k\right] = f_i(x^k, y_i^k),\ \mathbb{E}\left[\|s_i^{k+1} - f_i(x^k, y_i^k)\|^2|\mathscr{F}_k\right] \leq \sigma_{f,0}^2.$$

$\bigcup_{i=1}^n \left\{u_{x,i}^{k+1}, u_{y,i}^{k+1}, v_i^{k+1}, H_i^{k+1}, J_i^{k+1}\right\} \cup \left\{s^{k+1}\right\}$ *are conditionally independent under $\mathscr{F}_k$.*

We have the following convergence theorem of `MORMA-SOBA`.

Theorem 6 *Suppose Assumptions 1, 2 (for all f_i, g_i) and Assumption 3 hold. Then there exist positive constants $c_1, c_2, c_3, \tau_x, \tau_\lambda > 0$ such that if $\alpha_k \equiv \Theta(1/\sqrt{nK}), \beta_k = c_1\alpha_k, \gamma_k = c_2\alpha_k, \theta_k = c_3\alpha_k, \mu_\lambda < 1$ in Algorithm 2, then the iterates in Algorithm 2 satisfy*

$$\frac{1}{K}\sum_{k=0}^K \frac{1}{\tau_x^2}\mathbb{E}[\|\left(x^k - \Pi_{\mathcal{X}}\left(x^k - \tau_x\nabla\Psi_{\mu_\lambda}(x^k)\right)\right)\|^2] = \mathcal{O}\left(\frac{n^2}{\mu_\lambda^2\sqrt{K}}\right),$$

where $\Psi_{\mu_\lambda}(x) := \max_{\lambda\in\Delta_n} \Phi_{\mu_\lambda}(x,\lambda)$. That is to say, when uniformly randomly selecting a solution x^R from $\{x^1, \ldots, x^K\}$, the sample complexity (the total number of calls to stochastic oracles) of finding an ϵ-stationary point by Algorithm 2 is $\mathcal{O}(n^5\mu_\lambda^{-4}\epsilon^{-2})$.

Theorem 6 indicates that Algorithm 2 is capable of generating an ϵ-first-order stationary point of $\min_x \Psi_{\mu_\lambda}(x)$ with $K \gtrsim n^5\mu_\lambda^{-4}\epsilon^{-2}$. As $\mu_\lambda \to 0$, the problem (10) changes towards the nonconvex-concave problem (9) and the sample complexity becomes worse, which to some extent implies the difficulty of directly solving (9). We defer the proof details to Section 7.2. For Problem (9), we adopt the definition of ϵ-stationary point in Definition 3.5 in Lin et al. (2020b), and choose $\mu_\lambda = \mathcal{O}(\sqrt{\epsilon})$ to help shed light on the sample complexity.

Corollary 7 *Under the same setup of Theorem 6, setting $\mu_\lambda = \mathcal{O}(\sqrt{\epsilon})$, the sample complexity of finding an ϵ-stationary point of Problem (9) via Algorithm 2 is $\mathcal{O}(n^5\epsilon^{-4})$.*

Remark 8 *Note that in Theorem 6 we explicitly characterize the dependency on n and μ_λ in the convergence rate and the sample complexity. It is worth noting that two variants of stochastic gradient descent ascent (`SGDA`) algorithms for solving the nonconvex-strongly-concave min-max optimization problems (without bilevel substructures), have been studied in Lin et al. (2020a); Qiu et al. (2020). While such algorithms are not immediately applicable to solve* (10) *due to the presence of the additional bilevel substructure, it is instructive to compare to those methods assuming direct access to $y_i^*(x)$ in* (8)*. Specifically, we observe that the sample complexity of `SGDA` with batch size $M = \Theta(n^{1.5}\epsilon^{-1})$ in Lin et al. (2020a) and moving-average `SGDA` with $\mathcal{O}(1)$ batch size in Qiu et al. (2020) for solving* (10) *assuming direct access to $y_i^*(x)$ will be $\mathcal{O}\left(n^4\mu_\lambda^{-2}\epsilon^{-2}\right)$ and $\mathcal{O}\left(n^5\mu_\lambda^{-4}\epsilon^{-2}\right)$* respectively. Our results in Theorem 6 indicate that the sample complexity of the proposed algorithm `MORMA-SOBA` for solving min-max bilevel problems has the same dependency on n and μ_λ as the sample complexity of the moving-average `SGDA` introduced in Qiu et al. (2020) for solving min-max single-level problems, while also computing $y_i^*(x)$ instead of assuming direct access.*

5. Experiments

While our contributions primarily focus on theoretical aspects, we also conducted experiments to validate our results. We first compare the performance of `MA-SOBA` with other benchmark methods on two common tasks proposed in previous works (Ji et al., 2021; Hong et al., 2023; Dagréou et al., 2022), *hyperparameter optimization* for ℓ^2 penalized logistic regression and *data hyper-cleaning* on the corrupted MNIST data set. To demonstrate the practical performance of `MORMA-SOBA`, we then conduct experiments in *robust multi-task representation learning* introduced in Gu et al. (2023) on the FashionMNIST data set (Xiao et al., 2017).

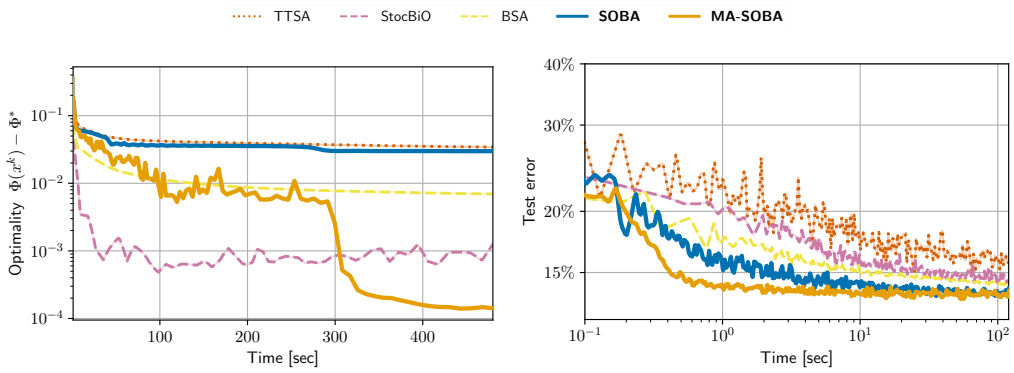
5.1 Experimental Details for MA-SOBA

Our experiments for `MA-SOBA` are performed with the aid of the recently developed package `Benchopt` (Moreau et al., 2022) and the open-sourced bilevel optimization benchmark†. For a fair comparison, we exclusively consider benchmark methods that do not utilize variance reduction techniques in Table 1: (i) `BSA` (Ghadimi and Wang, 2018); (ii) `stocBiO` (Ji et al., 2021); (iii) `TTSA` (Hong et al., 2023)/`ALSET` (Chen et al., 2021a); (iv) `SOBA` (Dagréou et al., 2022). Noting that `ALSET` only differs from `TTSA` regarding time scales, we use `TTSA` to represent this class of approach. Also, we omit the comparison with `AmIGO` (Arbel and Mairal, 2022) below, given that it is essentially a double-loop `SOBA` with increasing batch sizes. The tunable parameters in benchmark methods are selected in the same manner as those in `benchmark_bilevel`†.

Setup. We strictly adhere to the settings provided in `benchmark_bilevel`, as detailed in Appendix B.1 of Dagréou et al. (2022). The previous results and setups of Dagréou et al. (2022) have also been available in `https://benchopt.github.io/results/benchmark_bilevel.html`. For completeness, we provide a summary of the setup below.

*Note that $\Phi_{\mu_\lambda}(x,\lambda)$ in (9) is quadratic in λ, and these two sample complexities are obtained under this special case, i.e., $\nabla_2^2 f(x,y) = -\mu\mathbf{I}$ applied to Lin et al. (2020a); Qiu et al. (2020).

†`https://github.com/benchopt/benchmark_bilevel`



(a) ℓ^2 Penalized Logistic Regression on IJCNN1 (b) Data Hyper-Clearning on MNIST

Figure 1: Comparison of `MA-SOBA` with other stochastic bilevel optimization methods without using variance reduction techniques. For each algorithm, we plot the median performance over 10 runs. **Left:** Hyperparameter optimization for ℓ^2 penalized logistic regression on IJCNN1 data set. **Right.** Data hyper-cleaning on MNIST with $p = 0.5$ (corruption rate).

- To avoid redundant computations, we utilize oracles for functions F_ξ, G_ϕ, which provide access to quantities such as $\nabla_1 F_\xi(x,y)$, $\nabla_2 F_\xi(x,y)$, $\nabla_2 G_\phi(x,y)$, $\nabla^2_{22} G_\phi(x,y)v$, and $\nabla^2_{12} G_\phi(x,y)v$, although this may violate the independence assumption in Assumption 2.
- In all our experiments, we employ a batch size of 64 for all methods, even for `BSA` and `AmIGO` that theoretically require increasing batch sizes.
- For methods involving an inner loop (`stocBiO`, `BSA`, `AmIGO`), we perform 10 inner steps per each outer iteration as proposed in those papers.
- For methods that involve Neumann approximation for Hessian-vector product (such as `BSA`, `TTSA`, `SUSTAIN`, and `MRBO`), we perform 10 steps of the subroutine per outer iteration. For `AmIGO`, we perform 10 steps of SGD to approximate the inversion of the linear system.
- The step sizes and momentum parameters used in all benchmark algorithms are directly adopted from the fine-tuned parameters provided by Dagréou et al. (2022). From a grid search, we select the best constant step sizes for `MO-SOBA`.

We have excluded `SRBA` (Dagréou et al., 2023) from the benchmark due to its limited reported improvement over `SABA`.

5.1.1 Hyperparameter Optimization on IJCNN1

In the first task, we fit a multi-regularized logistic regression model (for binary classification), and select the regularization parameters (one hyperparameter per feature) on the IJCNN1 data set[‡]. The functions f and g of the problem (1) are the average logistic loss on the validation set and training set respectively, with ℓ^2 regularization for g. Specifically, the problem can be formulated as:

[‡] `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`

$$\min_{\nu\in\mathbb{R}^d} \quad \Phi(\nu) \coloneqq \underbrace{\mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}_{\text{val}}} \left[\ell\left(\langle \omega^*(\nu), X\rangle, Y\right)\right]}_{f(\nu,\omega^*(\nu))}$$

$$\text{s.t.} \quad \omega^*(\nu) = \arg\min_{\omega\in\mathbb{R}^d} \underbrace{\mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}_{\text{train}}} \left[\ell\left(\langle \omega, X\rangle, Y\right)\right] + \frac{1}{2}\omega^\top \operatorname{diag}\left(e^{\nu_1},\ldots,e^{\nu_d}\right)\omega}_{g(\nu,\omega)}.$$

In this case, $|\mathcal{D}_{\text{train}}| = 49,990$, $|\mathcal{D}_{\text{val}}| = 91,701$, and $d = 22$. For each sample, the covariate and label are denoted as (X, Y), where $X \in \mathbb{R}^{22}$ and $Y \in \{0, 1\}$. The inner variable ($\omega \in \mathbb{R}^{22}$) is the regression coefficient. The outer variable ($\nu \in \mathbb{R}^{22}$) is a vector of regularization parameters. The loss function $\ell(y', y) = -y\log(y') - (1-y)\log(1-y')$ is the log loss.

In Figure 1(a), we plot the suboptimality gap against the runtime for each method. Surprisingly, we observed that `MA-SOBA` achieves lower objective values after several iterations compared to all benchmark methods. This improvement can be attributed to the convergence of average hypergradients $\{h^k\}$. These findings demonstrate the practical superiority of our algorithm framework, even with the same sample complexity results.

To supplement the comparison, we conducted additional experiments that involved comparing all benchmark methods, including the variance reduction based method. In Figure 2, we plot the suboptimality gap $(\Phi(x) - \Phi^*)$ against runtime and the number of calls to oracles. Unfortunately, the previous results obtained for `MRBO` and `AmIGO` on the IJCNN1 data set are not reproducible at the moment due to some conflicts in the current developer version of `Benchopt`. As reported in Dagréou et al. (2022), `MRBO` exhibits similar performance to `SUSTAIN`, while the curve of `AmIGO` initially follows a similar trend as `SUSTAIN` and eventually reaches a similar level as `SABA` towards the end. Following a grid search, we have selected the parameters in `MA-SOBA` as $\alpha_k\tau = 0.02$, $\beta_k = \gamma_k = 0.01$, and $\theta_k = 0.1$. As shown in Figure 2, our proposed method `MA-SOBA` outperforms `SOBA` significantly, achieving a slightly lower suboptimality gap compared to the state-of-the-art variance reduction-based method `SABA`.

5.1.2 Data Hyper-Cleaning on MNIST

In the second task, we conduct data hyper-cleaning on the MNIST data set introduced in Franceschi et al. (2017). Data cleaning aims to train a multinomial logistic regression model on the corrupted training set and determine a weight for each training sample. These weights should approach zero for samples with corrupted labels. The data set is partitioned into a training set $\mathcal{D}_{\text{train}}$, a validation set $\mathcal{D}_{\text{val}}$, and a test set $\mathcal{D}_{\text{test}}$, where $|\mathcal{D}_{\text{train}}| = 20,000$, $|\mathcal{D}_{\text{val}}| = 5,000$, and $|\mathcal{D}_{\text{test}}| = 10,000$. Each sample is represented as a vector X of dimension 784, where the input image is flattened. The corresponding label takes values from the set $\{0, 1, \ldots, 9\}$. We use $Y \in \mathbb{R}^{10}$ to denote its one-hot encoding. Each sample in the training set is corrupted with probability p by replacing its label with a random label $\{0, 1, \ldots, 9\}$.

The task can be formulated into the bilevel optimization problem (1) with the inner variable y being the regression coefficients and the outer variable x being the sample weight. The `LL` function g is the sample-weighted cross-entropy loss on the corrupted training set with ℓ^2 regularization. The `UL` function f is the cross-entropy loss on the validation set.

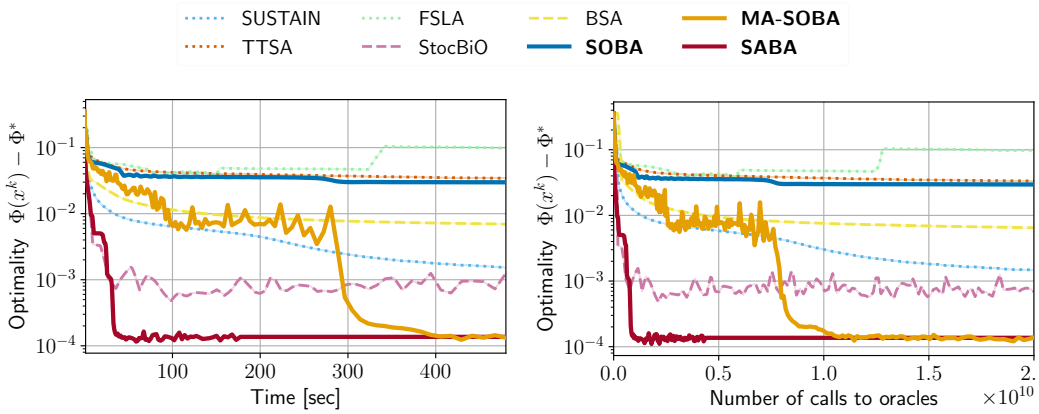


Figure 2: Comparison of `MA-SOBA` with other stochastic bilevel optimization methods in the problem of hyperparameter optimization for ℓ^2 regularized logistic regression on the IJCNN1 data set. We plot the median performance over 10 runs for each method. **Left:** Performance in runtime; **Right:** Performance in the number of gradient/Hessian(Jacobian)-vector products sampled.

Precisely, the task can be formulated into the bilevel optimization problem as below:

$$\min_{\nu \in \mathbb{R}^{|\mathcal{D}_{\text{train}}|}} \quad \Phi(\nu) := \underbrace{\mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}_{\text{val}}} [\ell(W^*(\nu)X, Y)]}_{f(\nu, W^*(\nu))}$$

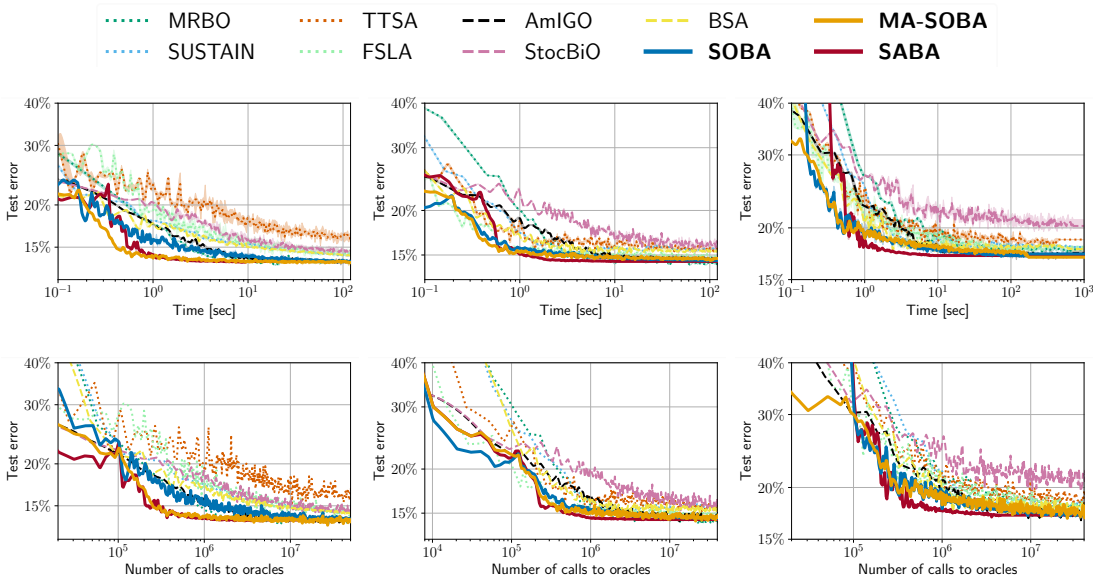
$$\text{s.t.} \quad W^*(\nu) = \arg\min_{\omega \in \mathbb{R}^d} \underbrace{\frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(X_i,Y_i)\sim\mathcal{D}_{\text{train}}} \sigma(\nu_i)\ell(WX_i, \overbrace{\widetilde{Y}_i}^{\text{corrupted}}) + C_r\|W\|^2}_{g(\nu, W)},$$

where the outer variable ($\nu \in \mathbb{R}^{20{,}000}$) is a vector of sample weights for the training set, the inner variable $W \in \mathbb{R}^{10\times 784}$, and ℓ is the cross entropy loss and σ is the sigmoid function. The regularization parameter $C_r = 0.2$ following Dagréou et al. (2022). The objective of data hyper-cleaning is to train a multinomial logistic regression model on the training set and determine a weight for each training sample using the validation set. The weights are designed to approach zero for corrupted samples, thereby aiding in the removal of these samples during the training process.

We report the test error in Figure 1(b). We observe that `MA-SOBA` outperforms other benchmark methods by achieving lower test errors faster.

To supplement the comparison, we conducted additional experiments that involved comparing all benchmark methods, including the variance reduction-based method. Following a grid search, we have selected the parameters in `MA-SOBA` as $\alpha_k\tau = 10^3$, $\beta_k = \gamma_k = 10^{-2}$, and $\theta_k = 10^{-1}$. In Figure 3, we plot the test error against runtime and the number of calls to oracles with different corruption probability $p \in \{0.5, 0.7, 0.9\}$. We observe that `MA-SOBA`

has comparable performance to the state-of-the-art method `SABA`. Remarkably, `MA-SOBA` is the fastest algorithm to reach the best test accuracy when $p = 0.5$.



(a) $p = 0.5$ (b) $p = 0.7$ (c) $p = 0.9$

Figure 3: Comparison of `MA-SOBA` with other stochastic bilevel optimization methods in the problem of data hyper-cleaning on the MNIST data set when the corruption probability $p \in \{0.5, 0.7, 0.9\}$. We plot the median performance over 10 runs for each method. **Top:** Performance in runtime; **Bottom:** Performance in the number of gradient/Hessian(Jacobian)-vector products sampled.

5.2 Experimental Details for MORMA-SOBA

To demonstrate the practical performance of `MORMA-SOBA` as compared to `MORBiT` (Gu et al., 2023), we conduct experiments in *robust multi-task representation learning* introduced in Gu et al. (2023) on the FashionMNIST data set (Xiao et al., 2017). We adopt the same setup as described in Gu et al. (2023), which can be summarized as follows.

Setup. We consider binary classification tasks generated from FashionMNIST where we select 8 "easy" tasks (lowest loss ~ 0.3 from independent training) and 2 "hard" tasks (lowest loss ~ 0.45 from independent training) for multi-objective robust representation learning:

- "easy" tasks: (0, 9), (1, 7), (2, 7), (2, 9), (4, 7), (4, 9), (3, 7), (3, 9)
- "hard" tasks: (0, 6), (2, 4)

For each task $i \in [10]$ above, we partition its data set into the training set $\mathcal{D}_i^{\text{train}}$, validation set $\mathcal{D}_i^{\text{val}}$, and test set $\mathcal{D}_i^{\text{test}}$. We also generate 7 (unseen) binary classification tasks for testing:

- "easy" tasks: (1, 9), (2, 5), (4, 5), (5, 6)
- "hard" tasks: (2, 6), (3, 6), (4, 6)

We train a shared representation network that maps the 784-dimensional (vectorized 28x28 images) input to a 100-dimensional space. To learn a shared representation and per-task models that generalize well on each task, we aim to solve the following problem:

$$\min_{E\in\mathbb{R}^{100\times 784}} \max_{1\le i\le n} \Phi_i(E) := \underbrace{\mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}_i^{\text{val}}}\left[\ell\left(W_i^*(E)\circ\overbrace{\text{ReLU}(EX)}^{\text{representation}}+b_i^*(E), Y\right)\right]}_{f_i(E,(W_i^*,b_i^*))}$$

$$\text{s.t.} \quad \begin{pmatrix} W_i^*(E) \\ b_i^*(E)\end{pmatrix} =$$

$$\underbrace{\mathop{\arg\min}_{W_i\in\mathbb{R}^{10\times 100}, b_i\in\mathbb{R}^{10}} \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}_i^{\text{train}}}\left[\ell\left(\overbrace{W_i}^{\text{weight}}\circ\text{ReLU}(EX)+\overbrace{b_i}^{\text{bias}}, Y\right)\right]+\rho\|W_i\|_F^2}_{g_i(E,(W_i,b_i))}, 1\le i\le n.$$

Each bilevel objective Φ_i in this setup represents a distinct binary classification "task" $i\in[n]$ with its own training and validation sets. The optimization variable is engaged in a shared representation network, parameterized by the outer variable $E\in\mathbb{R}^{100\times 784}$, along with per-task linear models parameterized by each inner variable (W_i, b_i). The `UL` function f_i is the average cross-entropy loss over the $\mathcal{D}_i^{\text{val}}$, and the `LL` function g_i is the ℓ^2 regularized cross-entropy loss over $\mathcal{D}_i^{\text{train}}$. Each sample is represented as a vector X of dimension 784, where the input image is flattened. The corresponding label takes values from the set $\{0, 1, \ldots, 9\}$. We use $Y\in\mathbb{R}^{10}$ to denote its one-hot encoding.

In the experiment, the regularization parameter in the `LL` function $\rho = 5\times 10^{-4}$. The implementation of `MORBiT` follows the same manner described in Gu et al. (2023). Specifically, the code of `MORBiT` (Gu et al., 2023) uses vanilla SGD with a learning rate scheduler and incorporates momentum and weight decay techniques to optimize each variable:

- Outer variable: learning rate = 0.01, momentum = 0.9, weight_decay = 10^{-4}
- Inner variable: learning rate = 0.01, momentum = 0.9, weight_decay = 10^{-4}
- Simplex variable: learning rate = 0.3, momentum = 0.9, weight_decay = 10^{-4}

In addition, `MORBiT` adopts a straightforward iterative auto-differentiation to calculate the hypergradient without using Neumann approximation of the Hession inversion.

For the implementation of `MORMA-SOBA`, the regularization parameter μ_λ in 10 is set to be 0.01. All remaining parameters are chosen as constant values, as listed below:

- Outer variable: $\tau_x = 1, \alpha_k = 0.02$,
- Inner variable: $\beta_k = 0.02$
- Auxiliary variable: $\gamma_k = 0.02$
- Simplex variable: $\tau_\lambda = 1, \alpha_k = 0.02$
- Average gradient: $\theta_k = 0.6$

Both evaluated methods use batch sizes of 8 and 128 to compute g_i for each inner step and f_i for each outer iteration, respectively.

In Figure 4, we compare our algorithm with the existing min-max bilevel algorithm `MORBiT` (Gu et al., 2023) in terms of the average loss ($(1/n)\sum_i \Phi_i$) and maximum loss ($\max_i \Phi_i$). The results demonstrate the superiority of `MORMA-SOBA` over `MORBiT` in terms of lowering both the max loss and average loss at a faster rate. In addition to Figure 4, which showcases the performance on 10 seen tasks used for representation learning, we present Figure 5. This figure displays the maximum/average loss values against the number of iterations on test sets consisting of 10 seen tasks and 7 unseen tasks. Our approach, `MORMA-SOBA`, demonstrates superior performance in terms of faster reduction of both maximum and average loss.

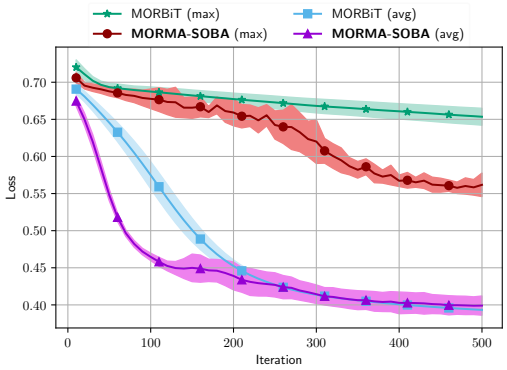


Figure 4: `MORMA-SOBA` ($\mu_\lambda = 0.01$) vs. `MORBiT` on robust multi-task representation learning.

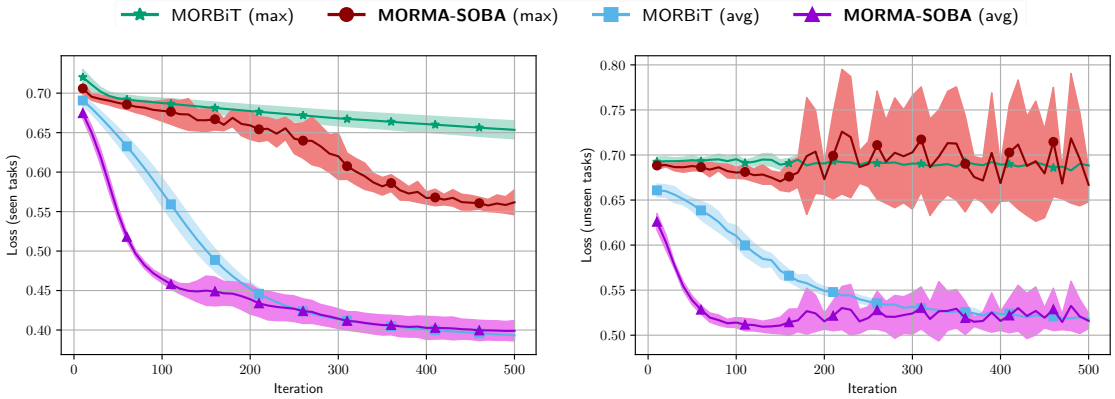


Figure 5: Comparison of `MORMA-SOBA` with `MORBiT` in the problem of multi-objective robust representation learning for binary classification tasks on the FashionMNIST data set. We aggregate the results over 10 runs for each method. **Left:** Performance on test sets of seen tasks; **Right:** Performance on unseen tasks.

5.3 Moving Average vs. Variance Reduction

Through empirical studies, we have demonstrated that our proposed method, `MA-SOBA`, achieves comparable performance to the state-of-the-art variance reduction-based approach `SABA` using `SAGA` updates (Defazio et al., 2014). In this context, we would like to highlight the key difference and relationship between these two methods.

We start with presenting the update rules of the sequence of estimated gradients $\{g^k\}$ for the variance reduction techniques `SAGA` (Defazio et al., 2014) and our moving-average method (`MA`) for the single-level problem:

SAGA (finite-sum:) $\min\ \frac{1}{n}\sum_{i=1}^{n} f_i(x)$

$$g^k = \nabla f_{i_k}(x^k) - \nabla f_{i_k}(\bar{x}_{i_k}) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(\bar{x}_j)$$

The `SAGA` update is designed for finite-sum problems with offline batch data. At each iteration k, the algorithm randomly selects an index $i_k \in [n]$ and updates the gradient variable g^k using a reference point $\bar{x}_{i_k}$, which corresponds to the last evaluated point for ∇f_{i_k}. However, it should be noted that `SAGA` requires storing the previously evaluated gradients $\nabla f_j(\bar{x}_j)$ in a table, which can be memory-intensive when sample size n or dimension d is large. In the finite-sum setting, there exist several other variance reduction methods, such as `SARAH` (Dagréou et al., 2023), that can be employed to further enhance the dependence on the number of samples, n, for bilevel optimization problems. However, the `SARAH`-type method requires double gradient evaluations on each iteration of x^k and x^{k-1}:

MA (expectation): $\min\ \mathbb{E}_\xi[f(x;\xi)]$

$$g^k = (1-\alpha_k)g^{k-1} + \alpha_k \nabla f(x^k;\xi^{k+1})$$

Unlike variance reduction techniques, the moving-average methods can solve the general expectation-form problem with online and streaming data using a simple update per iteration. In addition, the moving-average techniques offer two more advantages:

Theoretical Assumption. All variance reduction methods, including `SVRG` (Reddi et al., 2016), `SAGA` (Defazio et al., 2014), `SARAH` (Nguyen et al., 2017), `STORM` (Cutkosky and Orabona, 2019), and others, typically rely on assuming mean-squared smoothness assumptions. In particular, for stochastic optimization problems in the form of $\min_x\{f(x) = \mathbb{E}[F(x,\xi)]\}$, the definition of mean-squared smoothness (MSS) is: (MSS) $\mathbb{E}_\xi[\|\nabla F(x,\xi) - \nabla F(x',\xi)\|^2] \le L^2\|x-x'\|^2$. However, MSS is a stronger assumption than the general smoothness assumption on f: $\|\nabla f(x) - \nabla f(x')\| \le L\|x-x'\|$. By Jensen's inequality, we have that MSS is stronger than the general smoothness assumption on f: $\|\nabla f(x) - \nabla f(x')\|^2 \le \mathbb{E}_\xi[\|\nabla F(x,\xi) - \nabla F(x',\xi)\|^2]$. In this work, the theoretical results of the proposed methods are only built on the smoothness assumption on the `UL` and `LL` functions f, g without further assuming MSS on F_ξ and G_ϕ. It is worth noting that a clear distinction in the lower bounds of sample complexity for solving the single-level stochastic optimization has been proven in Arjevani et al. (2023). Specifically, they establish a separation under the MSS assumption on F_ξ and smoothness assumptions on f ($\mathcal{O}(\epsilon^{-1.5})$ vs. $\mathcal{O}(\epsilon^{-2})$). Thus, it is important to emphasize that `MA-SOBA` achieves the optimal sample complexity $\mathcal{O}(\epsilon^{-2})$ under our weaker assumptions.

Practical Implementation. Variance reduction methods often entail additional space complexity, require double-loop implementation or double oracle computations per iteration. These requirements can be unfavorable for large-scale problems with limited computing resources. For instance, in the second task, the runtime improvement achieved by using `SABA` is limited. This limitation can be attributed to the dimensionality of the variables ν (with a dimension of $20,000$) and W (with a dimension of 10×784). The benefit of using variance reduction methods is expected to be less significant for more complex problems involving computationally expensive oracle evaluations.

6. Conclusion

In this work, we propose a novel class of algorithms (`MA-SOBA`) for solving stochastic bilevel optimization problems in (1) by introducing the moving-average step to estimate the hypergradient. We present a refined convergence analysis of our algorithm, achieving the optimal sample complexity without relying on the high-order smoothness assumptions employed in the literature. Furthermore, we extend our algorithm framework to tackle a generic min-max bilevel optimization problem within the multi-objective setting, identifying and addressing the theoretical gap present in the literature.

Acknowledgments

We thank the authors of Gu et al. (2023) for clarifications regarding their paper. XC was supported in part by National Science Foundation (NSF) grants DMS-2053918 and CCF-1934568. KB was supported by NSF grant DMS-2053918.

7. Proofs

We will prove Theorems 3 and 6 in Section 7.1 and 7.2 respectively. In each section we will first establish the relations between the optimality measure (see V_k in Section 3.3) and the gradient mapping, which reduce the proof of main theorems to proving the convergence of primal variables (x^k in Theorem 3 or (x^k, λ^k) in Theorem 6) and dual variables (h^k in Theorem 3 or (h_x^k, h_λ^k) in Theorem 6). Then we will prove the hypergradient estimation error, primal convergence and dual convergence separately. In our notation convention, the superscript k usually denotes the iteration number and the subscript i represents variables related to functions f_i, g_i. $L_\#$ with being a function $\#$ denotes its Lipschitz constant.

Next we state some technical lemmas that will be used in both sections.

Lemma 9 (Lemma 10 in Qu and Li (2017).) *Suppose $f(x)$ is μ-strongly convex and L-smooth. For any x and $\gamma < \frac{2}{\mu+L}$, define $x^+ = x - \gamma\nabla f(x)$, $x^* = \arg\min f(x)$. Then we have $\|x^+ - x^*\| \le (1-\gamma\mu)\|x - x^*\|$.*

Lemma 10 *Define $\kappa = \max(L_{\nabla f}, L_{\nabla g})/\mu_g$, $z^*(x) = \left(\nabla_{22}^2 g(x, y^*(x))\right)^{-1} \nabla_2 f(x, y^*(x))$. Suppose Assumption 1 holds. Then $\Phi(x)$ is differentiable and $\nabla\Phi(x)$ is given by Then $\Phi(x), y^*(x)$, $z^*(x)$ are differentiable and $\nabla\Phi(x), y^*(x), z^*(x)$ are $L_{\nabla\Phi}, L_{y^*}, L_{z^*}$-Lipschitz continuous, and*

$$\nabla\Phi(x) = \nabla_1 f(x, y^*(x)) - \nabla_{12}^2 g(x, y^*(x)) \left(\nabla_{22}^2 g(x, y^*(x))\right)^{-1} \nabla_2 f(x, y^*(x)), \tag{11}$$

$$\nabla y^*(x) = -\nabla_{12}^2 g(x, y^*(x)) \left(\nabla_{22}^2 g(x, y^*(x))\right)^{-1}. \tag{12}$$

The constants are given by

$$L_{y^*} = \frac{L_{\nabla g}}{\mu_g} = \mathcal{O}(\kappa), \; L_{z^*} = \sqrt{1 + L_{y^*}^2}\left(\frac{L_{\nabla f}}{\mu_g} + \frac{L_f L_{\nabla_{22}^2 g}}{\mu_g^2}\right) = \mathcal{O}\left(\kappa^3\right),$$

$$L_{\nabla\Phi} = L_{\nabla f} + \frac{2L_{\nabla f}L_{\nabla g} + L_f^2 L_{\nabla^2 g}}{\mu_g} + \frac{2L_f L_{\nabla g} L_{\nabla^2 g} + L_{\nabla f} L_{\nabla g}^2}{\mu_g^2} + \frac{L_f L_{\nabla^2 g} L_{\nabla g}^2}{\mu_g^3} = \mathcal{O}\left(\kappa^3\right).$$

Moreover, we have

$$\|z^*(x)\| \leq \frac{L_f}{\mu_g}. \tag{13}$$

Proof See Lemma 2.2 in Ghadimi and Wang (2018) for the proof of (11) and (12), Lipschitz continuity of $\nabla\Phi$ and y^*. For the Lipschitz continuity of z^* we have for any $x, \tilde{x}$, we know

$$\begin{aligned}
&\|z^*(x) - z^*(\tilde{x})\| \\
=&\| \left(\nabla^2_{22} g(x, y^*(x))\right)^{-1} \nabla_2 f(x, y^*(x)) - \left(\nabla^2_{22} g(\tilde{x}, y^*(\tilde{x}))\right)^{-1} \nabla_2 f(\tilde{x}, y^*(\tilde{x}))\| \\
\leq&\| \left(\nabla^2_{22} g(x, y^*(x))\right)^{-1} \nabla_2 f(x, y^*(x)) - \left(\nabla^2_{22} g(\tilde{x}, y^*(\tilde{x}))\right)^{-1} \nabla_2 f(x, y^*(x))\| \\
&+ \| \left(\nabla^2_{22} g(\tilde{x}, y^*(\tilde{x}))\right)^{-1} \nabla_2 f(x, y^*(x)) - \left(\nabla^2_{22} g(\tilde{x}, y^*(\tilde{x}))\right)^{-1} \nabla_2 f(\tilde{x}, y^*(\tilde{x}))\| \\
\leq& L_f \| \left(\nabla^2_{22} g(x, y^*(x))\right)^{-1} \| \|\nabla^2_{22} g(x, y^*(x)) - \nabla^2_{22} g(\tilde{x}, y^*(\tilde{x}))\| \| \left(\nabla^2_{22} g(x, y^*(x))\right)^{-1} \| \\
&+ \frac{1}{\mu_g} \|\nabla_2 f(x, y^*(x)) - \nabla_2 f(\tilde{x}, y^*(\tilde{x}))\| \\
\leq& \frac{L_f L_{\nabla^2_{22} g}}{\mu_g^2} \sqrt{\|x - \tilde{x}\|^2 + \|y^*(x) - y^*(\tilde{x})\|^2} + \frac{L_{\nabla f}}{\mu_g} \sqrt{\|x - \tilde{x}\|^2 + \|y^*(x) - y^*(\tilde{x})\|^2} \\
\leq& L_{z^*} \|x - \tilde{x}\|,
\end{aligned}$$

where the first inequality uses triangle inequality, the second and third inequalities use Assumption 1, and the fourth inequality uses Lipschitz continuity of $y^*(x)$. The inequality in (13) holds since $g(x, \cdot)$ is μ_g-strongly convex and $\|\nabla_2 f(x, y^*(x))\| \leq L_f$ (Assumption 1). ■

Lemma 11 (Lemma 3.2 in Ghadimi et al. (2020)) *For any closed convex set $\mathcal{X}$, and the function $\eta_{\mathcal{X}}(x, h, \tau)$ defined in Section 3.3 is differentiable and $\nabla\eta_{\mathcal{X}}$ is $L_{\nabla\eta_{\mathcal{X}}}$-Lipschitz continuous, with the closed form expression and constant given by*

$$\nabla_1 \eta_{\mathcal{X}}(x, h, \tau) = -h + \frac{1}{\tau}(x - \bar{d}), \nabla_2 \eta_{\mathcal{X}}(x, h, \tau) = \bar{d} - x, L_{\nabla\eta_{\mathcal{X}}} = 2\sqrt{(1 + 1/\tau)^2 + (1 + \tau/2)^2},$$

where $\bar{d}$ is defined as $\bar{d} = \arg\min_{d \in \mathcal{X}} \{\langle h, d - x\rangle + \frac{1}{2\tau}\|d - x\|^2\} = \Pi_{\mathcal{X}}(x - \tau h)$, which satisfies

$$\left\langle h + \frac{1}{\tau}(\bar{d} - x), d - \bar{d} \right\rangle \geq 0, \quad \textit{for all } d \in \mathcal{X}. \tag{14}$$

7.1 Proof of Theorem 3

For simplicity, we summarize the notations that will be used in Section 7.1 as follows.

$$\begin{aligned}
&\kappa = \max(L_{\nabla f}, L_{\nabla g})/\mu_g, \ w^{k+1} = u_x^{k+1} - J^{k+1} z^k, \\
&y_*^k = y^*(x^k) = \arg\min_{y \in \mathbb{R}^{d_y}} g(x^k, y), \ z_*^k = \left(\nabla^2_{22} g(x^k, y_*^k)\right)^{-1} \nabla_2 f(x^k, y_*^k), \\
&\Phi(x) = f(x, y^*(x)), \ \eta_{\mathcal{X}}(x, h, \tau) = \min_{d \in X} \left\{ \langle h, d - x\rangle + \frac{1}{2\tau}\|d - x\|^2 \right\}.
\end{aligned} \tag{15}$$

In this section we suppose Assumptions 1 and 2 hold. We assume stepsizes in Algorithm 1 satisfy $\beta_k = c_1\alpha_k,\ \gamma_k = c_2\alpha_k, \theta_k = c_3\alpha_k$, where $c_1, c_2, c_3 > 0$ are constants to be determined. We will utilize the following merit function in our analysis:

$$W_k = \underbrace{\Phi(x^k) - \inf_{x\in\mathcal{X}} \Phi(x) - \frac{1}{c_3}\eta_{\mathcal{X}}(x^k, h^k, \tau)}_{W_{k,1}} + \underbrace{\frac{1}{c_1}\|y^k - y_*^k\|^2 + \frac{1}{c_2}\|z^k - z_*^k\|^2}_{W_{k,2}}.$$

By definition of $\eta_{\mathcal{X}}$, we can verify that $W_{k,1} \geq 0$. Moreover, as discussed in Section 3.3, we consider the following optimality measure:

$$V_k = \frac{1}{\tau^2}\|x_+^k - x^k\|^2 + \|h^k - \nabla\Phi(x^k)\|^2. \tag{16}$$

Next we characterize the relation between V_k and gradient mapping of problem 1.

Lemma 12 *Suppose Assumptions 1 and 2 hold. In Algorithm 1 we have*

$$\frac{1}{\tau^2}\|x^k - \Pi_{\mathcal{X}}\big(x^k - \tau\nabla\Phi(x^k)\big)\|^2 \leq 2V_k.$$

Proof Note that we have

$$\begin{aligned}\|x^k - \Pi_{\mathcal{X}}\big(x^k - \tau\nabla\Phi(x^k)\big)\|^2 \leq& 2\big(\|x_+^k - x^k\|^2 + \|\Pi_{\mathcal{X}}\big(x^k - \tau h^k\big) - \Pi_{\mathcal{X}}\big(x^k - \tau\nabla\Phi(x^k)\big)\|^2\big)\\ \leq& 2\big(\|x_+^k - x^k\|^2 + \tau^2\|h^k - \nabla\Phi(x^k)\|^2\big) = 2\tau^2 V_k,\end{aligned}$$

where the first inequality uses Cauchy-Schwarz inequality and the second inequality uses the non-expansiveness of projection onto a closed convex set. This completes the proof. ■

Then we bound the variance of w^{k+1} and $\|h^{k+1} - h^k\|$.

Lemma 13 *Suppose Assumptions 1 and 2 hold. In Algorithm 1 we have*

$$\begin{aligned}&\mathbb{E}\left[\|w^{k+1} - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2\right] \leq \sigma_{w,k+1}^2\\ &\sigma_{w,k+1}^2 := \sigma_w^2 + 2\sigma_{g,2}^2\mathbb{E}[\|z^k - z_*^k\|^2],\ \sigma_w^2 = \sigma_{f,1}^2 + \frac{2\sigma_{g,2}^2 L_f^2}{\mu_g^2}, \qquad (17)\\ &\mathbb{E}[\|h^{k+1} - h^k\|^2] \leq \sigma_{h,k}^2,\\ &\sigma_{h,k}^2 := 2\theta_k^2\mathbb{E}\left[\|h^k - \nabla\Phi(x^k)\|^2 + \|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Phi(x^k)\|^2\right] + \theta_k^2\sigma_{w,k+1}^2. \qquad (18)\end{aligned}$$

Proof We first consider w^k. Note that

$$w^{k+1} - \mathbb{E}[w^{k+1}|\mathscr{F}_k] = u_x^{k+1} - \mathbb{E}[u_x^{k+1}|\mathscr{F}_k] - \left(J^{k+1} - \mathbb{E}\left[J^{k+1}|\mathscr{F}_k\right]\right)z^k.$$

Hence we know

$$\mathbb{E}\left[\|w^{k+1} - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k\right]$$

$$
\begin{aligned}
=&\mathbb{E}\left[\|u_x^{k+1}-\mathbb{E}[u_x^{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k\right]+\mathbb{E}\left[\|J^{k+1}-\mathbb{E}\left[J^{k+1}|\mathscr{F}_k\right]\|^2|\mathscr{F}_k\right]\|z^k\|^2\\
\le&\sigma_{f,1}^2+2\sigma_{g,2}^2\|z_*^k\|^2+2\sigma_{g,2}^2\|z^k-z_*^k\|^2\le\sigma_{f,1}^2+\frac{2\sigma_{g,2}^2L_f^2}{\mu_g^2}+2\sigma_{g,2}^2\|z^k-z_*^k\|^2,
\end{aligned}
$$

where the first equality uses independence, the first inequality uses Cauchy-Schwarz inequality, and the second inequality uses (13). This proves (17). Next for $\|h^{k+1}-h^k\|$ we have

$$
\begin{aligned}
&\mathbb{E}\left[\|h^{k+1}-h^k\|^2|\mathscr{F}_k\right]\\
=&\theta_k^2\mathbb{E}\left[\|h^k-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k\right]+\theta_k^2\mathbb{E}\left[\|w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k\right]\\
\le&2\theta_k^2\mathbb{E}\left[\|h^k-\nabla\Phi(x^k)\|^2|\mathscr{F}_k\right]+2\theta_k^2\mathbb{E}\left[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\|^2|\mathscr{F}_k\right]+\theta_k^2\sigma_{w,k+1}^2,
\end{aligned}
$$

which proves of (18) by taking expectation on both sides. ■

Remark 14 *We would like to highlight that in* (17)*, we explicitly characterize the upper bound of the variance of w^{k+1}, which contains $\mathbb{E}\left[\|z^k-z_*^k\|^2\right]$ and requires further analysis. In contrast, Assumption 3.7 in Dagréou et al. (2022) directly assumes the second moment of D_x^t is uniformly bounded, i.e., $\mathbb{E}\left[\|D_x^t\|^2\right]\le B_x^2$ for some constant $B_x\ge 0$. Note that D_x^t in Dagréou et al. (2022) is the same as our w^{k+1} (see* (6)*, line 5 of Algorithm 1 and definition of w^{k+1} in* (15)*). The second moment bound can directly imply the variance bound, i.e., $\mathbb{E}\left[\|D_x^t-\mathbb{E}\left[D_x^t\right]\|^2\right]\le\mathbb{E}\left[\|D_x^t\|^2\right]\le B_x^2$. This implies that some stronger assumptions are needed to guarantee Assumption 3.7 in Dagréou et al. (2022), as also pointed out by the authors (see discussions right below it). Instead, our refined analysis does not require that.*

7.1.1 Hypergradient Estimation Error

Note that Assumptions 3.1 and 3.2 in Dagréou et al. (2022) state that the upper-level function f is twice differentiable, the lower-level function g is three times differentiable and $\nabla^2 f, \nabla^3 g$ are Lipschitz continuous so that z_*^k, as a function of x^k (see (15)), is smooth, which is a crucial condition for (31) and (81) in Dagréou et al. (2022) ($v^*(x^t)$ in their notation), which follows the analysis in Equation (49) in Chen et al. (2021a). In this section we show that, by incorporating the moving-average technique recently introduced to decentralized bilevel optimization (Chen et al., 2023b), we can remove this additional assumption. We have the following lemma characterizing the error induced by y^k and z^k.

Lemma 15 *Suppose Assumptions 1 and 2 hold. If the stepsizes satisfy*

$$
\beta_k<\frac{2}{\mu_g+L_{\nabla g}},\ \gamma_k\le\min\left(\frac{1}{4\mu_g},\ \frac{0.06\mu_g}{\sigma_{g,2}^2}\right),\tag{19}
$$

then in Algorithm 1 we have

$$
\begin{aligned}
\sum_{k=0}^K\alpha_k\mathbb{E}[\|y^k-y_*^k\|^2]&\le C_{yx}\sum_{k=0}^K\alpha_k\mathbb{E}[\|x_+^k-x^k\|^2]+C_{y,0}+C_{y,1}\left(\sum_{k=0}^K\alpha_k^2\right)\\
\sum_{k=0}^K\alpha_k\mathbb{E}[\|z^k-z_*^k\|^2]&\le C_{zx}\sum_{k=0}^K\alpha_k\mathbb{E}[\|x_+^k-x^k\|^2]+C_{z,0}+C_{z,1}\left(\sum_{k=0}^K\alpha_k^2\right).
\end{aligned}\tag{20}
$$

where the constants are defined as

$$C_{yx} = \frac{2L_{y^*}^2}{c_1^2\mu_g^2},\ C_{y,0} = \frac{1}{c_1\mu_g}\mathbb{E}\left[\|y^0 - y_*^0\|^2\right],\ C_{y,1} = \frac{2c_1\sigma_{g,1}^2}{\mu_g},$$
$$C_{zx} = \frac{5L_f^2}{\mu_g^2}\left(\frac{L_{\nabla_{22}^2 g}^2}{\mu_g^2}+1\right)\frac{2L_{y^*}^2}{c_1^2\mu_g^2} + \frac{4L_{z^*}^2}{c_2^2\mu_g^2},$$
$$C_{z,0} = \frac{5L_f^2}{\mu_g^2}\left(\frac{L_{\nabla_{22}^2 g}^2}{\mu_g^2}+1\right)\cdot\frac{1}{c_1\mu_g}\mathbb{E}\left[\|y^0 - y_*^0\|^2\right] + \frac{1}{c_2\mu_g}\mathbb{E}\left[\|z^0 - z_*^0\|^2\right],$$
$$C_{z,1} = \frac{5L_f^2}{\mu_g^2}\left(\frac{L_{\nabla_{22}^2 g}^2}{\mu_g^2}+1\right)\cdot\frac{2c_1\sigma_{g,1}^2}{\mu_g} + \frac{2c_2\sigma_w^2}{\mu_g}.$$

Proof We first consider the error induced by y^k. We have

$$\begin{aligned}\|y^{k+1} - y_*^{k+1}\|^2 &\le (1+\beta_k\mu_g)\|y^{k+1} - y_*^k\|^2 + \left(1+\frac{1}{\beta_k\mu_g}\right)\|y_*^{k+1} - y_*^k\|^2 \\ &\le (1+\beta_k\mu_g)\|y^{k+1} - y_*^k\|^2 + \left(\frac{\alpha_k^2}{\beta_k\mu_g}+\alpha_k^2\right)L_{y^*}^2\|x_+^k - x^k\|^2, \qquad (21)\end{aligned}$$

where the first inequality uses Cauchy-Schwarz inequality: $\|u+v\|^2 \le (1+c)\big(\|u\|^2 + \frac{1}{c}\|v\|^2\big)$, for any vectors u, v and constant $c > 0$. Thanks to the moving-average step of x^k, our analysis of $\|y_*^{k+1} - y_*^k\|$ is simplified comparing to that in Chen et al. (2021a). Also,

$$\begin{aligned}&\mathbb{E}\left[\|y^{k+1} - y_*^k\|^2|\mathscr{F}_k\right] = \mathbb{E}\left[\|y^k - \beta_k\nabla_2 g(x^k,y^k) - y_*^k - \beta_k(v^{k+1} - \nabla_2 g(x^k,y^k))\|^2|\mathscr{F}_k\right] \\ \le&\|y^k - \beta_k\nabla_2 g(x^k,y^k) - y_*^k\|^2 + \beta_k^2\sigma_{g,1}^2 \le (1-\beta_k\mu_g)^2\|y^k - y_*^k\|^2 + \beta_k^2\sigma_{g,1}^2, \qquad (22)\end{aligned}$$

where the first inequality uses Assumption (2) and Lemma 9, and the second inequality uses Lemma 9 (which requires strong convexity of g, Lipschtiz continuity of $\nabla_2 g$, and the first inequality in (19)). Combining (21) and (22), we know

$$\begin{aligned}&\mathbb{E}\left[\|y^{k+1} - y_*^{k+1}\|^2|\mathscr{F}_k\right] \\ \le&(1+\beta_k\mu_g)(1-\beta_k\mu_g)^2\|y^k - y_*^k\|^2 + \left(\frac{\alpha_k^2}{\beta_k\mu_g}+\alpha_k^2\right)L_{y^*}^2\|x_+^k - x^k\|^2 + (1+\beta_k\mu_g)\beta_k^2\sigma_{g,1}^2 \\ \le&(1-\beta_k\mu_g)\|y^k - y_*^k\|^2 + \frac{2\alpha_k^2 L_{y^*}^2}{\beta_k\mu_g}\|x_+^k - x^k\|^2 + 2\beta_k^2\sigma_{g,1}^2.\end{aligned}$$

where the second inequality uses $\beta_k < \frac{2}{\mu_g + L_{\nabla g}} \le \frac{1}{\mu_g}$. Taking summation ($k$ from 0 to K) on both sides and taking expectation, we know

$$\sum_{k=0}^K \beta_k\mu_g\mathbb{E}[\|y^k - y_*^k\|^2] \le \mathbb{E}\left[\|y^0 - y_*^0\|^2\right] + \sum_{k=0}^K \frac{2\alpha_k^2 L_{y^*}^2}{\beta_k\mu_g}\mathbb{E}[\|x_+^k - x^k\|^2] + \sum_{k=0}^K 2\beta_k^2\sigma_{g,1}^2,$$

which proves the first inequality in (20) by dividing $c_1\mu_g$ on both sides. Next we analyze the error induced by z^k. Our analysis is substantially different from Dagréou et al. (2022):

$$\|z^{k+1} - z_*^{k+1}\|^2 \le \left(1+\frac{\gamma_k\mu_g}{3}\right)\|z^{k+1} - z_*^k\|^2 + \left(1+\frac{3}{\gamma_k\mu_g}\right)\|z_*^{k+1} - z_*^k\|^2$$

$$\leq \left(1+\frac{\gamma_k\mu_g}{3}\right)\|z^{k+1}-z_*^k\|^2+\left(\frac{3\alpha_k^2}{\gamma_k\mu_g}+\alpha_k^2\right)L_{z^*}^2\|x_+^k-x^k\|^2 \tag{23}$$

where we use Cauchy-Schwarz inequality in the first and second inequality, we use the facts that ∇y^* is Lipschitz continuous. For $\|z^{k+1}-z_*^k\|$, we may follow the analysis of SGD under the strongly convex setting:

$$\begin{aligned}z^{k+1}-z_*^k=z^k-\gamma_k(H^kz^k-u_y^k)-z_*^k=z^k-\gamma_k\nabla_{22}^2g(x^k,y^k)z^k+\gamma_k\nabla_2f(x^k,y^k)-z_*^k\\-\gamma_k(H^{k+1}-\nabla_{22}^2g(x^k,y^k))z^k+\gamma_k(u_y^k-\nabla_2f(x^k,y^k))\end{aligned}$$

which gives

$$\begin{aligned}&\mathbb{E}\left[\|z^{k+1}-z_*^k\|^2|\mathscr{F}_k\right]\\
\leq&\|z^k-\gamma_k\nabla_{22}^2g(x^k,y^k)z^k+\gamma_k\nabla_2f(x^k,y^k)-z_*^k\|^2+\gamma_k^2\sigma_{g,2}^2\|z^k\|^2+\gamma_k^2\sigma_{f,1}^2\\
=&\|(I-\gamma_k\nabla_{22}^2g(x^k,y^k))(z^k-z_*^k)-\gamma_k(\nabla_{22}^2g(x^k,y^k)z_*^k-\nabla_2f(x^k,y^k))\|^2+\gamma_k^2\sigma_{g,2}^2\|z^k\|^2+\gamma_k^2\sigma_{f,1}^2\\
\leq&\left(1+\frac{\gamma_k\mu_g}{2}\right)\|(I-\gamma_k\nabla_{22}^2g(x^k,y^k))(z^k-z_*^k)\|^2\\
&+\left(1+\frac{2}{\gamma_k\mu_g}\right)\|\gamma_k\left(\nabla_{22}^2g(x^k,y^k)z_*^k-\nabla_{22}^2g(x^k,y_*^k)z_*^k+\nabla_2f(x^k,y_*^k)-\nabla_2f(x^k,y^k)\right)\|^2\\
&+2\gamma_k^2\sigma_{g,2}^2\left(\|z^k-z_*^k\|^2+\|z_*^k\|^2\right)+\gamma_k^2\sigma_{f,1}^2\\
\leq&\left(\left(1+\frac{\gamma_k\mu_g}{2}\right)(1-\gamma_k\mu_g)^2+2\gamma_k^2\sigma_{g,2}^2\right)\|z^k-z_*^k\|^2\\
&+\left(\frac{4\gamma_k}{\mu_g}+2\gamma_k^2\right)\left(L_{\nabla_{22}^2g}^2\|z_*^k\|^2+L_{\nabla_2f}^2\right)\|y^k-y_*^k\|^2+2\gamma_k^2\sigma_{g,2}^2\|z_*^k\|^2+\gamma_k^2\sigma_{f,1}^2.\\
\leq&\left(1-\frac{4\gamma_k\mu_g}{3}\right)\|z^k-z_*^k\|^2+\left(\frac{4\gamma_k}{\mu_g}+2\gamma_k^2\right)\left(\frac{L_{\nabla_{22}^2g}^2L_f^2}{\mu_g^2}+L_f^2\right)\|y^k-y_*^k\|^2+\left(\frac{2\sigma_{g,2}^2L_f^2}{\mu_g^2}+\sigma_{f,1}^2\right)\gamma_k^2,\end{aligned} \tag{24}$$

where the first inequality uses Assumption 2, the second inequality uses Cauchy-Schwarz inequality and the definition of z_*^k, the third inequality uses Cauchy-Schwarz inequality and the fact that g is μ_g-strongly convex, and the fourth inequality uses Cauchy-Schwarz inequality, (13) and $-\frac{\gamma_k\mu_g}{6}+2\gamma_k^2\sigma_{g,2}^2+\frac{\gamma_k^3\mu_g^3}{2}\leq 0$, which is a direct result from the bound of γ_k in (19). It is worth noting that our estimation can be viewed as a refined version of (72) - (75) in Dagréou et al. (2022) Combining (23) and (24) we may obtain

$$\begin{aligned}&\mathbb{E}\left[\|z^{k+1}-z_*^{k+1}\|^2|\mathscr{F}_k\right]\\
\leq&\left(1+\frac{\gamma_k\mu_g}{3}\right)\mathbb{E}\left[\|z^{k+1}-z_*^k\|^2|\mathscr{F}_k\right]+\left(\frac{3\alpha_k^2}{\gamma_k\mu_g}+\alpha_k^2\right)L_{z^*}^2\|x_+^k-x^k\|^2\\
\leq&\left(1+\frac{\gamma_k\mu_g}{3}\right)\left[\left(1-\frac{4\gamma_k\mu_g}{3}\right)\|z^k-z_*^k\|^2+\left(\frac{4\gamma_k}{\mu_g}+2\gamma_k^2\right)\left(\frac{L_{\nabla_{22}^2g}^2L_f^2}{\mu_g^2}+L_f^2\right)\|y^k-y_*^k\|^2\right]\\
&+\left(1+\frac{\gamma_k\mu_g}{3}\right)\left(\frac{2\sigma_{g,2}^2L_f^2}{\mu_g^2}+\sigma_{f,1}^2\right)\gamma_k^2+\left(\frac{3\alpha_k^2}{\gamma_k\mu_g}+\alpha_k^2\right)L_{z^*}^2\|x_+^k-x^k\|^2\\
=&\left(1-\gamma_k\mu_g\right)\|z^k-z_*^k\|^2+\left(\frac{4\gamma_k}{\mu_g}+\frac{10\gamma_k^2}{3}+\frac{2\gamma_k^3\mu_g}{3}\right)\left(\frac{L_{\nabla_{22}^2g}^2L_f^2}{\mu_g^2}+L_f^2\right)\|y^k-y_*^k\|^2\\
&+\sigma_w^2\left(\gamma_k^2+\frac{\gamma_k^3\mu_g}{3}\right)+\left(\frac{3\alpha_k^2}{\gamma_k\mu_g}+\alpha_k^2\right)L_{z^*}^2\|x_+^k-x^k\|^2\end{aligned}$$

$$\leq (1-\gamma_k\mu_g)\,\|z^k - z_*^k\|^2 + \frac{5\gamma_k L_f^2}{\mu_g}\left(\frac{L_{\nabla_{22}^2 g}^2}{\mu_g^2}+1\right)\|y^k - y_*^k\|^2 + 2\sigma_w^2\gamma_k^2 + \frac{4\alpha_k^2 L_{z^*}^2}{\gamma_k\mu_g}\|x_+^k - x^k\|^2,$$

where the equality uses the definition of σ_w^2 in (17) and the third inequality uses $\gamma_k\mu_g \leq \frac{1}{4}$. Taking summation ($k$ from 0 to K) and expectation, we know

$$\begin{aligned}\sum_{k=0}^K \gamma_k\mu_g\mathbb{E}[\|z^k - z_*^k\|^2] \leq& \mathbb{E}\left[\|z^0 - z_*^0\|^2\right] + \sum_{k=0}^K \frac{5\gamma_k L_f^2}{\mu_g}\left(\frac{L_{\nabla_{22}^2 g}^2}{\mu_g^2}+1\right)\mathbb{E}[\|y^k - y_*^k\|^2] \\ &+ \sum_{k=0}^K 2\sigma_w^2\gamma_k^2 + \sum_{k=0}^K \frac{4\alpha_k^2 L_{z^*}^2}{\gamma_k\mu_g}\mathbb{E}[\|x_+^k - x^k\|^2].\end{aligned}$$

This completes the proof of the second inequality in (20) by dividing $c_2\mu_g$ on both sides and replacing $\sum_{k=0}^K \alpha_k\mathbb{E}[\|y^k - y_*^k\|^2]$ with its upper bound in (20). ■

Lemma 16 *Suppose Assumptions 1 and 2 hold. We have*

$$\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Phi(x^k)\|^2 \leq 3\big(\big(L_{\nabla f}^2 + L_{\nabla^2 g}^2\big)\|y^k - y_*^k\|^2 + L_{\nabla g}^2\|z^k - z_*^k\|^2\big),$$

Proof Note that we have the following decomposition:

$$\begin{aligned}&\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Phi(x^k) \\ =&\mathbb{E}[u_x^{k+1}|\mathscr{F}_k] - \nabla_1 f(x^k, y_*^k) - \left(\mathbb{E}\left[J^{k+1}|\mathscr{F}_k\right] z^k - \nabla_{12}^2 g(x^k, y_*^k) z_*^k\right) \\ =&\nabla_1 f(x^k, y^k) - \nabla_1 f(x^k, y_*^k) - \nabla_{12}^2 g(x^k, y^k)\left(z^k - z_*^k\right) - \big(\nabla_{12}^2 g(x^k, y^k) - \nabla_{12}^2 g(x^k, y_*^k)\big) z_*^k.\end{aligned}$$

which, together with Cauchy-Schwarz inequality, implies the conclusion:

$$\begin{aligned}\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Phi(x^k)\|^2 \leq& 3\|\nabla_1 f(x^k, y^k) - \nabla_1 f(x^k, y_*^k)\|^2 + 3\|\nabla_{12}^2 g(x^k, y^k)(z^k - z_*^k)\|^2 \\ &+ 3\|\big(\nabla_{12}^2 g(x^k, y^k) - \nabla_{12}^2 g(x^k, y_*^k)\big) z_*^k\|^2 \\ \leq& 3\left(\big(L_{\nabla f}^2 + L_{\nabla^2 g}^2\big)\|y^k - y_*^k\|^2 + L_{\nabla g}^2\|z^k - z_*^k\|^2\right).\end{aligned}$$

■

7.1.2 Primal Convergence

Lemma 17 *Suppose Assumptions 1 and 2 hold. If*

$$\alpha_k \leq \min\left(\frac{\tau^2}{20c_3},\ \frac{c_3}{2\tau\left(c_3 L_{\nabla\Phi} + L_{\nabla\eta_{\mathcal{X}}}\right)},\ 1\right),\ \tau < 1,\ c_3 \leq \frac{1}{10}, \tag{25}$$

then in Algorithm 1 we have

$$\sum_{k=0}^K \frac{\alpha_k}{\tau^2}\mathbb{E}[\|x_+^k - x^k\|^2] \leq \frac{2}{\tau}\mathbb{E}\left[W_{0,1}\right] + 3\sum_{k=0}^K \alpha_k\mathbb{E}\left[\|\nabla\Phi(x^k) - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2\right]$$

$$+\frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|h^k-\nabla\Phi(x^k)\|^2]+\sum_{k=0}^{K}\left(\alpha_k^2\sigma_{g,2}^2\mathbb{E}[\|z^k-z_*^k\|^2]+\alpha_k^2\sigma_w^2\right). \tag{26}$$

Proof The $L_{\nabla\Phi}$-smoothness of $\Phi(x)$ and $L_{\nabla\eta_{\mathcal{X}}}$-smoothness of $\eta_{\mathcal{X}}$ (Lemmas 10, 11) imply

$$\Phi(x^{k+1})-\Phi(x^k)\le\alpha_k\left\langle\nabla\Phi(x^k),x_+^k-x^k\right\rangle+\frac{L_{\nabla\Phi}}{2}\|x^{k+1}-x^k\|^2 \tag{27}$$

and

$$\begin{aligned}
&\eta_{\mathcal{X}}(x^k,h^k,\tau)-\eta_{\mathcal{X}}(x^{k+1},h^{k+1},\tau)\\
\le&\left\langle-h^k+\frac{1}{\tau}(x^k-x_+^k),x^k-x^{k+1}\right\rangle+\left\langle x_+^k-x^k,h^k-h^{k+1}\right\rangle\\
&+\frac{L_{\nabla\eta_{\mathcal{X}}}}{2}\left(\|x^{k+1}-x^k\|^2+\|h^{k+1}-h^k\|^2\right)\\
=&\alpha_k\left\langle h^k,x_+^k-x^k\right\rangle+\frac{\alpha_k}{\tau}\|x_+^k-x^k\|^2+\theta_k\left\langle h^k,x_+^k-x^k\right\rangle-\theta_k\left\langle w^{k+1},x_+^k-x^k\right\rangle\\
&+\frac{L_{\nabla\eta_{\mathcal{X}}}}{2}\left(\|x^{k+1}-x^k\|^2+\|h^{k+1}-h^k\|^2\right)\\
\le&-\frac{\theta_k}{\tau}\|x_+^k-x^k\|^2-\theta_k\left\langle w^{k+1},x_+^k-x^k\right\rangle+\frac{L_{\nabla\eta_{\mathcal{X}}}}{2}\left(\|x^{k+1}-x^k\|^2+\|h^{k+1}-h^k\|^2\right),
\end{aligned} \tag{28}$$

where the first inequality uses $L_{\nabla\eta_{\mathcal{X}}}$-smoothness of $\nabla\eta_{\mathcal{X}}$, and the second inequality uses the optimality condition (14) (with $d=x^k$). Hence by computing (27) + (28)/c_3 and taking conditional expectation with respect to $\mathscr{F}_k$ we know

$$\begin{aligned}
&\frac{\alpha_k}{\tau}\|x_+^k-x^k\|^2\\
\le&\frac{1}{c_3}\left(\mathbb{E}\left[\eta_{\mathcal{X}}(x^{k+1},h^{k+1},\tau)|\mathscr{F}_k\right]-\eta_{\mathcal{X}}(x^k,h^k,\tau)\right)+\Phi(x^k)-\mathbb{E}\left[\Phi(x^{k+1})|\mathscr{F}_k\right]\\
&+\alpha_k\left\langle\nabla\Phi(x^k)-\mathbb{E}[w^{k+1}|\mathscr{F}_k],x_+^k-x^k\right\rangle+\frac{(c_3L_{\nabla\Phi}+L_{\nabla\eta_{\mathcal{X}}})}{2c_3}\|x^{k+1}-x^k\|^2\\
&+\frac{L_{\nabla\eta_{\mathcal{X}}}}{2c_3}\mathbb{E}\left[\|h^{k+1}-h^k\|^2|\mathscr{F}_k\right]\\
=&W_{k,1}-\mathbb{E}\left[W_{k+1,1}|\mathscr{F}_k\right]+\alpha_k\left\langle\nabla\Phi(x^k)-\mathbb{E}[w^{k+1}|\mathscr{F}_k],x_+^k-x^k\right\rangle\\
&+\frac{(c_3L_{\nabla\Phi}+L_{\nabla\eta_{\mathcal{X}}})}{2c_3}\|x^{k+1}-x^k\|^2+\frac{L_{\nabla\eta_{\mathcal{X}}}}{2c_3}\mathbb{E}\left[\|h^{k+1}-h^k\|^2|\mathscr{F}_k\right]\\
\le&W_{k,1}-\mathbb{E}\left[W_{k+1,1}|\mathscr{F}_k\right]+\alpha_k\left(\tau\|\nabla\Phi(x^k)-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2+\frac{1}{4\tau}\|x_+^k-x^k\|^2\right)\\
&+\frac{\alpha_k}{4\tau}\|x_+^k-x^k\|^2+\frac{5}{2c_3\tau}\mathbb{E}\left[\|h^{k+1}-h^k\|^2|\mathscr{F}_k\right],
\end{aligned} \tag{29}$$

where the second inequality uses Young's inequality and $\frac{\alpha_k^2(c_3L_{\nabla\Phi}+L_{\nabla\eta_{\mathcal{X}}})}{2c_3}\le\frac{\alpha_k}{4\tau}$, $L_{\nabla\eta_{\mathcal{X}}}<\frac{5}{\tau}$ when (25) holds. Note that by (18) we know

$$\frac{5}{c_3\tau^2}\mathbb{E}[\|h^{k+1}-h^k\|^2]\le\frac{10c_3\alpha_k^2}{\tau^2}\mathbb{E}\left[\|h^k-\nabla\Phi(x^k)\|^2+\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\|^2\right]$$

$$
\begin{aligned}
&+\frac{5c_3\alpha_k^2}{\tau^2}\sigma_w^2+\frac{10c_3\alpha_k^2\sigma_{g,2}^2}{\tau^2}\mathbb{E}[\|z^k-z_*^k\|^2].\\
\leq&\frac{\alpha_k}{2}\mathbb{E}[\|h^k-\nabla\Phi(x^k)\|^2]+\alpha_k\mathbb{E}\left[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\|^2\right]\\
&+\alpha_k^2\sigma_w^2+\alpha_k^2\sigma_{g,2}^2\mathbb{E}[\|z^k-z_*^k\|^2], \qquad (30)
\end{aligned}
$$

where the second inequality uses (25). Taking summation and expectation on both sides of (29) and using (30), we obtain (26). ■

7.1.3 Dual Convergence

Lemma 18 *Suppose Assumptions 1 and 2 hold. In Algorithm 1 we have*

$$
\begin{aligned}
&\sum_{k=0}^K\alpha_k\mathbb{E}[\|h^k-\nabla\Phi(x^k)\|^2]\\
\leq&\frac{1}{c_3}\mathbb{E}\left[\|h^0-\nabla\Phi(x^0)\|^2\right]+2\sum_{k=0}^K\alpha_k\mathbb{E}\left[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\|^2\right]\\
&+\frac{2L_{\nabla\Phi}^2}{c_3^2}\sum_{k=0}^K\alpha_k\mathbb{E}[\|x_+^k-x^k\|^2]+2c_3\sigma_{g,2}^2\sum_{k=0}^K\alpha_k^2\mathbb{E}[\|z^k-z_*^k\|^2]+\sum_{k=0}^K c_3\alpha_k^2\sigma_w^2. \qquad (31)
\end{aligned}
$$

Proof Note that by moving-average update of h^k, we have

$$
\begin{aligned}
&h^{k+1}-\nabla\Phi(x^{k+1})\\
=&(1-\theta_k)h^k+\theta_k(w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k])+\theta_k\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^{k+1})\\
=&(1-\theta_k)(h^k-\nabla\Phi(x^k))+\theta_k(\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k))+\nabla\Phi(x^k)-\nabla\Phi(x^{k+1})\\
&+\theta_k(w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k])
\end{aligned}
$$

Hence we know

$$
\begin{aligned}
&\mathbb{E}\left[\|h^{k+1}-\nabla\Phi(x^{k+1})\|^2|\mathscr{F}_k\right]\\
=&\|(1-\theta_k)(h^k-\nabla\Phi(x^k))+\theta_k(\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k))+\nabla\Phi(x^k)-\nabla\Phi(x^{k+1})\|^2\\
&+\theta_k^2\mathbb{E}\left[\|w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k\right]\\
\leq&(1-\theta_k)\|h^k-\nabla\Phi(x^k)\|^2+\theta_k\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)+\frac{1}{\theta_k}(\nabla\Phi(x^k)-\nabla\Phi(x^{k+1}))\|^2+\theta_k^2\sigma_{w,k+1}^2\\
\leq&(1-\theta_k)\|h^k-\nabla\Phi(x^k)\|^2+2\theta_k\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\|^2+\frac{2}{\theta_k}\|\nabla\Phi(x^k)-\nabla\Phi(x^{k+1})\|^2+\theta_k^2\sigma_{w,k+1}^2\\
\leq&(1-\theta_k)\|h^k-\nabla\Phi(x^k)\|^2+2\theta_k\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\|^2+\frac{2\alpha_k^2L_{\nabla\Phi}^2}{\theta_k}\|x_+^k-x^k\|^2+\theta_k^2\sigma_{w,k+1}^2, \qquad (32)
\end{aligned}
$$

where the first equality uses the fact that x^k, h^k, x^{k+1}, are all $\mathscr{F}_k$-measurable and are independent of w^{k+1} given $\mathscr{F}_k$, the first inequality uses the convexity of $\|\cdot\|^2$ and (17), the second inequality uses Cauchy-Schwarz inequality, the third inequality uses the Lipschitz

continuity of $\nabla\Phi$ in Lemma 19, and the update rules of x^{k+1}. Taking summation, expectation on both sides of (32), dividing c_3 and using (17), we know (31) holds. ■

7.1.4 Proof of Theorem 3

Now we are ready to prove Theorem 3. From Lemma 12 we know it suffices to bound V_k. By definition of V_k in (16), (26) and (31) we have

$$\begin{aligned}
&\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[V_k\right]=\sum_{k=0}^{K}\Big(\frac{\alpha_k}{\tau^2}\mathbb{E}[\|x_+^k-x^k\|^2]+\alpha_k\mathbb{E}[\|h^k-\nabla\Phi(x^k)\|^2]\Big) \qquad (33)\\
\leq&\frac{2L_{\nabla\Phi}^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|x_+^k-x^k\|^2]+\frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|h^k-\nabla\Phi(x^k)\|^2]\\
&+5\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\|\nabla\Phi(x^k)-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2\right]+(1+2c_3)\sigma_{g,2}^2\sum_{k=0}^{K}\alpha_k^2\mathbb{E}[\|z^k-z_*^k\|^2]\\
&+\frac{2}{\tau}\mathbb{E}\left[W_{0,1}\right]+\frac{1}{c_3}\mathbb{E}\left[\|h^0-\nabla\Phi(x^0)\|^2\right]+(1+c_3)\sigma_w^2\Big(\sum_{k=0}^{K}\alpha_k^2\Big),\\
\leq&\frac{2L_{\nabla\Phi}^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|x_+^k-x^k\|^2]+\frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|h^k-\nabla\Phi(x^k)\|^2]\\
&+15\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\big(L_{\nabla f}^2+L_{\nabla^2 g}^2\big)\|y^k-y_*^k\|^2+L_{\nabla g}^2\|z^k-z_*^k\|^2\right]+L_{\nabla g}^2\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|z^k-z_*^k\|^2]\\
&+\frac{2}{\tau}\mathbb{E}\left[W_{0,1}\right]+\frac{1}{c_3}\mathbb{E}\left[\|h^0-\nabla\Phi(x^0)\|^2\right]+(1+c_3)\sigma_w^2\Big(\sum_{k=0}^{K}\alpha_k^2\Big)\\
\leq&C_{vx}\tau^2\sum_{k=0}^{K}\frac{\alpha_k}{\tau^2}\mathbb{E}[\|x_+^k-x^k\|^2]+C_{vh}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|h^k-\nabla\Phi(x^k)\|^2]+C_{v,0}+C_{v,1}\Big(\sum_{k=0}^{K}\alpha_k^2\Big),
\end{aligned}$$

where we assume

$$(1+2c_3)\sigma_{g,2}^2\alpha_k\leq L_{\nabla g}^2, \qquad (34)$$

in the second inequality. The constants are defined as

$$\begin{aligned}
C_{vx}&=15\big(L_{\nabla f}^2+L_{\nabla^2 g}^2\big)C_{yx}+16L_{\nabla g}^2C_{zx}+\frac{2L_{\nabla\Phi}^2}{c_3^2},\ C_{vh}=\frac{1}{2},\\
C_{v,0}&=15\big(L_{\nabla f}^2+L_{\nabla^2 g}^2\big)C_{y,0}+16L_{\nabla g}^2C_{z,0}+\frac{2}{\tau}\mathbb{E}\left[W_{0,1}\right]+\frac{1}{c_3}\mathbb{E}\left[\|h^0-\nabla\Phi(x^0)\|^2\right],\\
C_{v,1}&=15\big(L_{\nabla f}^2+L_{\nabla^2 g}^2\big)C_{y,1}+16L_{\nabla g}^2C_{z,1}+(1+c_3)\sigma_w^2.
\end{aligned}$$

Using constants defined in Lemma 15, we know

$$C_{vx}=\mathcal{O}\Big(\frac{\kappa^8}{c_1^2}+\frac{\kappa^4}{c_2^2}+\frac{\kappa^6}{c_3^2}\Big),C_{vh}=\mathcal{O}(1),C_{v,0}=\mathcal{O}\Big(\frac{\kappa^5}{c_1}+\frac{\kappa^2}{c_2}+\frac{1}{\tau}\Big),C_{v,1}=\mathcal{O}\big(c_1\kappa^5+c_2\kappa^2\big).$$

Hence we can pick $\alpha_k \equiv \Theta(1/\sqrt{K}),\ \tau = \Theta\left(\kappa^{-4}\right),\ c_1 = \mathcal{O}(1),\ c_2 = \mathcal{O}(1),\ c_3 = \mathcal{O}(1)$ so that the conditions ((19), (25) and (34)) in previous lemmas hold, and $\tau = \Theta(\kappa^{-4})$ such that

$$C_{vx}\tau^2 = \mathcal{O}(\kappa^8\tau^2) \le \frac{1}{2}.$$

Plugging in all the constants in (33), we have

$$\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}[V_k] \le \frac{1}{2K}\big(\sum_{k=0}^{K}\frac{1}{\tau^2}\mathbb{E}[\|x_+^k - x^k\|^2] + \sum_{k=0}^{K}\mathbb{E}[\|h^k - \nabla\Phi(x^k)\|^2]\big) + \mathcal{O}\big(\frac{\kappa^5}{\sqrt{K}}\big).$$

Then we have $\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\left[V_k\right] = \mathcal{O}(\kappa^5/\sqrt{K})$. which, together with Lemma 12, proves Theorem 3.

7.2 Proof of Theorem 6

In this section we present our proof of Theorem 6. For simplicity, we summarize the notations that will be used in our proof as follows.

$$L_{\nabla f} = \max_{1\le i\le n} L_{\nabla f_i},\ L_{\nabla g} = \max_{1\le i\le n} L_{\nabla g_i},\ L_{\nabla^2 g_i} = \max_{1\le i\le n} L_{\nabla^2 g_i},\ \mu_g = \max_{1\le i\le n}\mu_{g_i},$$

$$\kappa = \max(L_{\nabla f}, L_{\nabla g})/\mu_g,\ u_x^{k+1} = \sum_{i=1}^{n} u_{x,i}^{k+1},\ w^{k+1} = \sum_{i=1}^{n}\lambda_i^k\big(u_{x,i}^{k+1} - J_i^{k+1}z_i^k\big),$$

$$\lambda_*^k = \lambda_*(x^k) = \arg\max_{\lambda\in\Delta_n}\Phi_{\mu_\lambda}(x^k,\lambda),\ y_{*,i}^k = y_i^*(x^k) = \arg\min_{y\in\mathbb{R}^{d_y}} g_i(x^k,y),$$

$$\Phi_i(x) = f_i(x, y_i^*(x)),\ \Phi^k = \Big(\Phi_1(x^k),...,\Phi_n(x^k)\Big)^\top, z_{*,i}^k = \big(\nabla_{22}^2 g_i(x^k, y_{*,i}^k)\big)^{-1}\nabla_2 f_i(x^k, y_{*,i}^k),$$

$$\Psi(x) = \max_{\lambda\in\Delta_n}\Phi_{\mu_\lambda}(x,\lambda) = \max_{\lambda\in\Delta_n}\left(\sum_{i=1}^{n}\lambda_i\Phi_i(x) - \frac{\mu_\lambda}{2}\|\lambda - \frac{\mathbf{1}_n}{n}\|^2\right),$$

$$\eta_X(x,h,\tau) = \min_{d\in X}\big\{\ \langle h, d-x\rangle + \frac{1}{2\tau}\|d-x\|^2\big\},\ \text{where } X = \mathcal{X} \text{ or } \Delta_n.$$

In this subsection we suppose Assumptions 1, 2 hold for all f_i, g_i and Assumption 3 holds. We suppose stepsizes in Algorithm 2 satisfy $\beta_k = c_1\alpha_k,\ \gamma_k = c_2\alpha_k, \theta_k = c_3\alpha_k$, where $c_1, c_2, c_3 > 0$ are constants to be determined. We will utilize the following merit function in our analysis:

$$\tilde{W}_k = \tilde{W}_{k,1} + \tilde{W}_{k,2}, \tilde{W}_{k,1} = \tilde{W}_{k,1}^{(1)} + \tilde{W}_{k,1}^{(2)}, \tilde{W}_{k,1}^{(1)} = \Psi(x^k) - \Phi_{\mu_\lambda}(x^k,\lambda^k) - \frac{1}{c_3}\eta_{\Delta_n}(\lambda^k, -h_\lambda^k, \tau_\lambda)$$

$$\tilde{W}_{k,1}^{(2)} = \Psi(x^k) - \inf_{x\in\mathcal{X}}\Psi(x) - \frac{1}{c_3}\eta_{\mathcal{X}}(x^k, h_x^k, \tau_x), \tilde{W}_{k,2} = \sum_{i=1}^{n}\big(\frac{1}{c_1}\|y_i^k - y_{*,i}^k\|^2 + \frac{1}{c_2}\|z_i^k - z_{*,i}^k\|^2\big).$$

By definition of $\Psi, \eta_{\mathcal{X}}, \eta_{\Delta_n}$, we can verify that $\tilde{W}_{k,1}^{(1)} \ge 0, \tilde{W}_{k,1}^{(2)} \ge 0$. Moreover, as discussed in Section 4.2, we consider the following optimality measure:

$$\tilde{V}_k = \underbrace{\frac{1}{\tau_x^2}\|x_+^k - x^k\|^2 + \|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2}_{\tilde{V}_{k,1}\text{: Optimality of min problem}} + \underbrace{\frac{1}{\tau_\lambda^2}\|\lambda_+^k - \lambda^k\|^2 + \|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2}_{\tilde{V}_{k,2}\text{: Optimality of max problem}}. \tag{35}$$

The following lemma provides some smoothness of functions that we will use in our proof.

Lemma 19 *Functions* $\nabla\Psi(\cdot), \nabla_1\Phi_{\mu_\lambda}(\cdot,\lambda), \nabla_1\Phi(\cdot,\lambda), \nabla_1\Phi_{\mu_\lambda}(x,\cdot), \nabla_1\Phi(x,\cdot), \nabla_2\Phi_{\mu_\lambda}(\cdot,\lambda)$, $\nabla_2\Phi_{\mu_\lambda}(x,\cdot)$ *are* $L_{\nabla\Psi}, L_{\nabla\Phi}, L_{\nabla\Phi}, L_{\nabla_1\Phi_{\mu_\lambda}}, L_{\nabla_1\Phi_{\mu_\lambda}}, L_{\nabla_2\Phi_{\mu_\lambda}}, \mu_\lambda$*-Lipschitz continuous respectively, with constants given by* $L_{\nabla\Psi} = \frac{n}{\mu_\lambda}\left(L_\Phi^2 + b_\Phi L_{\nabla\Phi}\right) + L_{\nabla\Phi}$, $L_{\nabla_1\Phi_{\mu_\lambda}} = L_{\nabla_2\Phi_{\mu_\lambda}} = \sqrt{n}L_\Phi$.

Proof For $\nabla\Psi$ we first notice that the nonconvex-strongly-concave problem in (10) can be reformulated as a bilevel problem:

$$\min_{x\in\mathcal{X}} \Psi(x) = \Phi_{\mu_\lambda}(x,\lambda^*(x)) \text{ s.t. } \lambda^*(x) = \arg\min_{\lambda\in\Delta_n} \left(-\Phi_{\mu_\lambda}(x,\lambda)\right) = \frac{\mu_\lambda}{2}\|\lambda - \frac{\mathbf{1}_n}{n}\|^2 - \sum_{i=1}^n \lambda_i\Phi_i(x).$$

By Lemma 10 we know

$$\begin{aligned}
\nabla\Psi(x) =& \nabla_1\Phi_{\mu_\lambda}(x,\lambda^*(x)) - \nabla_{12}^2\Phi_{\mu_\lambda}(x,\lambda^*(x))\left(\nabla_{22}^2\Phi_{\mu_\lambda}(x,\lambda^*(x))\right)^{-1}\nabla_2\Phi_{\mu_\lambda}(x,\lambda^*(x)) \\
=& \sum_{i=1}^n \lambda_i^*(x)\nabla\Phi_i(x) + \frac{1}{\mu_\lambda}\left(\nabla\Phi_1(x),...,\nabla\Phi_n(x)\right)\left[\begin{pmatrix}\Phi_1(x)\\ \vdots \\ \Phi_n(x)\end{pmatrix} - \mu_\lambda\left(\lambda^*(x) - \frac{\mathbf{1}_n}{n}\right)\right] \\
=& \frac{1}{\mu_\lambda}\sum_{i=1}^n \Phi_i(x)\nabla\Phi_i(x) + \frac{1}{n}\sum_{i=1}^n \nabla\Phi_i(x),
\end{aligned}$$

from which we know $\nabla\Psi(\cdot)$ is $L_{\nabla\Psi}$-Lipschitz continuous since

$$\begin{aligned}
&\|\Phi_i(x)\nabla\Phi_i(x) - \Phi_i(\tilde{x})\nabla\Phi_i(\tilde{x})\| \\
\le& \|\Phi_i(x)\nabla\Phi_i(x) - \Phi_i(x)\nabla\Phi_i(\tilde{x})\| + \|\Phi_i(x)\nabla\Phi_i(\tilde{x}) - \Phi_i(\tilde{x})\nabla\Phi_i(\tilde{x})\| \\
\le& \left(L_\Phi^2 + b_\Phi L_{\nabla\Phi}\right)\|x-y\|.
\end{aligned}$$

Note that for any fixed $\lambda\in\Delta_n$ and $x,\tilde{x}\in\mathcal{X}$, we have

$$\nabla_1\Phi_{\mu_\lambda}(x,\lambda) = \nabla_1\Phi(x,\lambda) = \sum_{i=1}^n \lambda_i\nabla\Phi_i(x), \tag{36}$$

$$\|\nabla_1\Phi_{\mu_\lambda}(x,\lambda) - \nabla_1\Phi_{\mu_\lambda}(\tilde{x},\lambda)\| = \|\sum_{i=1}^n \lambda_i\left(\nabla\Phi_i(x) - \nabla\Phi_i(\tilde{x})\right)\| \le L_{\nabla\Phi}\|x-\tilde{x}\|. \tag{37}$$

Similarly, for any fixed $x\in\mathcal{X}$ and $\lambda,\tilde{\lambda}\in\Delta_n$ we know

$$\|\nabla_1\Phi_{\mu_\lambda}(x,\lambda) - \nabla_1\Phi_{\mu_\lambda}(x,\tilde{\lambda})\| = \|\sum_{i=1}^n(\lambda_i - \tilde{\lambda}_i)\nabla\Phi_i(x)\| \le \sqrt{n}L_\Phi\|\lambda-\tilde{\lambda}\|. \tag{38}$$

(36), (37) and (38) imply $\nabla_1\Phi_{\mu_\lambda}(\cdot,\lambda), \nabla_1\Phi(\cdot,\lambda)$ are $L_{\nabla\Phi}$-Lipschitz continuous and $\nabla_1\Phi(x,\cdot)$, $\nabla_1\Phi_{\mu_\lambda}(x,\cdot)$ are $L_{\nabla_1\Phi_{\mu_\lambda}}$-Lipschitz continuous. Finally, for $\nabla_2\Phi_{\mu_\lambda}(x,\lambda)$ we have $\nabla_2\Phi_{\mu_\lambda}(x,\lambda)$ $= \left(\Phi_1(x),...,\Phi_n(x)\right)^\top - \mu_\lambda\left(\lambda - \frac{\mathbf{1}_n}{n}\right)$, and thus functions $\nabla_2\Phi_{\mu_\lambda}(\cdot,\lambda), \nabla_2\Phi_{\mu_\lambda}(x,\cdot)$ are $\sqrt{n}L_\Phi$, μ_λ-Lipschitz continuous respectively. ■

Next we present a technical lemma that will be used in analyzing the strongly convex function over a closed convex set.

Lemma 20 *Suppose $f(x)$ is μ-strongly convex and L-smooth over a closed convex set $\mathcal{X}$. For any $\tau \le \frac{1}{L}$ define $x_+ = \Pi_{\mathcal{X}}(x - \tau\nabla f(x))$ and $x_* = \arg\min_{x\in\mathcal{X}} f(x)$, we have $\big(1 - \sqrt{1-\tau\mu}\big)\|x - x_*\| \le \|x - x_+\|$.*

Proof By Corollary 2.2.4 in Nesterov (2018) we know

$$\begin{aligned}\frac{1}{\tau}\langle x - x_+, x - x_*\rangle &\ge \frac{1}{2\tau}\|x - x_+\|^2 + \frac{\mu}{2}\|x - x_*\|^2 + \frac{\mu}{2}\|x_+ - x_*\|^2\\ &= \left(\frac{1}{2\tau} + \frac{\mu}{2}\right)\|x - x_+\|^2 + \mu\|x - x_*\|^2 - \mu\langle x - x_+, x - x_*\rangle\end{aligned}$$

which implies $\|x - x_+\|\|x - x_*\| \ge \langle x - x_+, x - x_*\rangle \ge \frac{1}{2}\|x - x_+\|^2 + r\|x - x_*\|^2$, where $r = \frac{\mu}{\frac{1}{\tau}+\mu} \le \frac{1}{2}$. Applying Young's inequality to the left hand side of the above inequality, we know $\frac{1+\sqrt{1-2r}}{4r}\|x - x_+\|^2 + \frac{r}{1+\sqrt{1-2r}}\|x - x_*\|^2 \ge \frac{1}{2}\|x - x_+\|^2 + r\|x - x_*\|^2$, which gives $\|x - x_+\| \ge \big(1 - \sqrt{1-2r}\big)\|x - x_*\| \ge \big(1 - \sqrt{1-\tau\mu}\big)\|x - x_*\|$. This completes the proof. ■

The next lemma shows the relation between the stationarity used in Theorem 6 and our measure of optimality $\tilde{V}_k$ in (35).

Lemma 21 *Suppose Assumptions 1, 2 hold for all f_i, g_i and Assumption 3 holds. If $\tau_\lambda\mu_\lambda = 1$, then in Algorithm 2 we have*

$$\begin{aligned}&\frac{1}{\tau_x^2}\|x^k - \Pi_{\mathcal{X}}\big(x^k - \tau_x\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\big)\|^2 \le 2\left(\frac{1}{\tau_x^2}\|x_+^k - x^k\|^2 + \|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2\right),\\ &\|\lambda^k - \lambda_*^k\|^2 \le \frac{2}{\mu_\lambda^2}\left(\frac{1}{\tau_\lambda^2}\|\lambda_+^k - \lambda^k\|^2 + \|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2\right),\end{aligned}$$

which imply $\|\frac{1}{\tau_x}\big(x^k - \Pi_{\mathcal{X}}\big(x^k - \tau_x\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\big)\big)\|^2 + \|\lambda^k - \lambda_*^k\|^2 \le \max\left(2, \frac{2}{\mu_\lambda^2}\right)\tilde{V}_k$.

Proof The first inequality follows (12):

$$\begin{aligned}&\|x^k - \Pi_{\mathcal{X}}\big(x^k - \tau_x\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\big)\|^2\\ \le&2\big(\|x_+^k - x^k\|^2 + \|\Pi_{\mathcal{X}}\big(x^k - \tau_x h_x^k\big) - \Pi_{\mathcal{X}}\big(x^k - \tau_x\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\big)\|^2\big)\\ \le&2\big(\|x_+^k - x^k\|^2 + \tau_x^2\|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2\big),\end{aligned}$$

where the first inequality uses Cauchy-Schwarz inequality and the second inequality uses the non-expansiveness of projection onto a closed convex set. Note $\lambda_*^k = \arg\min_{\lambda\in\Delta_n}\Phi_{\mu_\lambda}(x^k,\lambda)$ is a minimizer (over the probability simplex) of a μ_λ-smooth and μ_λ-strongly convex function $\Phi_{\mu_\lambda}(x^k,\cdot)$. Hence we know from Lemma 20 that

$$\begin{aligned}&\mu_\lambda^2\|\lambda_*^k - \lambda^k\|^2\\ \le&\tau_\lambda^{-2}\big(1 + \sqrt{1-\tau_\lambda\mu_\lambda}\big)^2\|\lambda^k - \Pi_{\Delta_n}\big(\lambda^k + \tau_\lambda\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\big)\|^2\\ \le&2\tau_\lambda^{-2}\big(1 + \sqrt{1-\tau_\lambda\mu_\lambda}\big)^2\big(\|\lambda_+^k - \lambda^k\|^2 + \|\Pi_{\Delta_n}(\lambda^k + \tau_\lambda h_\lambda^k) - \Pi_{\Delta_n}(\lambda^k + \tau_\lambda\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k))\|^2\big)\\ \le&2\tau_\lambda^{-2}\big(1 + \sqrt{1-\tau_\lambda\mu_\lambda}\big)^2\big(\|\lambda_+^k - \lambda^k\|^2 + \tau_\lambda^2\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2\big),\end{aligned}$$

where the second inequality uses Cauchy-Schwarz inequality and the third inequality uses non-expansiveness of the projection operator. Setting $\tau_\lambda \mu_\lambda = 1$ completes the proof. ■

Lemma 22 *Suppose Assumptions 1, 2 hold for all f_i, g_i and Assumption 3 holds. In Algorithm 2 we have*

$$\mathbb{E}\left[\|w^{k+1} - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2\right] \le \sigma_{w,k+1}^2 \tag{39}$$

$$\mathbb{E}[\|h_x^{k+1} - h_x^k\|^2] \le \sigma_{h_x,k}^2, \quad \mathbb{E}[\|h_\lambda^{k+1} - h_\lambda^k\|^2] \le \sigma_{h_\lambda,k}^2, \tag{40}$$

$$\sigma_{w,k+1}^2 := \sigma_w^2 + 2\sigma_{g,2}^2 \sum_{i=1}^n \mathbb{E}[\lambda_i^k \|z_i^k - z_{*,i}^k\|^2], \sigma_w^2 = \sigma_{f,1}^2 + \frac{2\sigma_{g,2}^2 L_f^2}{\mu_g^2}$$

$$\sigma_{h_x,k}^2 := 2\theta_k^2 \mathbb{E}[\|h_x^k - \nabla_1 \Phi_{\mu_\lambda}(x^k, \lambda^k)\|^2 + \|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1 \Phi_{\mu_\lambda}(x^k, \lambda^k)\|^2] + \theta_k^2 \sigma_{w,k+1}^2$$

$$\sigma_{h_\lambda,k}^2 := \theta_k^2 \mathbb{E}[\|h_\lambda^k - \nabla_2 \Phi_{\mu_\lambda}(x^k, \lambda^k)\|^2] + n\theta_k^2 \sigma_{f,0}^2.$$

Proof We first consider w^k. Note that

$$w^{k+1} - \mathbb{E}[w^{k+1}|\mathscr{F}_k] = \sum_{i=1}^n \lambda_i^k \left(u_{x,i}^{k+1} - \mathbb{E}[u_{x,i}^{k+1}|\mathscr{F}_k] - \left(J_i^{k+1} - \mathbb{E}[J_i^{k+1}|\mathscr{F}_k]\right) z_i^k\right).$$

Hence we know

$$\begin{aligned}
&\mathbb{E}\left[\|w^{k+1} - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k\right] \\
=&\sum_{i=1}^n \left(\lambda_i^k\right)^2 \left(\mathbb{E}\left[\|u_{x,i}^k - \mathbb{E}[u_{x,i}^k|\mathscr{F}_k]\|^2|\mathscr{F}_k\right] + \mathbb{E}\left[\|J_i^{k+1} - \mathbb{E}[J_i^{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k\right] \|z_i^k\|^2\right) \\
\le&\sum_{i=1}^n \lambda_i^k \left(\sigma_{f,1}^2 + 2\sigma_{g,2}^2 \|z_{*,i}^k\|^2 + 2\sigma_{g,2}^2 \|z_i^k - z_{*,i}^k\|^2\right) \\
\le&\sigma_{f,1}^2 + \frac{2\sigma_{g,2}^2 L_f^2}{\mu_g^2} + 2\sigma_{g,2}^2 \sum_{i=1}^n \lambda_i^k \|z_i^k - z_{*,i}^k\|^2.
\end{aligned}$$

Taking expectation on both sides proves (39). Next for $\|h_x^{k+1} - h_x^k\|$ we have

$$\begin{aligned}
&\mathbb{E}\left[\|h_x^{k+1} - h_x^k\|^2|\mathscr{F}_k\right] = \theta_k^2 \mathbb{E}\left[\|h_x^k - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k\right] + \theta_k^2 \mathbb{E}\left[\|w^{k+1} - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k\right] \\
&\le 2\theta_k^2 \mathbb{E}\left[\|h_x^k - \nabla_1 \Phi(x^k, \lambda^k)\|^2|\mathscr{F}_k\right] + 2\theta_k^2 \mathbb{E}\left[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1 \Phi(x^k, \lambda^k)\|^2|\mathscr{F}_k\right] + \theta_k^2 \sigma_{w,k+1}^2,
\end{aligned}$$

which proves the first inequality of (40). Similarly we have

$$\begin{aligned}
&\mathbb{E}\left[\|h_\lambda^{k+1} - h_\lambda^k\|^2|\mathscr{F}_k\right] \\
=&\theta_k^2 \mathbb{E}\left[\|h_\lambda^k - \mathbb{E}[s^{k+1}|\mathscr{F}_k] + \mu_\lambda(\lambda^k - \frac{\mathbf{1}_n}{n})\|^2|\mathscr{F}_k\right] + \theta_k^2 \mathbb{E}\left[\|s^{k+1} - \mathbb{E}[s^{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k\right] \\
\le&\theta_k^2 \mathbb{E}\left[\|h_\lambda^k - \nabla_2 \Phi_{\mu_\lambda}(x^k, \lambda^k)\|^2|\mathscr{F}_k\right] + n\theta_k^2 \sigma_{f,0}^2,
\end{aligned}$$

which proves the second inequality of (40). ■

7.2.1 Hypergradient Estimation Error

Lemma 23 *Suppose Assumptions 1, 2 hold for all f_i, g_i and Assumption 3 holds. In Algorithm 2 if the stepsizes satisfy*

$$\beta_k < \frac{2}{\mu_g + L_{\nabla g}},\ \gamma_k \leq \min\left(\frac{1}{4\mu_g},\ \frac{0.06\mu_g}{\sigma_{g,2}^2}\right), \tag{41}$$

then we have

$$\sum_{k=0}^{K}\alpha_k\mathbb{E}\Big[\sum_{i=1}^{n}\|y_i^k - y_{*,i}^k\|^2\Big] \leq nC_{yx}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|x_+^k - x^k\|^2] + \sum_{i=1}^{n}C_{y_i,0} + nC_{y,1}\Big(\sum_{k=0}^{K}\alpha_k^2\Big)$$
$$\sum_{k=0}^{K}\alpha_k\mathbb{E}\Big[\sum_{i=1}^{n}\|z_i^k - z_{*,i}^k\|^2\Big] \leq nC_{zx}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|x_+^k - x^k\|^2] + \sum_{i=1}^{n}C_{z_i,0} + nC_{z,1}\Big(\sum_{k=0}^{K}\alpha_k^2\Big)$$

where constants $C_{yx}, C_{y,1}, C_{zx}, C_{z,1}$ are defined in Lemma 15. $C_{y_i,0}, C_{z_i,0}$ are defined as

$$C_{y_i,0} = \frac{1}{c_1\mu_g}\mathbb{E}\left[\|y_i^0 - y_{*,i}^0\|^2\right],\ C_{z_i,0} = \frac{5L_f^2}{\mu_g^2}\Big(\frac{L_{\nabla_{22}^2 g}^2}{\mu_g^2} + 1\Big)C_{y_i,0} + \frac{1}{c_2\mu_g}\mathbb{E}\left[\|z_i^0 - z_{*,i}^0\|^2\right].$$

Proof Note that the proof follows almost the same reasoning in Lemma 15. Since Assumptions 1 and 2 hold for all f_i, g_i, by replacing y^k, y_*^k, z^k, z_*^k with $y_i^k, y_{*,i}^k, z_i^k, z_{*,i}^k$ respectively, we have similar results hold for each $1 \leq i \leq n$,

$$\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|y_i^k - y_{*,i}^k\|^2] \leq C_{yx}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|x_+^k - x^k\|^2] + C_{y_i,0} + C_{y,1}\Big(\sum_{k=0}^{K}\alpha_k^2\Big),$$
$$\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|z_i^k - z_{*,i}^k\|^2] \leq C_{zx}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|x_+^k - x^k\|^2] + C_{z_i,0} + C_{z,1}\Big(\sum_{k=0}^{K}\alpha_k^2\Big). \tag{42}$$

Taking summation on both sides of (42), we complete the proof. ■

The next lemma shows that the inequalities above will be used in the error analysis of $\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi(x^k, \lambda^k)\|$.

Lemma 24 *Suppose Assumptions 1, 2 hold for all f_i, g_i and Assumption 3 holds. We have*

$$\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi_{\mu_\lambda}(x^k, \lambda^k)\|^2 \leq \sum_{i=1}^{n}3\lambda_i^k\left\{\big(L_{\nabla f}^2 + L_{\nabla^2 g}^2\big)\|y_i^k - y_{*,i}^k\|^2 + L_{\nabla g}^2\|z_i^k - z_{*,i}^k\|^2\right\},$$
$$\begin{aligned}\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Psi(x^k)\|^2 \leq &\sum_{i=1}^{n}4\lambda_i^k\left\{\big(L_{\nabla f}^2 + L_{\nabla^2 g}^2\big)\|y_i^k - y_{*,i}^k\|^2 + L_{\nabla g}^2\|z_i^k - z_{*,i}^k\|^2\right\}\\ &+ 8nL_\Phi^2\left\{\|\lambda_+^k - \lambda^k\|^2 + \frac{1}{\mu_\lambda^2}\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k, \lambda^k)\|^2\right\}.\end{aligned}$$

Proof Note that we have the following decomposition:

$$\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi_{\mu_\lambda}(x^k, \lambda^k)$$

$$
\begin{aligned}
=&\mathbb{E}[u_x^{k+1}|\mathscr{F}_k] - \sum_{i=1}^n \lambda_i^k \nabla_1 f_i(x^k, y_{*,i}^k) - \sum_{i=1}^n \lambda_i^k \left(\mathbb{E}[J_i^{k+1}|\mathscr{F}_k] z_i^k - \nabla_{12}^2 g_i(x^k, y_{*,i}^k) z_{*,i}^k\right) \\
=&\sum_{i=1}^n \lambda_i^k \Big\{ \nabla_1 f_i(x^k, y_i^k) - \nabla_1 f_i(x^k, y_{*,i}^k) - \nabla_{12}^2 g_i(x^k, y_i^k)\left(z_i^k - z_{*,i}^k\right) \\
&\qquad - \left[\nabla_{12}^2 g_i(x^k, y_i^k) - \nabla_{12}^2 g_i(x^k, y_{*,i}^k)\right] z_{*,i}^k \Big\}. \qquad (43)
\end{aligned}
$$

which, together with Cauchy-Schwarz inequality, implies

$$
\begin{aligned}
&\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1 \Phi_{\mu_\lambda}(x^k, \lambda^k)\|^2 \\
\le& 3\|\sum_{i=1}^n \lambda_i^k \big(\nabla_1 f_i(x^k, y_i^k) - \nabla_1 f_i(x^k, y_{*,i}^k)\big)\|^2 + 3\|\sum_{i=1}^n \lambda_i^k \nabla_{12}^2 g_i(x^k, y_i^k)\big(z_i^k - z_{*,i}^k\big)\|^2 \\
&+ 3\|\sum_{i=1}^n \big(\nabla_{12}^2 g_i(x^k, y_i^k) - \nabla_{12}^2 g_i(x^k, y_{*,i}^k)\big) z_{*,i}^k\|^2 \\
\le& \sum_{i=1}^n 3\lambda_i^k \big(\big(L_{\nabla f}^2 + L_{\nabla^2 g}^2\big)\|y_i^k - y_{*,i}^k\|^2 + L_{\nabla g}^2 \|z_i^k - z_{*,i}^k\|^2\big).
\end{aligned}
$$

Similarly we have $\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Psi(x^k) = \mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi_{\mu_\lambda}(x^k, \lambda^k) + \nabla_1\Phi_{\mu_\lambda}(x^k, \lambda^k) - \nabla_1\Phi_{\mu_\lambda}(x^k, \lambda_*^k)$. Applying Cauchy-Schwarz inequality, Assumption 1 and Lemma 19 to the above equation and (43), we know

$$
\begin{aligned}
&\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Psi(x^k)\|^2 \\
\le& 4\|\sum_{i=1}^n \lambda_i^k \big(\nabla_1 f_i(x^k, y_i^k) - \nabla_1 f_i(x^k, y_{*,i}^k)\big)\|^2 + 4\|\sum_{i=1}^n \lambda_i^k \nabla_{12}^2 g_i(x^k, y_i^k)\big(z_i^k - z_{*,i}^k\big)\|^2 \\
&+ 4\|\sum_{i=1}^n \big(\nabla_{12}^2 g_i(x^k, y_i^k) - \nabla_{12}^2 g_i(x^k, y_{*,i}^k)\big) z_{*,i}^k\|^2 + 4\|\nabla_1\Phi(x^k, \lambda^k) - \nabla_1\Phi(x^k, \lambda_*^k)\|^2 \\
\le& \sum_{i=1}^n 4\lambda_i^k \Big\{ \big(L_{\nabla f}^2 + L_{\nabla^2 g}^2\big)\|y_i^k - y_{*,i}^k\|^2 + L_{\nabla g}^2 \|z_i^k - z_{*,i}^k\|^2 \Big\} + 4nL_\Phi^2 \|\lambda^k - \lambda_*^k\|^2,
\end{aligned}
$$

which together with Lemma 21 completes the proof. ■

7.2.2 Primal Convergence

Lemma 25 *Suppose Assumptions 1, 2 hold for all f_i, g_i and Assumption 3 holds. If*

$$
\begin{aligned}
&\alpha_k \le \min\left(\frac{\tau_x^2}{20c_3},\ \frac{c_3}{2\tau_x\left(c_3 L_{\nabla\Phi} + L_{\nabla\eta_{\mathcal{X}}}\right)},\ \frac{c_3}{4\tau_\lambda(L_{\nabla\eta_{\Delta_n}} + c_3\mu_\lambda)},\ \frac{n\tau_\lambda L_\Phi^2}{L_\Psi + L_{\nabla\Phi}},\ 1\right), \\
&\tau_x < 1,\ \tau_\lambda\mu_\lambda = 1,\ c_3 \le \min\left(\frac{1}{10},\ \frac{1}{8(\mu_\lambda+1)^2}\right), \qquad (44)
\end{aligned}
$$

then in Algorithm 2 we have

$$
\begin{aligned}
&\sum_{k=0}^{K}\frac{\alpha_k}{\tau_x^2}\mathbb{E}[\|x_+^k-x^k\|^2]\\
\leq&\frac{2}{\tau_x}\mathbb{E}\left[\tilde{W}_{0,1}^{(1)}\right]+2\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Psi(x^k)\|^2]\\
&+\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2\right]+\frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|h_x^k-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]\\
&+\sigma_{g,2}^2\sum_{k=0}^{K}\alpha_k^2\mathbb{E}\Big[\sum_{i=1}^{n}\lambda_i^k\|z_i^k-z_{*,i}^k\|^2\Big]+\sigma_w^2\sum_{k=0}^{K}\alpha_k^2,\\
&\sum_{k=0}^{K}\frac{\alpha_k}{\tau_\lambda^2}\mathbb{E}[\|\lambda_+^k-\lambda^k\|^2]\\
\leq&\frac{2}{\tau_\lambda}\mathbb{E}\left[\tilde{W}_{0,1}^{(2)}\right]+\frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]+4L_f^2\sum_{k=0}^{K}\alpha_k\mathbb{E}\Big[\sum_{i=1}^{n}\|y_i^k-y_{*,i}^k\|^2\Big]\\
&+13nL_\Phi^2\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|x_+^k-x^k\|^2]+n\sigma_{f,0}^2\sum_{k=0}^{K}\alpha_k^2. \qquad (45)
\end{aligned}
$$

Proof The proof of the first inequality in (45) is almost the same as that in (17). Note that by replacing $\Phi, h^k, W_{k,1}$ with $\Psi, h_x^k, \tilde{W}_{k,1}$, we know

$$
\begin{aligned}
&\frac{\alpha_k}{\tau_x}\|x_+^k-x^k\|^2\\
\leq&\tilde{W}_{k,1}^{(1)}-\mathbb{E}\big[\tilde{W}_{k+1,1}^{(1)}|\mathscr{F}_k\big]+\alpha_k\big(\tau_x\|\nabla\Psi(x^k)-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2+\frac{1}{4\tau_x}\|x_+^k-x^k\|^2\big)\\
&+\frac{\alpha_k}{4\tau_x}\|x_+^k-x^k\|^2+\frac{5}{2c_3\tau_x}\mathbb{E}\left[\|h_x^{k+1}-h_x^k\|^2|\mathscr{F}_k\right], \qquad (46)
\end{aligned}
$$

Similar to (30), from (40) we have that

$$
\begin{aligned}
&\frac{5}{c_3\tau_x^2}\mathbb{E}[\|h_x^{k+1}-h_x^k\|^2]\\
\leq&\frac{10c_3\alpha_k^2}{\tau_x^2}\mathbb{E}\left[\|h_x^k-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2+\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2\right]+\frac{5c_3\alpha_k^2}{\tau_x^2}\sigma_w^2\\
&+\frac{10c_3\alpha_k^2\sigma_{g,2}^2}{\tau_x^2}\mathbb{E}\Big[\sum_{i=1}^{n}\lambda_i^k\|z_i^k-z_{*,i}^k\|^2\Big].\\
\leq&\frac{\alpha_k}{2}\mathbb{E}[\|h_x^k-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]+\alpha_k\mathbb{E}\left[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2\right]+\alpha_k^2\sigma_w^2\\
&+\alpha_k^2\sigma_{g,2}^2\mathbb{E}\Big[\sum_{i=1}^{n}\lambda_i^k\|z_i^k-z_{*,i}^k\|^2\Big], \qquad (47)
\end{aligned}
$$

where the second inequality uses (44). Taking summation and expectation on both sides of (46) and using (47), we obtain the first inequality in (45). For the second inequality in (45), the $L_{\nabla\Psi}$-smoothness of $\Psi(x)$ and $L_{\nabla\eta_{\mathcal{X}}}$-smoothness of $\eta_{\mathcal{X}}$ in Lemma 19 imply

$$\begin{aligned}
&\Psi(x^{k+1}) - \Psi(x^k) \le \alpha_k \left\langle \nabla\Psi(x^k), x_+^k - x^k \right\rangle + \frac{L_{\nabla\Psi}}{2}\|x^{k+1} - x^k\|^2, \qquad (48)\\
&\eta_{\Delta_n}(\lambda^k, -h_\lambda^k, \tau_\lambda) - \eta_{\Delta_n}(\lambda^{k+1}, -h_\lambda^{k+1}, \tau_\lambda)\\
\le& \left\langle h_\lambda^k + \frac{1}{\tau_\lambda}(\lambda^k - \lambda_+^k), \lambda^k - \lambda^{k+1} \right\rangle + \left\langle \lambda_+^k - \lambda^k, -h_\lambda^k + h_\lambda^{k+1} \right\rangle\\
&+ \frac{L_{\nabla\eta_{\Delta_n}}}{2}\left(\|\lambda^{k+1} - \lambda^k\|^2 + \| - h_\lambda^{k+1} + h_\lambda^k\|^2\right)\\
=&\alpha_k \left\langle -h_\lambda^k, \lambda_+^k - \lambda^k \right\rangle + \frac{\alpha_k}{\tau_\lambda}\|\lambda_+^k - \lambda^k\|^2 + \theta_k \left\langle \lambda_+^k - \lambda^k, s^{k+1} - h_\lambda^k - \mu_\lambda\big(\lambda^k - \frac{\mathbf{1}_n}{n}\big) \right\rangle\\
&+ \frac{L_{\nabla\eta_{\Delta_n}}}{2}\left(\|\lambda^{k+1} - \lambda^k\|^2 + \|h_\lambda^{k+1} - h_\lambda^k\|^2\right)\\
\le& -\frac{\theta_k}{\tau_\lambda}\|\lambda_+^k - \lambda^k\|^2 + \theta_k \left\langle s^{k+1} - \mu_\lambda\big(\lambda^k - \frac{\mathbf{1}_n}{n}\big), \lambda_+^k - \lambda^k \right\rangle\\
&+ \frac{L_{\nabla\eta_{\Delta_n}}}{2}\left(\|\lambda^{k+1} - \lambda^k\|^2 + \|h_\lambda^{k+1} - h_\lambda^k\|^2\right). \qquad (49)
\end{aligned}$$

We also have

$$\begin{aligned}
&\Phi_{\mu_\lambda}(x^k, \lambda^k) - \Phi_{\mu_\lambda}(x^{k+1}, \lambda^{k+1})\\
=&\sum_{i=1}^n \left(\lambda_i^k \Phi_i(x^k) - \lambda_i^{k+1}\Phi_i(x^{k+1})\right) + \frac{\mu_\lambda}{2}\|\lambda^{k+1} - \frac{\mathbf{1}_n}{n}\|^2 - \frac{\mu_\lambda}{2}\|\lambda^k - \frac{\mathbf{1}_n}{n}\|^2\\
=&\left\langle \lambda^k, \Phi^k \right\rangle - \left\langle \lambda^{k+1}, \Phi^{k+1} \right\rangle + \frac{\mu_\lambda}{2}\left(\|\lambda^{k+1} - \lambda^k + \lambda^k - \frac{\mathbf{1}_n}{n}\|^2 - \|\lambda^k - \frac{\mathbf{1}_n}{n}\|^2\right)\\
=&\left\langle \lambda^k - \lambda^{k+1}, \Phi^k \right\rangle + \left\langle \lambda^{k+1}, \Phi^k - \Phi^{k+1} \right\rangle + \mu_\lambda\alpha_k \left\langle \lambda^k - \frac{\mathbf{1}_n}{n}, \lambda_+^k - \lambda^k \right\rangle + \frac{\mu_\lambda}{2}\|\lambda^{k+1} - \lambda^k\|^2\\
=&\alpha_k \left\langle \lambda^k - \lambda_+^k, \mathbb{E}[s^{k+1}|\mathscr{F}_k] - \mu_\lambda\big(\lambda^k - \frac{\mathbf{1}_n}{n}\big) \right\rangle + \alpha_k \left\langle \lambda^k - \lambda_+^k, \Phi^k - \mathbb{E}[s^{k+1}|\mathscr{F}_k] \right\rangle\\
&+ \frac{\mu_\lambda}{2}\|\lambda^{k+1} - \lambda^k\|^2 + \left\langle \lambda^{k+1}, \Phi^k - \Phi^{k+1} \right\rangle\\
\le&\alpha_k \left\langle \lambda^k - \lambda_+^k, \mathbb{E}[s^{k+1}|\mathscr{F}_k] - \mu_\lambda\big(\lambda^k - \frac{\mathbf{1}_n}{n}\big) \right\rangle + \alpha_k \left\langle \lambda^k - \lambda_+^k, \Phi^k - \mathbb{E}[s^{k+1}|\mathscr{F}_k] \right\rangle\\
&+ \frac{\mu_\lambda}{2}\|\lambda^{k+1} - \lambda^k\|^2 - \alpha_k \left\langle \nabla_1\Phi(x^k, \lambda^k), x_+^k - x^k \right\rangle + \sqrt{n}L_\Phi\|\lambda^{k+1} - \lambda^k\|\|x_+^k - x^k\|\\
&+ \frac{L_{\nabla\Phi}}{2}\|x^{k+1} - x^k\|^2. \qquad (50)
\end{aligned}$$

where the inequality uses Lemma 19 and (c) in Assumption 1 to obtain

$$\left\langle \lambda^{k+1}, \Phi^k - \Phi^{k+1} \right\rangle = \sum_{i=1}^n \lambda_i^{k+1}(\Phi_i(x^k) - \Phi_i(x^{k+1}))$$

$$\leq \sum_{i=1}^{n} \lambda_i^{k+1} \big(\left\langle \nabla \Phi_i(x^k), x^k - x^{k+1} \right\rangle + \frac{L_{\nabla\Phi}}{2} \|x^k - x^{k+1}\|^2 \big)$$
$$\leq -\alpha_k \left\langle \nabla_1 \Phi(x^k, \lambda^k), x_+^k - x^k \right\rangle + \sqrt{n} L_\Phi \|\lambda^{k+1} - \lambda^k\| \|x_+^k - x^k\| + \frac{L_{\nabla\Phi}}{2} \|x^{k+1} - x^k\|^2.$$

Taking conditional expectation with respect to $\mathscr{F}_k$ on (48) + (49)/c_3 + (50), we know

$$\begin{aligned}
&\frac{\alpha_k}{\tau_\lambda} \|\lambda_+^k - \lambda^k\|^2 \\
\leq& \tilde{W}_{k,1}^{(2)} - \mathbb{E}\left[\tilde{W}_{k+1,1}^{(2)} | \mathscr{F}_k\right] + \alpha_k \left\langle \nabla\Psi(x^k) - \nabla_1 \Phi(x^k, \lambda^k), x_+^k - x^k \right\rangle \\
&+ \alpha_k \left\langle \lambda^k - \lambda_+^k, \Phi^k - \mathbb{E}[s^{k+1} | \mathscr{F}_k] \right\rangle + \frac{(L_{\nabla\Psi} + L_{\nabla\Phi})}{2} \|x^{k+1} - x^k\|^2 \\
&+ \frac{(L_{\nabla\eta_{\Delta_n}} + c_3\mu_\lambda)}{2c_3} \|\lambda^{k+1} - \lambda^k\|^2 + \sqrt{n} L_\Phi \|\lambda^{k+1} - \lambda^k\| \|x_+^k - x^k\| + \frac{L_{\nabla\eta_{\Delta_n}}}{2c_3} \mathbb{E}\left[\|h_\lambda^{k+1} - h_\lambda^k\|^2 | \mathscr{F}_k\right] \\
\leq& \tilde{W}_{k,1}^{(2)} - \mathbb{E}\left[\tilde{W}_{k+1,1}^{(2)} | \mathscr{F}_k\right] + \alpha_k \sqrt{n} L_\Phi \|\lambda^k - \lambda_*^k\| \|x_+^k - x^k\| \\
&+ \alpha_k L_f \|\lambda_+^k - \lambda^k\| \left(\sum_{i=1}^{n} \|y_i^k - y_{*,i}^k\|^2 \right)^{\frac{1}{2}} + \alpha_k \sqrt{n} L_\Phi \|\lambda_+^k - \lambda^k\| \|x_+^k - x^k\| \\
&+ \frac{\alpha_k^2 (L_{\nabla\Psi} + L_{\nabla\Phi})}{2} \|x_+^k - x^k\|^2 + \frac{\alpha_k^2 (L_{\nabla\eta_{\Delta_n}} + c_3\mu_\lambda)}{2c_3} \|\lambda_+^k - \lambda^k\|^2 + \frac{L_{\nabla\eta_{\Delta_n}}}{2c_3} \mathbb{E}\left[\|h_\lambda^{k+1} - h_\lambda^k\|^2 | \mathscr{F}_k\right] \\
\leq& \tilde{W}_{k,1}^{(2)} - \mathbb{E}\left[\tilde{W}_{k+1,1}^{(2)} | \mathscr{F}_k\right] + \alpha_k \left(\frac{1}{16\tau_\lambda} \|\lambda^k - \lambda_*^k\|^2 + 4n\tau_\lambda L_\Phi^2 \|x_+^k - x^k\|^2 \right) \\
&+ \alpha_k \left(\frac{1}{8\tau_\lambda} \|\lambda_+^k - \lambda^k\|^2 + 2\tau_\lambda L_f^2 \sum_{i=1}^{n} \|y_i^k - y_{*,i}^k\|^2 \right) + \alpha_k \left(\frac{1}{8\tau_\lambda} \|\lambda_+^k - \lambda^k\|^2 + 2n\tau_\lambda L_\Phi^2 \|x_+^k - x^k\|^2 \right) \\
&+ \frac{\alpha_k n \tau_\lambda L_\Phi^2}{2} \|x_+^k - x^k\|^2 + \frac{\alpha_k}{8\tau_\lambda} \|\lambda_+^k - \lambda^k\|^2 + \frac{L_{\nabla\eta_{\Delta_n}}}{2c_3} \mathbb{E}\left[\|h_\lambda^{k+1} - h_\lambda^k\|^2 | \mathscr{F}_k\right], \qquad (51)
\end{aligned}$$

where the second inequality uses Lemma 19, and the third inequality uses Young's inequality and the conditions on α_k (see (44)): $\frac{\alpha_k}{8\tau_\lambda} - \frac{\alpha_k^2 (L_{\nabla\eta_{\Delta_n}} + c_3\mu_\lambda)}{2c_3} \geq 0,\ \alpha_k^2 (L_{\nabla\Psi} + L_{\nabla\Phi}) \leq \alpha_k n \tau_\lambda L_\Phi^2$. Recall that in Lemma 21 we have

$$\|\lambda^k - \lambda_*^k\|^2 \leq 2\|\lambda_+^k - \lambda^k\|^2 + \frac{2}{\mu_\lambda^2} \|h_\lambda^k - \nabla_2 \Phi_{\mu_\lambda}(x^k, \lambda^k)\|^2, \qquad (52)$$

and by (40) we know

$$\begin{aligned}
\frac{L_{\nabla\eta_{\Delta_n}}}{c_3\tau_\lambda} \mathbb{E}[\|h_\lambda^{k+1} - h_\lambda^k\|^2] &\leq 2c_3 \alpha_k^2 (\mu_\lambda + 1)^2 \left(\mathbb{E}[\|h_\lambda^k - \nabla_2 \Phi_{\mu_\lambda}(x^k, \lambda^k)\|^2] + n\sigma_{f,0}^2 \right) \\
&\leq \frac{\alpha_k}{4} \mathbb{E}[\|h_\lambda^k - \nabla_2 \Phi_{\mu_\lambda}(x^k, \lambda^k)\|^2] + n\alpha_k^2 \sigma_{f,0}^2. \qquad (53)
\end{aligned}$$

where the second inequality uses $2c_3(\mu_\lambda + 1)^2 \leq \frac{1}{4},\ \alpha_k \leq 1$ in (44). By (51), (52), and (53):

$$\begin{aligned}
\frac{\alpha_k}{\tau_\lambda^2} \mathbb{E}[\|\lambda_+^k - \lambda^k\|^2] \leq& \frac{2}{\tau_\lambda} \mathbb{E}\left[\tilde{W}_{k,1}^{(2)} - \tilde{W}_{k+1,1}^{(2)}\right] + \frac{\alpha_k}{2} \mathbb{E}[\|h_\lambda^k - \nabla_2 \Phi_{\mu_\lambda}(x^k, \lambda^k)\|^2] \\
&+ 4\alpha_k L_f^2 \mathbb{E}\left[\sum_{i=1}^{n} \|y_i^k - y_{*,i}^k\|^2 \right] + 13\alpha_k n L_\Phi^2 \mathbb{E}[\|x_+^k - x^k\|^2] + n\alpha_k^2 \sigma_{f,0}^2,
\end{aligned}$$

which implies the second inequality in (45) by taking summation. ■

7.2.3 Dual Convergence

Lemma 26 *Suppose Assumptions 1, 2 hold for all f_i, g_i and Assumption 3 holds. In Algorithm 2 we have*

$$\begin{aligned}
&\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|h_x^k-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]\\
\le&\frac{1}{c_3}\mathbb{E}\left[\|h_x^0-\nabla_1\Phi_{\mu_\lambda}(x^0,\lambda^0)\|^2\right]+3\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2\right]\\
&+\frac{3L_{\nabla\Phi}^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|x_+^k-x^k\|^2]+\frac{3nL_\Phi^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|\lambda_+^k-\lambda^k\|^2]\\
&+2c_3\sigma_{g,2}^2\sum_{k=0}^{K}\alpha_k^2\mathbb{E}\left[\sum_{i=1}^{n}\lambda_i^k\|z_i^k-z_{*,i}^k\|^2\right]+c_3\sigma_w^2\sum_{k=0}^{K}\alpha_k^2,\\
&\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]\\
\le&\frac{1}{c_3}\mathbb{E}\left[\|h_\lambda^0-\nabla_2\Phi_{\mu_\lambda}(x^0,\lambda^0)\|^2\right]+3\alpha_kL_f^2\sum_{i=1}^{n}\mathbb{E}[\|y_i^k-y_{*,i}^k\|^2]\\
&+\frac{3nL_\Phi^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|x_+^k-x^k\|^2]+\frac{3\mu_\lambda^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|\lambda_+^k-\lambda^k\|^2]+nc_3\sigma_{f,0}^2\sum_{k=0}^{K}\alpha_k^2. \qquad (54)
\end{aligned}$$

Proof The proof is similar to that of Lemma 18, except that we now have another λ^k to handle. Since $\nabla_1\Phi(x,\lambda)=\nabla_1\Phi_{\mu_\lambda}(x,\lambda)$ for all (x,λ) (see (10)), for simplicity we omit subscript μ_λ in $\nabla_1\Phi_{\mu_\lambda}(x,\lambda)$ in proof. Note that by moving-average update of h_x^k, we have

$$\begin{aligned}
&h_x^{k+1}-\nabla_1\Phi(x^{k+1},\lambda^{k+1})\\
=&(1-\theta_k)h_x^k+\theta_k(w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k])+\theta_k\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi(x^{k+1},\lambda^{k+1})\\
=&(1-\theta_k)(h_x^k-\nabla_1\Phi(x^k,\lambda^k))+\theta_k(\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi(x^k,\lambda^k))\\
&+\nabla_1\Phi(x^k,\lambda^k)-\nabla_1\Phi(x^{k+1},\lambda^{k+1})+\theta_k(w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k])
\end{aligned}$$

Hence we know

$$\begin{aligned}
&\mathbb{E}\left[\|h_x^{k+1}-\nabla_1\Phi(x^{k+1},\lambda^{k+1})\|^2|\mathscr{F}_k\right]\\
=&\Big\|(1-\theta_k)(h_x^k-\nabla_1\Phi(x^k,\lambda^k))+\theta_k(\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi(x^k,\lambda^k))\\
&+\nabla_1\Phi(x^k,\lambda^k)-\nabla_1\Phi(x^{k+1},\lambda^{k+1})\Big\|^2+\theta_k^2\mathbb{E}\left[\|w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k\right]\\
\le&(1-\theta_k)\|h_x^k-\nabla_1\Phi(x^k,\lambda^k)\|^2+\theta_k^2\sigma_{w,k+1}^2
\end{aligned}$$

$$
\begin{aligned}
&+\theta_k\|(\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi(x^k,\lambda^k))+\frac{1}{\theta_k}(\nabla_1\Phi(x^k,\lambda^k)-\nabla_1\Phi(x^{k+1},\lambda^{k+1}))\|^2\\
\leq&(1-\theta_k)\|h_x^k-\nabla_1\Phi(x^k,\lambda^k)\|^2+3\theta_k\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi(x^k,\lambda^k)\|^2+\theta_k^2\sigma_{w,k+1}^2\\
&+\frac{3}{\theta_k}\|\nabla_1\Phi(x^k,\lambda^k)-\nabla_1\Phi(x^{k+1},\lambda^k)\|^2+\frac{3}{\theta_k}\|\nabla_1\Phi(x^{k+1},\lambda^k)-\nabla_1\Phi(x^{k+1},\lambda^{k+1})\|^2\\
\leq&(1-\theta_k)\|h_x^k-\nabla_1\Phi(x^k,\lambda^k)\|^2+3\theta_k\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi(x^k,\lambda^k)\|^2\\
&+\frac{3\alpha_k^2}{\theta_k}\left(L_{\nabla\Phi}^2\|x_+^k-x^k\|^2+nL_\Phi^2\|\lambda_+^k-\lambda^k\|^2\right)+\theta_k^2\sigma_{w,k+1}^2,
\end{aligned}
\tag{55}
$$

where the first equality uses the fact that $x^k,\lambda^k,h_x^k,x^{k+1},\lambda^{k+1}$ are all $\mathscr{F}_k$-measurable and are independent of w^{k+1} given $\mathscr{F}_k$, the first inequality uses the convexity of $\|\cdot\|^2$ and (39), the second inequality uses Cauchy-Schwarz inequality, the third inequality uses the Lipschitz continuity of $\nabla_1\Phi$ in Lemma 19, and the update rules of x^{k+1} and λ^{k+1}. Taking summation, expectation on both sides of (55), dividing c_3, and applying (39), we know the first inequality in (54) holds. Similarly we have

$$
\begin{aligned}
&h_\lambda^{k+1}-\nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1})\\
=&(1-\theta_k)h_\lambda^k+\theta_k\big(s^{k+1}-\mu_\lambda\lambda^k+\mu_\lambda\frac{\mathbf{1}_n}{n}\big)-\nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1})\\
=&(1-\theta_k)(h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k))+\theta\left(\mathbb{E}[s^{k+1}|\mathscr{F}_k]-\nabla_2\Phi(x^k,\lambda^k)\right)\\
&+\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)-\nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1})+\theta_k(s^{k+1}-\mathbb{E}[s^{k+1}|\mathscr{F}_k]).
\end{aligned}
$$

where the second equality uses $\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)=\nabla_2\Phi(x^k,\lambda^k)-\mu_\lambda\big(\lambda^k-\frac{\mathbf{1}_n}{n}\big)$. So we know

$$
\begin{aligned}
&\mathbb{E}\left[\|h_\lambda^{k+1}-\nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1})\|^2|\mathscr{F}_k\right]\\
=&\left\|(1-\theta_k)(h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k))+\theta\left(\mathbb{E}[s^{k+1}|\mathscr{F}_k]-\nabla_2\Phi(x^k,\lambda^k)\right)\right.\\
&\left.+\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)-\nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1})\right\|^2+\theta_k^2\mathbb{E}\left[\|s^{k+1}-\mathbb{E}[s^{k+1}|\mathscr{F}_k]\|^2|\mathscr{F}_k\right]\\
\leq&(1-\theta_k)\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2+n\theta_k^2\sigma_{f,0}^2\\
&+\theta_k\|\mathbb{E}[s^{k+1}|\mathscr{F}_k]-\nabla_2\Phi(x^k,\lambda^k)+\frac{1}{\theta_k}(\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)-\nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1}))\|^2\\
\leq&(1-\theta_k)\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2+3\theta_k\|\mathbb{E}[s^{k+1}|\mathscr{F}_k]-\nabla_2\Phi(x^k,\lambda^k)\|^2+n\theta_k^2\sigma_{f,0}^2\\
&+\frac{3}{\theta_k}(\|\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)-\nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^k)\|^2+\|\nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^k)-\nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1})\|^2)\\
\leq&(1-\theta_k)\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2+3\theta_kL_f^2\sum_{i=1}^n\|y_i^k-y_{*,i}^k\|^2\\
&+\frac{3\alpha_k^2}{\theta_k}\left(nL_\Phi^2\|x_+^k-x^k\|^2+\mu_\lambda^2\|\lambda_+^k-\lambda^k\|^2\right)+n\theta_k^2\sigma_{f,0}^2,
\end{aligned}
\tag{56}
$$

where the third inequality uses Lemma 19 and the fact that

$$
\mathbb{E}[s^{k+1}|\mathscr{F}_k]=\Big(f_1(x^k,y_1^k),...,f_n(x^k,y_n^k)\Big)^\top,\nabla_2\Phi(x^k,\lambda^k)=\Big(f_1(x^k,y_{*,1}^k),...,f_n(x^k,y_{*,n}^k)\Big)^\top
$$

Taking summation, expectation on both sides of (56), and dividing c_3, we know the second inequality in (54) holds. ■

7.2.4 Proof of Theorem 6 and Corollary 7

Now we are ready to present our main convergence results. Note that by Lemmas (25) and (26), for $\tilde{V}_{k,1}$ we have

$$\begin{aligned}
&\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,1}\big] = \sum_{k=0}^{K}\frac{\alpha_k}{\tau_x^2}\mathbb{E}[\|x_+^k - x^k\|^2] + \sum_{k=0}^{K}\alpha_k\mathbb{E}[\|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]\\
\leq&\frac{3L_{\nabla\Phi}^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|x_+^k - x^k\|^2] + \frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]\\
&+\frac{2}{\tau_x}\mathbb{E}\left[\tilde{W}_{0,1}^{(1)}\right] + \frac{1}{c_3}\mathbb{E}\left[\|h_x^0 - \nabla_1\Phi_{\mu_\lambda}(x^0,\lambda^0)\|^2\right] + 2\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Psi(x^k)\|^2]\\
&+4\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2\big] + \frac{3nL_\Phi^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|\lambda_+^k - \lambda^k\|^2]\\
&+(1+2c_3)\,\sigma_{g,2}^2\sum_{k=0}^{K}\alpha_k^2\mathbb{E}\bigg[\sum_{i=1}^{n}\lambda_i^k\|z_i^k - z_{*,i}^k\|^2\bigg] + (1+c_3)\,\sigma_w^2\bigg(\sum_{k=0}^{K}\alpha_k^2\bigg).
\end{aligned}\tag{57}$$

By Lemma 24 we know

$$\begin{aligned}
&4\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2\right] + 2\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Psi(x^k)\|^2]\\
\leq&\sum_{k=0}^{K}\alpha_k\mathbb{E}\bigg[\sum_{i=1}^{n}20\left(\left(L_{\nabla f}^2 + L_{\nabla^2 g}^2\right)\|y_i^k - y_{*,i}^k\|^2 + L_{\nabla g}^2\|z_i^k - z_{*,i}^k\|^2\right)\bigg]\\
&+\sum_{k=0}^{K}16nL_\Phi^2\alpha_k\mathbb{E}\big[\|\lambda_+^k - \lambda^k\|^2 + \frac{1}{\mu_\lambda^2}\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2\big].
\end{aligned}\tag{58}$$

Choosing

$$(1+2c_3)\,\sigma_{g,2}^2\alpha_k \leq L_{\nabla g}^2 \tag{59}$$

in (57), and using (58), we know

$$\begin{aligned}
\sum_{k=0}^{K}\alpha_k\mathbb{E}[\tilde{V}_{k,1}] \leq& C_{v_1,x}\tau_x^2\sum_{k=0}^{K}\frac{\alpha_k}{\tau_x^2}\mathbb{E}[\|x_+^k - x^k\|^2] + C_{v_1,h_x}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]\\
&+C_{v_1,\lambda}\tau_\lambda^2\sum_{k=0}^{K}\frac{\alpha_k}{\tau_\lambda^2}\mathbb{E}[\|\lambda_+^k - \lambda^k\|^2] + C_{v_1,h_\lambda}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]\\
&+C_{v_1,0} + C_{v_1,1}\bigg(\sum_{k=0}^{K}\alpha_k^2\bigg),
\end{aligned}\tag{60}$$

where the constants are defined as

$$
\begin{aligned}
C_{v_1,x} =& 20n\big(L_{\nabla f}^2 + L_{\nabla^2 g}^2\big)C_{yx} + 21nL_{\nabla g}^2 C_{zx} + \frac{3L_{\nabla\Phi}^2}{c_3^2},\ C_{v_1,h_x} = \frac{1}{2},\\
C_{v_1,\lambda} =& \left(16 + \frac{3}{c_3^2}\right) nL_\Phi^2,\ C_{v_1,h_\lambda} = \frac{16nL_\Phi^2}{\mu_\lambda^2},\\
C_{v_1,0} =& 20\big(L_{\nabla f}^2 + L_{\nabla^2 g}^2\big)\left(\sum_{i=1}^n C_{y_i,0}\right) + 21L_{\nabla g}^2\left(\sum_{i=1}^n C_{z_i,0}\right) + \frac{2}{\tau_x}\mathbb{E}\left[\tilde{W}_{0,1}^{(1)}\right]\\
&+ \frac{1}{c_3}\mathbb{E}\left[\|h_x^0 - \nabla_1\Phi_{\mu_\lambda}(x^0,\lambda^0)\|^2\right],\\
C_{v_1,1} =& 20n\big(L_{\nabla f}^2 + L_{\nabla^2 g}^2\big)C_{y,1} + 21nL_{\nabla g}^2 C_{z,1}.
\end{aligned}
$$

For $\tilde{V}_{k,2}$ we have

$$
\begin{aligned}
&\sum_{k=0}^K \alpha_k\mathbb{E}\big[\tilde{V}_{k,2}\big] = \sum_{k=0}^K \frac{\alpha_k}{\tau_\lambda^2}\mathbb{E}[\|\lambda_+^k - \lambda^k\|^2] + \sum_{k=0}^K \alpha_k\mathbb{E}[\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]\\
\le&\frac{3\mu_\lambda^2}{c_3^2}\sum_{k=0}^K \alpha_k\mathbb{E}[\|\lambda_+^k - \lambda^k\|^2] + \frac{1}{2}\sum_{k=0}^K \alpha_k\mathbb{E}[\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]\\
&+ \frac{2}{\tau_\lambda}\mathbb{E}\left[\tilde{W}_{0,1}^{(2)}\right] + \frac{1}{c_3}\mathbb{E}\left[\|h_\lambda^0 - \nabla_2\Phi_{\mu_\lambda}(x^0,\lambda^0)\|^2\right] + 7L_f^2\sum_{k=0}^K \alpha_k\mathbb{E}\left[\sum_{i=1}^n \|y_i^k - y_{*,i}^k\|^2\right]\\
&+ \left(13 + \frac{3}{c_3^2}\right) nL_\Phi^2\sum_{k=0}^K \alpha_k\mathbb{E}[\|x_+^k - x^k\|^2] + n(1+c_3)\sigma_{f,0}^2\left(\sum_{k=0}^K \alpha_k^2\right),
\end{aligned}
$$

which implies

$$
\begin{aligned}
\sum_{k=0}^K \alpha_k\mathbb{E}\big[\tilde{V}_{k,2}\big] \le& C_{v_2,x}\tau_x^2\sum_{k=0}^K \frac{\alpha_k}{\tau_x^2}\mathbb{E}[\|x_+^k - x^k\|^2] + C_{v_2,h_x}\sum_{k=0}^K \alpha_k\mathbb{E}[\|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]\\
&+ C_{v_2,\lambda}\tau_\lambda^2\sum_{k=0}^K \frac{\alpha_k}{\tau_\lambda^2}\mathbb{E}[\|\lambda_+^k - \lambda^k\|^2] + C_{v_2,h_\lambda}\sum_{k=0}^K \alpha_k\mathbb{E}[\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\|^2]\\
&+ C_{v_2,0} + C_{v_2,1}\left(\sum_{k=0}^K \alpha_k^2\right) \qquad (61)
\end{aligned}
$$

where the constants are defined as

$$
\begin{aligned}
C_{v_2,x} =& 7nL_f^2 C_{yx} + \left(13 + \frac{3}{c_3^2}\right) nL_\Phi^2,\ C_{v_2,h_x} = 0,\ C_{v_2,\lambda} = \frac{3\mu_\lambda^2}{c_3^2},\ C_{v_2,h_\lambda} = \frac{1}{2},\\
C_{v_2,0} =& 7L_f^2\left(\sum_{i=1}^n C_{y_i,0}\right) + \frac{2}{\tau_\lambda}\mathbb{E}\left[\tilde{W}_{0,1}^{(2)}\right] + \frac{1}{c_3}\mathbb{E}\left[\|h_\lambda^0 - \nabla_2\Phi_{\mu_\lambda}(x^0,\lambda^0)\|^2\right]\\
C_{v_2,1} =& 7nL_f^2 C_{y,1} + n(1+c_3)\sigma_{f,0}^2.
\end{aligned}
$$

According to the definition of the constants in Lemmas 15 and 23, we could obtain (for simplicity we omit the dependency on κ here)

$$C_{v_1,x} = \mathcal{O}\left(\frac{n}{c_1^2} + \frac{n}{c_2^2} + \frac{1}{c_3^2}\right), C_{v_1,h_x} = \frac{1}{2} = \mathcal{O}(1), C_{v_1,\lambda} = \mathcal{O}\left(n + \frac{n}{c_3^2}\right), C_{v_1,h_\lambda} = \mathcal{O}\left(\frac{n}{\mu_\lambda^2}\right),$$
$$C_{v_1,0} = \mathcal{O}\left(\frac{n}{c_1} + \frac{n}{c_2} + \frac{1}{c_3} + \frac{1}{\tau_x} + \frac{1}{c_3\tau_x}\right), \ C_{v_1,1} = \mathcal{O}(nc_1 + nc_2),$$
$$C_{v_2,x} = \mathcal{O}\left(\frac{n}{c_1^2} + n + \frac{n}{c_3^2}\right), \ C_{v_2,h_x} = 0, \ C_{v_2,\lambda} = \mathcal{O}\left(\frac{1}{c_3^2}\right), \ C_{v_2,h_\lambda} = \frac{1}{2} = \mathcal{O}(1),$$
$$C_{v_2,0} = \mathcal{O}\left(\frac{n}{c_1} + \frac{1}{c_3}\right), \ C_{v_2,1} = \mathcal{O}(nc_1 + n + nc_3).$$

Hence we can pick $\alpha_k, c_1, c_2, c_3, \tau_x, \tau_\lambda$ such that $\alpha_k \equiv \Theta(1/\sqrt{nK})$, $c_1 = c_2 = \sqrt{n}$, $c_3 = \Theta(1)$, $\tau_x = \mathcal{O}(\mu_\lambda/n)$, $\tau_\lambda = 1/\mu_\lambda$, which leads to $C_{v_1,x}\tau_x^2 \le \frac{1}{2}$, $C_{v_2,x}C_{v_1,\lambda}\tau_x^2\tau_\lambda^2 \le \frac{1}{8}$, $C_{v_2,\lambda}\tau_\lambda^2 \le \frac{1}{2}$, and the conditions ((41), (44), and (59)) in previous lemmas hold. Moreover, using the above conditions in (60) and (61), we can get

$$\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,1}\big] \le \frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,1}\big] + C_{v_1,\lambda}\tau_\lambda^2\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,2}\big] + \mathcal{O}(n)$$
$$\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,2}\big] \le \frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,2}\big] + C_{v_2,x}\tau_x^2\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,1}\big] + \mathcal{O}\left(\sqrt{n}\right).$$

Combining the above two inequalities, we have $\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\big[\tilde{V}_{k,1}\big] = \mathcal{O}(n^2/\mu_\lambda^2\sqrt{K})$ and $\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\big[\tilde{V}_{k,2}\big] = \mathcal{O}(n/\sqrt{K})$, which completes the proof of Theorem 6 since we have

$$\begin{aligned}
&\frac{1}{\tau_x^2}\mathbb{E}[\|x^k - \Pi_{\mathcal{X}}\big(x^k - \tau_x\nabla\Psi_{\mu_\lambda}(x^k)\big)\|^2] \\
\le&\frac{2}{\tau_x^2}\mathbb{E}[\|x^k - \Pi_{\mathcal{X}}\big(x^k - \tau_x\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\big)\|^2] \\
&+ \frac{2}{\tau_x^2}\mathbb{E}[\|\Pi_{\mathcal{X}}\big(x^k - \tau_x\nabla\Psi_{\mu_\lambda}(x^k)\big) - \Pi_{\mathcal{X}}\big(x^k - \tau_x\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\big)\|^2] \\
\le&\frac{2}{\tau_x^2}\mathbb{E}[\|x^k - \Pi_{\mathcal{X}}\big(x^k - \tau_x\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\big)\|^2] + 2nL_\Phi^2\mathbb{E}[\|\lambda^k - \lambda_*^k\|^2] \le 4\mathbb{E}[\tilde{V}_{k,1}] + \frac{4nL_\Phi^2}{\mu_\lambda^2}\mathbb{E}[\tilde{V}_{k,2}]
\end{aligned}$$

where the second inequality uses non-expansiveness of projection operator and $\sqrt{n}L_\Phi$-Lipschitz continuity of $\nabla_1\Phi_{\mu_\lambda}(x,\cdot)$ in Lemma 19. Note that we have n^2 in the numerator since we explicitly write out the Lipschitz constant $L_{\nabla_1\Phi_{\mu_\lambda}}$.

To prove Corollary 7, we notice that by choosing $\mu_\lambda = \mathcal{O}(\sqrt{\epsilon})$, we have $\|\nabla\Phi_{\mu_\lambda}(x,\lambda) - \nabla\Phi(x,\lambda)\|^2 \le \mu_\lambda^2\|\lambda - \frac{\mathbf{1}_n}{n}\|^2 \le \mu_\lambda^2 = \mathcal{O}(\epsilon)$, and thus under the same setup of Theorem 6, we know from Section D.2 of Lin et al. (2020b) that any ϵ-stationary point of Problem (10) is an ϵ-stationary point of Problem (9). Hence the corresponding sample complexity is $\mathcal{O}(n^5\epsilon^{-4})$.

References

Zeeshan Akhtar, Amrit Singh Bedi, Srujan Teja Thomdapu, and Ketan Rajawat. Projection-free stochastic bi-level optimization. *IEEE Transactions on Signal Processing*, 70:6332–6347, 2022.

Michael Arbel and Julien Mairal. Amortized implicit differentiation for stochastic bilevel optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=3PN4iyXBeF.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299. PMLR, 2020.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.

Krishnakumar Balasubramanian, Saeed Ghadimi, and Anthony Nguyen. Stochastic multilevel composition optimization algorithms with level-independent convergence rates. *SIAM Journal on Optimization*, 32(2):519–544, 2022.

Yoshua Bengio. Gradient-based optimization of hyperparameters. *Neural computation*, 12 (8):1889–1900, 2000.

Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxnZh0ct7.

Quentin Bertrand, Quentin Klopfenstein, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization. In *International Conference on Machine Learning*, pages 810–821. PMLR, 2020.

Jerome Bracken and James T McGill. Mathematical programs with optimization problems in the constraints. *Operations research*, 21(1):37–44, 1973.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International conference on machine learning*, pages 654–663. PMLR, 2017.

Lesi Chen, Jing Xu, and Jingzhao Zhang. On bilevel optimization without lower-level strong convexity. *arXiv preprint arXiv:2301.00712*, 2023a.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34, 2021a.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021b.

Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2466–2488. PMLR, 2022.

Xuxing Chen, Minhui Huang, Shiqian Ma, and Krishna Balasubramanian. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *International Conference on Machine Learning*, pages 4641–4671. PMLR, 2023b.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. *Advances in neural information processing systems*, 32, 2019.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=wlEOsQ917F`.

Mathieu Dagréou, Thomas Moreau, Samuel Vaiter, and Pierre Ablin. A lower bound and a near-optimal algorithm for bilevel empirical risk minimization. *arXiv e-prints*, pages arXiv–2302, 2023.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pages 1165–1173. PMLR, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018.

Hongchang Gao, Xiaoqian Wang, Lei Luo, and Xinghua Shi. On the convergence of stochastic compositional gradient descent ascent method. In *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

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

Saeed Ghadimi, Andrzej Ruszczynski, and Mengdi Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020.

Tommaso Giovannelli, Griffin Kent, and Luis Nunes Vicente. Inexact bilevel stochastic gradient methods for constrained and unconstrained lower-level problems. *arXiv preprint arXiv:2110.00604*, 2021.

Tommaso Giovannelli, Griffin Kent, and Luis Nunes Vicente. Bilevel optimization with a multi-objective lower-level problem: Risk-neutral and risk-averse formulations. *arXiv preprint arXiv:2302.05540*, 2023.

Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. On differentiating parameterized Argmin and Argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.

Riccardo Grazzi, Massimiliano Pontil, and Saverio Salzo. Bilevel optimization with a lower-level contraction: Optimal sample complexity without warm-start. *Journal of Machine Learning Research*, 24(167):1–37, 2023. URL `http://jmlr.org/papers/v24/22-1043.html`.

Alex Gu, Songtao Lu, Parikshit Ram, and Tsui-Wei Weng. Min-max multi-objective bilevel optimization with applications in robust machine learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=PvDY71zKsvP`.

Zhishuai Guo, Quanqi Hu, Lijun Zhang, and Tianbao Yang. Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021a.

Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the ADAM family and beyond. *arXiv preprint arXiv:2104.14840*, 2021b.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Quanqi Hu, Yongjian Zhong, and Tianbao Yang. Multi-block min-max bilevel optimization with applications in multi-task deep auc maximization. *Advances in Neural Information Processing Systems*, 35:29552–29565, 2022.

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

Minhui Huang, Xuxing Chen, Kaiyi Ji, Shiqian Ma, and Lifeng Lai. Efficiently escaping saddle points in bilevel optimization. *arXiv preprint arXiv:2202.03684v3*, 2023.

Kaiyi Ji, Jason D Lee, Yingbin Liang, and H Vincent Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. *Advances in Neural Information Processing Systems*, 33:11490–11500, 2020.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pages 4882–4892. PMLR, 2021.

Ruichen Jiang, Nazanin Abolfazli, Aryan Mokhtari, and Erfan Yazdandoost Hamedani. A conditional gradient-based method for simple bilevel optimization with convex lower-level problem. In *International Conference on Artificial Intelligence and Statistics*, pages 10305–10323. PMLR, 2023.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283, 2021.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages 18083–18113. PMLR, 2023.

Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

Junyi Li, Bin Gu, and Heng Huang. A fully single loop algorithm for bilevel optimization without Hessian inverse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7426–7434, 2022.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020a.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020b.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. BOME! Bilevel Optimization Made Easy: A Simple First-Order Approach. *Advances in Neural Information Processing Systems*, 35:17248–17262, 2022.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations*, 2019.

Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34:8662–8675, 2021.

Risheng Liu, Yaohua Liu, Wei Yao, Shangzhi Zeng, and Jin Zhang. Averaged method of multipliers for bi-level optimization without lower-level strong convexity. *arXiv preprint arXiv:2302.03407*, 2023.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015.

Thomas Moreau, Mathurin Massias, Alexandre Gramfort, Pierre Ablin, Pierre-Antoine Bannier, Benjamin Charlier, Mathieu Dagréou, Tom Dupre la Tour, Ghislain Durif, and Cassio F Dantas. Benchopt: Reproducible, efficient and collaborative optimization benchmarks. *Advances in Neural Information Processing Systems*, 35:25404–25421, 2022.

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

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.

Barak A Pearlmutter. Fast exact multiplication by the Hessian. *Neural computation*, 6(1): 147–160, 1994.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pages 737–746. PMLR, 2016.

Qi Qian, Shenghuo Zhu, Jiasheng Tang, Rong Jin, Baigui Sun, and Hao Li. Robust optimization over multiple domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4739–4746, 2019.

Shuang Qiu, Zhuoran Yang, Xiaohan Wei, Jieping Ye, and Zhaoran Wang. Single-timescale stochastic nonconvex-concave optimization for smooth nonlinear TD-learning. *arXiv preprint arXiv:2008.10103*, 2020.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR, 2016.

Cédric Rommel, Thomas Moreau, Joseph Paillard, and Alexandre Gramfort. Cadda: Class-wise automatic differentiable data augmentation for eeg signals. In *ICLR 2022-International Conference on Learning Representations*, 2022.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. *arXiv preprint arXiv:2302.05185*, 2023.

Daouda Sow, Kaiyi Ji, Ziwei Guan, and Yingbin Liang. A constrained optimization approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022a.

Daouda Sow, Kaiyi Ji, and Yingbin Liang. On the convergence theory for hessian-free bilevel algorithms. *Advances in Neural Information Processing Systems*, 35:4136–4149, 2022b.

Gilbert W Stewart. *Matrix algorithms: volume 1: basic decompositions*. SIAM, 1998.

Ioannis Tsaknakis, Prashant Khanduri, and Mingyi Hong. An implicit gradient-type method for linearly constrained bilevel problems. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5438–5442. IEEE, 2022.

Jingkang Wang, Tianyun Zhang, Sijia Liu, Pin-Yu Chen, Jiacen Xu, Makan Fardad, and Bo Li. Adversarial attack generation empowered by min-max optimization. *Advances in Neural Information Processing Systems*, 34:16020–16033, 2021.

Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161:419–449, 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Quan Xiao, Han Shen, Wotao Yin, and Tianyi Chen. Alternating projected sgd for equality-constrained bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 987–1023. PMLR, 2023.

Tesi Xiao, Krishnakumar Balasubramanian, and Saeed Ghadimi. A projection-free algorithm for constrained stochastic multi-level composition optimization. In *Advances in Neural Information Processing Systems*, volume 35, pages 19984–19996, 2022.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34:13670–13682, 2021.

Shuoguang Yang, Mengdi Wang, and Ethan X Fang. Multilevel stochastic gradient methods for nested composition optimization. *SIAM Journal on Optimization*, 29(1):616–659, 2019.

Yihua Zhang, Yuguang Yao, Parikshit Ram, Pu Zhao, Tianlong Chen, Mingyi Hong, Yanzhi Wang, and Sijia Liu. Advancing model pruning via bi-level optimization. In *Advances in Neural Information Processing Systems*, 2022.

Appendix A. Discussions on the prior works

A.1 Regarding Gu et al. (2023)

In this section, we discuss several issues in the current form of Gu et al. (2023), which introduces a **M**ulti-**O**bjective **R**obust **Bi**level **T**wo-timescale optimization algorithm (`MORBiT`).

The primary issue in the current analysis of `MORBiT` arises from the *ambiguity* and *inconsistency* regarding the *expectation and filtration*. As a consequence, the current form of the paper was unable to demonstrate $\mathbb{E}[\max_{i\in[n]} \|y_i^k - y_i^*(x^{(k-1)})\|^2] \leq \tilde{\mathcal{O}}(\sqrt{n}K^{-2/5})$ claimed in Theorem 1 (10b) of Gu et al. (2023). The subsequent arguments are incorrect. We discuss some mistakes made in Gu et al. (2023) as follows.

We start by looking at Lemma 8 (informal) and Lemma 14 (formal) in Gu et al. (2023) that characterize the upper bound of $\mathcal{L}^{(k+1)} - \mathcal{L}^{(k)}$ where $\mathcal{L}^{(k)} = \mathbb{E}[\sum_{i=1}^n \lambda_i^{(k)} \ell_i(x^{(k)})]$. Here, the function ℓ_i is the function $\Phi_i(x)$ in our notation. The paper incorrectly asserted that $\mathcal{L}^k = \sum_{i=1}^n \lambda_i^{(k)} \mathbb{E}[\ell_i(x^{(k)})]$. To see why, let $\mathscr{F}_k$ denote the sigma algebra generated by all iterates (x, y, λ) with superscripts not greater than k. It is important to note that both $\{\lambda_i^{(k)}\}$ and $x^{(k)}$ are random objectives given the filtration $\mathscr{F}_k$. The ambiguity lies in the lack of clarity regarding the randomness over which the expectation operation is performed. In fact,

we can rewrite the claim of Lemma 14 in Gu et al. (2023) without hiding the randomness. Let $\mathcal{L}^{(k)} = \sum_{i=1}^n \lambda_i^{(k)} \ell_i(x^{(k)})$. Then, we have

$$\mathcal{L}^{(k+1)} - \mathcal{L}^{(k)} \leq \mathcal{O}(\alpha) \underbrace{\left(\sum_{i=1}^n \lambda_i^k \|y_i^{k+1} - y_i^*(x^{(k)})\| \right)^2}_{\leq \max_{i\in[n]} \|y_i^{k+1} - y_i^*(x^{(k)})\|^2} - \frac{1}{\alpha}\|x^{k+1} - x^k\|^2 + \mathcal{O}(\gamma n) + \mathcal{O}(\alpha)\|h_x^{(k)} - \mathbb{E}[h_x^{(k)} \mid \mathscr{F}_k]\|^2, \tag{62}$$

where α, β, γ are step sizes for x, y, and λ respectively. We hide the dependency for constants in their assumptions for simplicity. In addition, we want to emphasize that, unlike our notation, $h_x^{(k)}$ and $h_\lambda^{(k)}$ are stochastic gradients at step k. Therefore, $h_x^{(k)}$ and $h_\lambda^{(k)}$ are random objects given $\mathscr{F}_k$. By taking expectations over all the randomness above, we can see that Lemma 14 in Gu et al. (2023) is incorrect because it writes in the form of $\max \mathbb{E}[\cdot]$ instead of $\mathbb{E}[\max(\cdot)]$. Therefore, the subsequent arguments regarding the convergence of x, y, λ are incorrect, at least in the current form.

Regardless of the error, one may be able to proceed with the proof by utilizing Eq.(62) since our ultimate goal is to demonstrate the convergence of $\mathbb{E}[\max_{i\in[n]} \|y_i^k - y_i^*(x^{(k-1)})\|^2]$. One possible direction is to utilize the basic recursive inequality of $\max_{i\in[n]} \|y_i^{k+1} - y_i^*(x^{(k)})\|^2$. Observe that for each $i \in [n]$, we can establish the following inequality similar to Lemma 13 in Gu et al. (2023) without hiding the randomness:

$$\|y_i^{(k+1)} - y_i^*(x^{(k)})\|^2 \leq (1 - \mathcal{O}(\mu_g\beta)) \|y_i^{(k)} - y_i^*(x^{(k-1)})\|^2 + \mathcal{O}\left(\frac{1}{\mu_g\beta}\right)\|x^k - x^{k-1}\|^2$$
$$+ \mathcal{O}(\beta^2)\|h_{y,i}^{(k)} - \mathbb{E}[h_{y,i}^{(k)}|\mathscr{F}_k]\|^2 + \mathcal{O}(\beta)\left\langle y_i^{(k)} - y_i^*(x^{(k-1)}), h_{y,i}^{(k)} - \mathbb{E}[h_{y,i}^{(k)}|\mathscr{F}_k]\right\rangle$$

However, the order of taking the expectation over the randomness and the maximum over $i \in [n]$ adds complexity to the problem. The last inner-product term can only be zero when first taking expectation given $\mathscr{F}_k$. When applying Young's inequality to bound this term, it inevitably introduces terms such as $\mathcal{O}(\beta)\|h_{y,i}^{(k)} - \mathbb{E}[h_{y,i}^{(k)}|\mathscr{F}_k]\|^2$ or $\mathcal{O}(1)\|y_i^{(k)} - y_i^*(x^{(k-1)})\|^2$, which make it challenging to proceed further with the convergence analysis.

Finally, we remark about the choice of the stationarity condition used in Gu et al. (2023). Although the algorithmic aspect in Gu et al. (2023) is motivated by Lin et al. (2020a), the notion of stationarity for λ in Gu et al. (2023) is different from Lin et al. (2020a). Under the notion of stationarity in Lin et al. (2020a) (Definition 3.7) $\Phi_{1/2\ell}(\cdot)$ is the Moreau envelope of $\Phi(\cdot)$, which is defined after taking the max over y (i.e., λ in our notation) in Definition 3.5 in Lin et al. (2020a), and a point x is ϵ-stationarity when $\|\nabla\Phi_{1/2\ell}(x)\| \leq \epsilon$. It is unclear if (10a) and (10c) in Gu et al. (2023) will imply similar convergence results under the notion of stationarity in Definition 3.7 in Lin et al. (2020a).

A.2 Regarding Hu et al. (2022)

Hu et al. (2022) considered a multi-block min-max bilevel optimization, which shares similarity with Problem (10) we consider. However, we note that their Assumption 2.2 on the LL

function g_i requires $\nabla^2_{22} g_i(x, y; \zeta) \succeq \mu_g I$, and is much stronger than ours and that in Gu et al. (2023). For example, for any $0 < \mu_g < L_g$ and

$$\nabla^2_{22} g_i(x, y_i; \zeta) = \begin{pmatrix} 2L_g & 0 \\ 0 & 0 \end{pmatrix} \text{ or } \begin{pmatrix} 0 & 0 \\ 0 & 2\mu_g \end{pmatrix} \text{ with equal probability}$$

indicates that $\nabla^2_{22} g_i(x, y_i) = \operatorname{diag}(L_g, \mu_g) \succeq \mu_g I$ can hold even if $\nabla^2_{22} g_i(x, y_i; \zeta) \succeq \mu_g I$ does not hold for any ζ. Further, they do not characterize the dependence on μ_g in the final complexity. Hence we omit a detailed comparison with Hu et al. (2022).