

A Provably Convergent Plug-and-Play Framework for Stochastic Bilevel Optimization*

Tianshu Chu^{†§}

CHUTS@EMAILS.BJUT.EDU.CN

Dachuan Xu[†]

XUDC@BJUT.EDU.CN

Wei Yao^{||§}

YAOW@SUSTECH.EDU.CN

Chengming Yu^{‡§}

YUCM@BUPT.EDU.CN

Jin Zhang^{¶§&}

ZHANGJ9@SUSTECH.EDU.CN

[†] INSTITUTE OF OPERATIONS RESEARCH AND INFORMATION ENGINEERING, BEIJING UNIVERSITY OF TECHNOLOGY, BEIJING, CHINA

^{||} SCHOOL OF ARTIFICIAL INTELLIGENCE, SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY, SHENZHEN, CHINA

[§] NATIONAL CENTER FOR APPLIED MATHEMATICS SHENZHEN, SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY, SHENZHEN, CHINA

[‡] SCHOOL OF SCIENCE, BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS, BEIJING, CHINA

[¶] DEPARTMENT OF MATHEMATICS, SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY, SHENZHEN, CHINA

[&] DETECTION INSTITUTE FOR ADVANCED TECHNOLOGY LONGHUA-SHENZHEN (DIATLHSZ), SHENZHEN, CHINA

Editor: Joseph Salmon

Abstract

Bilevel optimization has recently attracted significant attention in machine learning due to its wide range of applications and advanced hierarchical optimization capabilities. In this paper, we propose a plug-and-play framework, named PnPBO, for developing and analyzing stochastic bilevel optimization methods. This framework integrates both modern unbiased and biased stochastic estimators into the single-loop bilevel optimization framework introduced in Dagr  ou et al. (2022), with several improvements. In the implementation of PnPBO, all stochastic estimators for different variables can be independently incorporated, and an additional moving average technique is applied when using an unbiased estimator for the upper-level variable. In the theoretical analysis, we provide a unified convergence and complexity analysis for PnPBO, demonstrating that the adaptation of various stochastic estimators (including PAGE, ZeroSARAH, and mixed strategies) within the PnPBO framework achieves optimal sample complexity. Specifically, in the finite-sum setting, the resulting complexity matches the lower bound in Dagr  ou et al. (2024) and is comparable to that of single-level optimization (Zhou and Gu, 2019). This resolves the open question of whether the optimal complexity bounds for solving bilevel optimization are identical to those for single-level optimization. Finally, we empirically validate our framework, demonstrating its effectiveness on several benchmark problems and confirming our theoretical findings.

Keywords: bilevel optimization, stochastic optimization, plug-and-play, sample complexity

*. This work was partially presented at International Conference on Machine Learning (ICML) 2024 (Chu et al. (2024)). Corresponding authors: Jin Zhang.

1. Introduction

Bilevel optimization (BLO) effectively addresses challenges arising from hierarchical optimization, where the decision variables in the upper level are also involved in the lower level. In recent years, BLO has gained increasing attention due to its extensive and effective applications, including hyperparameter optimization (Franceschi et al., 2018), meta-learning (Ji et al., 2020), continual learning (Hao et al., 2023), and reinforcement learning (Shen et al., 2024).

In this paper, we focus on the nonconvex-strongly-convex BLO problem under classical assumptions, formulated as follows

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

where $H(x)$ denotes the total objective function, also referred to as the value function (Dagr  ou et al., 2022, 2024; Chen et al., 2024). The upper-level (UL) objective $f(x, y)$ is L^f -smooth and possibly nonconvex. The lower-level (LL) objective $g(x, y)$ is L_1^g -smooth and strongly convex in y , and its gradient ∇g is L_2^g -smooth. As is the case in many applications of interest in machine learning (Shalev-Shwartz and Ben-David, 2014), the UL and LL objective functions take the finite-sum form

$$f(x, y) = \frac{1}{n} \sum_{i=1}^n F_i(x, y), \quad g(x, y) = \frac{1}{m} \sum_{j=1}^m G_j(x, y). \quad (2)$$

A widely used and effective strategy for solving the BLO problem in (1) involves using implicit differentiation (Pedregosa, 2016; Liao et al., 2018; Lorraine et al., 2020; Li et al., 2022), which leads to the following expression for $\nabla H(x)$ (known as the hypergradient)

$$\nabla H(x) = \nabla_1 f(x, y^*(x)) - \nabla_{12}^2 g(x, y^*(x)) [\nabla_{22}^2 g(x, y^*(x))]^{-1} \nabla_2 f(x, y^*(x)).$$

The bilevel structure typically makes the objective $H(x)$ nonconvex, except in a few special cases. Therefore, our goal is to develop efficient gradient-based algorithms to find an ϵ -stationary point of $H(x)$, specifically, a point x that satisfies the inequality $\mathbb{E} \|\nabla H(x)\|^2 \leq \epsilon$ in the stochastic setting (Ghadimi and Lan, 2013). To achieve this, a promising approach is to adapt stochastic methods from single-level optimization, which we refer to as stochastic estimators, to the bilevel optimization context. However, estimating hypergradients requires solving LL problems and computing inverse Hessian-vector products.

To tackle these challenges, Ghadimi and Wang (2018) use multi-step gradient descent to approximately solve the LL problem and incorporate a truncated Neumann series to approximate the Hessian inversion. Subsequent studies (Yang et al., 2021; Khanduri et al., 2021) leverage variance reduction techniques, such as STORM (Cutkosky and Orabona, 2019) or SARAH (Pham et al., 2020), improving the sample complexity to $\tilde{\mathcal{O}}(\epsilon^{-1.5})$, where the $\tilde{\mathcal{O}}$ notation hides a factor of $\log(\epsilon^{-1})$. Notably, methods using Neumann series approximations introduce an additional $\tilde{\mathcal{O}}(1)$ factor in both sample complexity¹ and batch size.

Another approach to computing the inverse Hessian-vector product is to solve the corresponding linear system. For example, by employing a warm-start strategy and performing

1. The sample complexity refers to the number of calls made to stochastic gradients and Hessian (Jacobian)-vector products required to obtain an ϵ -stationary point.

one or more SGD steps to solve LL problems and the linear systems, stochastic algorithms such as AmIGO (Arbel and Mairal, 2022) and SOBA (Dagr  ou et al., 2022) achieve a sample complexity of $\mathcal{O}(\epsilon^{-2})$. However, AmIGO requires a batch size that scales with ϵ^{-1} , and SOBA depends on a stronger smoothness condition. Thus, there is a gap in the complexity analysis between stochastic bilevel and single-level optimization when implementing the SGD gradient estimator. Recently, this gap has been effectively addressed by MA-SOBA (Chen et al., 2024) through the incorporation of an additional moving average (MA) technique for the UL variable in SOBA.

For other stochastic gradient estimators, SABA (Dagr  ou et al., 2022) integrates the stochastic estimator SAGA (Defazio et al., 2014), and SRBA (Dagr  ou et al., 2024) employs SARAH to propose algorithms with lower sample complexity. As indicated by the results of SABA in Appendix D of Dagr  ou et al. (2022), the sample complexity of SABA, under classical smoothness assumptions, differs from the performance of its stochastic estimator SAGA in single-level optimization by a $\mathcal{O}((n + m)^{1/3})$ factor. Additionally, the theoretical analysis of SRBA requires stronger smoothness conditions to achieve a better sample complexity of $\mathcal{O}((n + m)^{1/2}\epsilon^{-1})$.

In this paper, we address several open and important questions in stochastic bilevel optimization. First, we observe that existing algorithms and their theoretical analyses typically rely on a specific stochastic estimator, with case-by-case theoretical analysis. Therefore, the following question arises:

Q1: Can we propose a unified algorithm design and theoretical analysis framework in which stochastic estimators can be integrated flexibly and independently?

Furthermore, observing that there remains a gap between stochastic bilevel and single-level optimization when using SAGA, the following question needs to be addressed:

Q2: Is it possible to bridge the gap between stochastic bilevel and single-level optimization when using SAGA?

In addition, noting that SRBA (Dagr  ou et al., 2024) utilizes a double-loop structure and requires stronger smoothness conditions, one fundamental question arises:

Q3: How can we develop a fully single-loop algorithm for solving stochastic bilevel optimization problems that achieves optimal sample complexity of $\mathcal{O}((n + m)^{1/2}\epsilon^{-1})$ under standard smoothness assumptions in the finite-sum setting?

1.1 Contribution

We summarize the major contributions of this work as follows.

- **A unified plug-and-play framework.** We propose a flexible and unified framework for stochastic BLO, referred to as PnPBO. At its core, PnPBO consists of three key features: (1) It integrates both unbiased and biased stochastic estimators into the single-loop BLO algorithm framework presented in Dagr  ou et al. (2022), allowing independent incorporation of stochastic estimators for different variables; (2) An additional MA technique is suggested when an unbiased estimator is used for the UL variable; (3) Clipping is applied to the implicit variable, following standard techniques in deep learning practices.
- **New algorithms with optimal sample complexity.** The PnPBO framework is versatile, accommodating all existing single-loop BLO algorithms. Within this

framework, we employ distinct designs of stochastic estimator combinations, resulting in some specific instances of PnPBO: (1) SPABA, SFFBA, and MSEBA, which use the stochastic estimators PAGE (Li et al., 2021a), ZeroSARAH (Li et al., 2021b), and a mixed strategy, achieving optimal sample complexity that matches the finite-sum lower bound in Dagr  ou et al. (2024); (2) MA-SABA, which bridges the gap between single-level and BLO when using SAGA; (3) Generalized SPABA and SRMBA: These are instances of the extended PnPBO framework applied to the expectation setting, where Generalized SPABA matches the lower bound in Arjevani et al. (2023) and SRMBA is near-optimal under existing bounds. See Tables 1 and 2 for details.

- **A unified and sharp convergence and complexity analysis.** We provide a unified convergence and complexity analysis for PnPBO, applicable to all variants of PnPBO with distinct combinations of stochastic estimators. Although the stochastic estimators for different variables can be independently plugged in, our theoretical analysis shows that the step size (learning rate) strategies for these variables must be coupled to ensure convergence. For further details regarding the explicit coupling step size conditions, refer to Theorems 14, 15 and Remark 13. Under this unified framework, we can derive convergence results for all proposed algorithms.
- **Empirical validation.** We validate our framework empirically, demonstrating its effectiveness on several benchmark problems and confirming the theoretical findings.

Algorithm	Stochastic Estimator	Sample Complexity	UL, LL	Gap	Reference
SABA	SAGA	$\mathcal{O}((n+m)\epsilon^{-1})$	$C_L^{1,1}, C_L^{2,2}$	✓	Dagr��ou et al. (2022)
		$\mathcal{O}((n+m)^{2/3}\epsilon^{-1})$	$C_L^{2,2}, C_L^{3,3}$	✓	
MA-SABA	SAGA+MA & SAGA	$\mathcal{O}((n+m)^{2/3}\epsilon^{-1})$	$C_L^{1,1}, C_L^{2,2}$	✗	Theorem 3.5*
SRBA	SARAH	$\mathcal{O}((n+m)^{1/2}\epsilon^{-1})$	$C_L^{2,2}, C_L^{3,3}$	✓	Dagr��ou et al. (2024)
SPABA	PAGE	$\mathcal{O}((n+m)^{1/2}\epsilon^{-1})$	$C_L^{1,1}, C_L^{2,2}$	✗	Theorem 3.7*
SFFBA	ZeroSARAH	$\mathcal{O}((n+m)^{1/2}\epsilon^{-1})$	$C_L^{1,1}, C_L^{2,2}$	✗	Theorem 16
MSEBA	ZeroSARAH & PAGE	$\mathcal{O}((n+m)^{1/2}\epsilon^{-1})$	$C_L^{1,1}, C_L^{2,2}$	✗	Theorem 18
Lower Bound: $\Omega((n+m)^{1/2}\epsilon^{-1})$ (Dagr��ou et al., 2024)					

Table 1: Comparison of stochastic bilevel optimization algorithms in the finite-sum setting.

Following Dagr  ou et al. (2024), the finite-sum complexity comparison is made in the regime $n+m \lesssim \epsilon^{-2}$. In this regime, the one-time $O(n+m)$ initialization cost is subsumed by $O((n+m)^{1/2}\epsilon^{-1})$. The symbol * indicates that this result has been published in our conference paper (Chu et al., 2024). $C_L^{p,p}$ means that for all $i \in [n]$, $F_i \in C^p$, and $\nabla^k F_i$ is Lipschitz continuous for all $0 < k \leq p$. The same holds for G_j for all $j \in [m]$. The symbol ✓ indicates a gap in either sample complexity or required conditions when using this stochastic estimator for single-level versus BLO. We omit algorithms such as SOBA(Dagr  ou et al., 2022), SUSTAIN (Khanduri et al., 2021), MRBO and VRBO (Yang et al., 2021), which have a sample complexity of at best $\tilde{O}(\epsilon^{-1.5})$. These algorithms, designed for objective functions in the expectation setting, are further discussed in Section 6.3.

1.2 Related Work

Stochastic Estimators and Lower Bounds. Single-level stochastic optimization has a variety of algorithms, which we refer to as stochastic estimators. A classic method is SGD (Sinha and Griscik, 1971; Bottou, 2010), which requires $\mathcal{O}(\epsilon^{-2})$ sample complexity to reach an ϵ -stationary point in non-convex optimization. Variants of SGD, such as reshuffled versions (Shamir, 2016) and other modifications (Lan, 2020), have been proposed to improve its performance. While SGD is widely used due to its simplicity, its variance remains constant. To address this issue, Johnson and Zhang (2013) introduced variance reduction techniques and proposed SVRG, which inspired more advanced algorithms like SAG (Roux et al., 2012), SAGA (Defazio et al., 2014), and STORM (Cutkosky and Orabona, 2019).

The finite-sum structure is widely applied in single-level optimization to develop first-order algorithms that converge faster. Several methods achieve optimal sample complexity, including SARAH (Pham et al., 2020), SPIDER (Fang et al., 2018), SNVRG (Zhou et al., 2020), PAGE (Li et al., 2021a), ZeroSARAH (Li et al., 2021b), SILVER (Oko et al., 2024), and SpiderBoost (Wang et al., 2019). The sample complexity lower bound $\Omega(\sqrt{n}\epsilon^{-1})$ is established in (Fang et al., 2018; Zhou and Gu, 2019).

Bilevel Optimization Methods. Other advances in stochastic bilevel optimization include: Li and Huang (2024) incorporated a without-replacement sampling strategy into SOBA. Recently, several algorithms using only first-order information have been proposed (Kwon et al., 2023; Yang et al., 2023; Chen et al., 2025). However, these algorithms either lack non-asymptotic convergence guarantees, have less competitive convergence rates compared to our results, or are limited to deterministic settings. Sow et al. (2022); Gu et al. (2021) proposed zeroth-order algorithms, but they do not have a convergence rate guarantee. Shen et al. (2025); Huang (2023); Kwon et al. (2024) proposed stochastic algorithms for bilevel optimization with a non-strongly convex LL objective.

2. Preliminaries

2.1 Notations

Let $N = n + m$. To simplify the notation, we introduce the following notation $D_{k;I,J}^x := \frac{1}{|I|} \sum_{i \in I} \nabla_1 F_i(x_k, y_k) - \frac{1}{|J|} \sum_{j \in J} \nabla_{12}^2 G_j(x_k, y_k) z_k$, where I and J are the sets of indices sampled for functions f and g at the k -th iteration. Similarly, $D_{k;J}^y$ and $D_{k;I,J}^z$ are defined. Additionally, let $y_k^* := y^*(x_k)$ and $z_k^* := z^*(x_k)$. The clipping function on z with radius $R > 0$ is defined as $\text{Clip}(z; R) = \min\{1, R/\|z\|\} \cdot z$.

2.2 Assumptions

We present the assumptions for the BLO problem under study.

- Assumption 1** (a) For any x , there exists $C^f > 0$ such that $\|\nabla_2 f(x, y^*(x))\| \leq C^f$;
 (b) For all $i \in [n]$, the function $\nabla F_i(x, y)$ is L^f -Lipschitz continuous in (x, y) ;
 (c) The function $H(x)$ is bounded from below, i.e., $H_* := \inf_{x \in \mathbb{R}^{d_x}} H(x) > -\infty$.

- Assumption 2** (a) For any x , $g(x, \cdot)$ is μ -strongly convex;
 (b) For all $j \in [m]$, the functions $\nabla G_j(x, y)$ and $\nabla^2 G_j(x, y)$ are L_1^g and L_2^g Lipschitz continuous in (x, y) , respectively.

Remark 1 *Assumptions 1(b) and 2(b) are stochastic conditions that many algorithms rely on to achieve enhanced complexity results (Fang et al., 2018; Li et al., 2021a; Dagr  ou et al., 2022). They imply that $\nabla f(x, y)$, $\nabla g(x, y)$, and $\nabla^2 g(x, y)$ are Lipschitz continuous with L^f , L_1^g , and L_2^g , respectively.*

2.3 Hypergradient Decoupling and Single-Loop Algorithm

We revisit the decoupling approach to address the challenge of computing the hypergradient. For a given x , we compute an approximation y of the LL minimizer $y^*(x) \in \arg \min_y g(x, y)$. Let the inverse Hessian-vector product $[\nabla_{22}^2 g(x, y^*(x))]^{-1} \nabla_2 f(x, y^*(x))$ be denoted by $z^*(x)$. To obtain a computable approximation of $z^*(x)$ in a single-loop scheme, we introduce an implicit variable z , which is computed as an inexact solution to the following quadratic minimization problem evaluated at the available pair (x, y) :

$$\min_{z \in \mathbb{R}^{d_y}} \frac{1}{2} \langle \nabla_{22}^2 g(x, y) z, z \rangle - \langle \nabla_2 f(x, y), z \rangle.$$

The process of solving the BLO problem can be viewed as consisting of three update modules: one for the UL variable x , one for the LL variable y , and one for the implicit variable z . Based on a warm-start strategy, the single-loop decoupling algorithms (Arbel and Mairal, 2022; Dagr  ou et al., 2022; Chen et al., 2024) are structured around these update directions and can be summarized as follows.

-
- 1: **for** $k = 0$ **to** $K - 1$ **do**
 - 2: Update $x_{k+1} = x_k - \alpha_k \tilde{D}_k^x$,
 where $\tilde{D}_k^x$ is an approximation of $D_k^x := \nabla_1 f(x_k, y_k) - \nabla_{12}^2 g(x_k, y_k) z_k$;
 - 3: Update $y_{k+1} = y_k - \beta_k \tilde{D}_k^y$,
 where $\tilde{D}_k^y$ is an approximation of $D_k^y := \nabla_2 g(x_k, y_k)$;
 - 4: Update $z_{k+1} = z_k - \gamma_k \tilde{D}_k^z$,
 where $\tilde{D}_k^z$ is an approximation of $D_k^z := \nabla_{22}^2 g(x_k, y_k) z_k - \nabla_2 f(x_k, y_k)$.
 - 5: **end for**
-

3. Method

In this section, we introduce a unified plug-and-play framework, PnPBO, enabling the independent integration of modern stochastic estimators and offering guidance on their usage. Furthermore, based on this framework, we propose two new algorithms as illustrative examples.

3.1 PnPBO: A Unified Plug-and-Play Framework

We now introduce PnPBO, a flexible and unified framework for stochastic bilevel optimization, as outlined in Algorithm 1. PnPBO consists of three modules: Lines 4-6 for updating the UL variable x , Lines 7-8 for updating the LL variable y , and Lines 9-10 for updating the implicit variable z . These modules operate in parallel and are independent. Unlike existing algorithms that rely on specific variance reduction techniques, this framework supports the

independent integration of different stochastic estimators $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. They can be either unbiased or biased.

For the implicit variable z , we apply clipping with a radius R in Line 10, where $R := C^f/\mu$ serves as the upper bound for $\|z^*(x)\|$ in Lemma 19. This technique has been introduced and used in bilevel optimization (Hu et al., 2023; Dagr  ou et al., 2024) to control the implicit iterates z_k , which allows the analysis to proceed without an a priori boundedness assumption on z . From a computational perspective, clipping is widely adopted in deep learning practice (Zhang et al., 2020; Bhattacharjee et al., 2025), and it does not introduce significant additional computation.

When $\mathcal{A}$ is an unbiased estimator, we recommend employing the MA technique. Importantly, unbiasedness of $\mathcal{A}$ (as an estimator of D_k^x) does not in general imply an unbiased hypergradient estimate, since D_k^x itself is an approximation of $\nabla H(x_k)$, so that in general $\mathbb{E}[\hat{v}_k^x] \neq \nabla H(x_k)$. We therefore introduce MA by incorporating the historical direction v_{k-1}^x into the current update to reduce the hypergradient estimation error and stabilize the x -update, which allows a larger step size α_k and leads to improved sample complexity.

Algorithm 1 PnPBO: A Unified Plug-and-Play Framework

- 1: **Input:** Initializations v_{-1}^x , ρ_{-1} , (x_{-1}, y_{-1}, z_{-1}) and (x_0, y_0, z_0) , number of total iterations K , stochastic estimators $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, step sizes $\{\alpha_k\}$, $\{\beta_k\}$, $\{\gamma_k\}$, MA weight $\{\rho_k\}$, clipping radius R .
 - 2: **for** $k = 0$ **to** $K - 1$ **do**
 - 3: Sample $\mathcal{S}_k^f$ for f and $\mathcal{S}_k^g$ for g ;
 - 4: Compute an estimate $\hat{v}_k^x$ of D_k^x using the stochastic estimator $\mathcal{A}$ with $\mathcal{S}_k^f$ and $\mathcal{S}_k^g$;
 - 5:
$$v_k^x = \begin{cases} \hat{v}_k^x, & \text{if } \mathcal{A} \text{ is a biased estimator of } D_k^x; \\ (1 - \rho_{k-1})v_{k-1}^x + \rho_{k-1}\hat{v}_k^x, & \text{if } \mathcal{A} \text{ is an unbiased estimator of } D_k^x; \end{cases}$$
 - 6: Update
$$x_{k+1} \leftarrow x_k - \alpha_k v_k^x;$$
 - 7: Compute an estimate v_k^y of D_k^y using the stochastic estimator $\mathcal{B}$ with $\mathcal{S}_k^g$;
 - 8: Update
$$y_{k+1} \leftarrow y_k - \beta_k v_k^y;$$
 - 9: Compute an estimate v_k^z of D_k^z using the stochastic estimator $\mathcal{C}$ with $\mathcal{S}_k^f$ and $\mathcal{S}_k^g$;
 - 10: Update
$$z_{k+1} \leftarrow \text{Clip}(z_k - \gamma_k v_k^z; R).$$
 - 11: **end for**
-

3.2 Instantiation of PnPBO

We present two algorithms as specific implementations of PnPBO. Below, we describe the stochastic estimators used and the resulting estimates of the three update directions after independent embedding.

3.2.1 SFFBA: STOCHASTIC FULL GRADIENT FREE BILEVEL ALGORITHM

We introduce a new algorithm, called SFFBA, which achieves optimal sample complexity in the finite-sum setting. SFFBA is an adaptation of the ZeroSARAH algorithm (Li et al., 2021b), originally designed for non-convex optimization, and tailored for the bilevel setting. The key advantage of this stochastic estimator is that it eliminates the need for multiple full gradient evaluations. We apply it to all three variable update modules.

Let $w_{k,i} = (w_{k,i}^x, w_{k,i}^y)$ for $i \in [n]$ and $\tilde{w}_{k,j} = (\tilde{w}_{k,j}^x, \tilde{w}_{k,j}^y, \tilde{w}_{k,j}^z)$ for $j \in [m]$ represent two memory variables, corresponding to the calls to functions f and g . For each iteration $k \geq 1$, we sample sets $I \subset [n]$ and $J \subset [m]$ for f and g with a minibatch size of b . The update rules for $w_{k,i}$ are as follows: if $i \in I$, then $w_{k,i}$ is updated to (x_k, y_k) ; if $i \notin I$, $w_{k,i}$ retains its previous value $w_{k-1,i}$. Similarly, for $\tilde{w}_{k,j}$, the update rule follows the same structure.

The update direction for the variable x is $v_k^x = \hat{v}_k^x$ with

$$\hat{v}_k^x = (1 - \bar{\rho}_k) (v_{k-1}^x - D_{k-1;I,J}^x) + D_{k;I,J}^x + \bar{\rho}_k (\hat{D}_{k-1;[n],[m]}^x - \hat{D}_{k-1;I,J}^x),$$

where $\bar{\rho}_k$ is the momentum parameter, and we define the notation as

$$\hat{D}_{k;I,J}^x := \frac{1}{|I|} \sum_{i \in I} \nabla_1 F_i(w_{k,i}^x, w_{k,i}^y) - \frac{1}{|J|} \sum_{j \in J} \nabla_{12}^2 G_j(\tilde{w}_{k,j}^x, \tilde{w}_{k,j}^y) \tilde{w}_{k,j}^z.$$

Similarly for $\hat{D}_{k;J}^y$ and $\hat{D}_{k;I,J}^z$. The update directions for y and z follow a similar approach

$$\begin{aligned} v_k^y &= (1 - \bar{\rho}_k) (v_{k-1}^y - D_{k-1;J}^y) + D_{k;J}^y + \bar{\rho}_k (\hat{D}_{k-1;[m]}^y - \hat{D}_{k-1;J}^y), \\ v_k^z &= (1 - \bar{\rho}_k) (v_{k-1}^z - D_{k-1;I,J}^z) + D_{k;I,J}^z + \bar{\rho}_k (\hat{D}_{k-1;[n],[m]}^z - \hat{D}_{k-1;I,J}^z). \end{aligned} \quad (3)$$

Following ZeroSARAH (Li et al., 2021b, Algorithm 2 and Corollaries 1–2), two initialization settings can be considered. The first uses a full batch at $k = 0$. Specifically, for SFFBA, we use the one-time full initialization, for all $i \in [n]$ and $j \in [m]$,

$$w_{0,i} = (x_0, y_0), \quad \tilde{w}_{0,j} = (x_0, y_0, z_0), \quad v_0^x = \hat{v}_0^x = D_0^x, \quad v_0^y = D_0^y, \quad v_0^z = D_0^z. \quad (4)$$

No full-gradient evaluation is required thereafter. The second uses a minibatch at $k = 0$ and avoids full initialization, as in ZeroSARAH (Li et al., 2021b, Corollary 2). In this setting, the initial update directions are computed from the sampled minibatch rather than the full data set.

3.2.2 MSEBA: MULTIPLE STOCHASTIC ESTIMATORS BILEVEL ALGORITHM

Next, we introduce MSEBA, an algorithm that integrates different stochastic estimators, including PAGE and ZeroSARAH. This algorithm demonstrates the flexibility of PnPBO and provides an example for its plug-and-play implementation.

We choose the stochastic estimators $\mathcal{A}$ and $\mathcal{C}$ for the UL variable x and implicit variable z update modules, respectively, as PAGE. Specifically, for $k \geq 1$, the updates are as follows

$$v_k^x = \hat{v}_k^x = \begin{cases} D_{k;[n],[m]}^x, & \text{w.p. } p, \\ v_{k-1}^x + D_{k;I,J}^x - D_{k-1;I,J}^x, & \text{w.p. } 1 - p. \end{cases}$$

That is, with probability (w.p.) p , we compute the full gradient D_k^x , and w.p. $1 - p$, we update the previous iteration's direction using the samples. The update direction for

z is constructed analogously: with probability p , $v_k^z = D_k^z$, and with probability $1 - p$, $v_k^z = v_{k-1}^z + D_{k;I,J}^z - D_{k-1;I,J}^z$. The stochastic estimator $\mathcal{B}$ is chosen as ZeroSARAH, and the construction of its update direction is given by (3). One initialization setting for MSEBA is the following one-time full initialization, where $\tilde{w}_{0,j}$ is initialized for all $j \in [m]$:

$$v_0^x = \hat{v}_0^x = D_0^x, \quad v_0^y = D_0^y, \quad v_0^z = D_0^z, \quad \tilde{w}_{0,j} = (x_0, y_0, z_0). \quad (5)$$

For the ZeroSARAH estimator $\mathcal{B}$, the minibatch initialization described in Section 3.2.1 can alternatively be used at $k = 0$, thereby avoiding full initialization of $\mathcal{B}$.

Remark 2 *As established in Section 6.2, MSEBA achieves optimal sample complexity in the finite-sum setting. This algorithm serves as an example to demonstrate the theoretical feasibility of integrating different stochastic estimators into the three update modules, highlighting the flexibility of PnPBO. It encourages users to select estimators based on problem-specific characteristics or practical outcomes.*

4. Theoretical Analysis Framework

In this section, we present a unified and sharp convergence and complexity analysis framework. Such a framework would help clarify how the variance properties of stochastic estimators, combined with step size conditions, contribute to the final convergence rates and sample complexities. The algorithm framework PnPBO and its convergence analysis framework are illustrated in Figure 1.

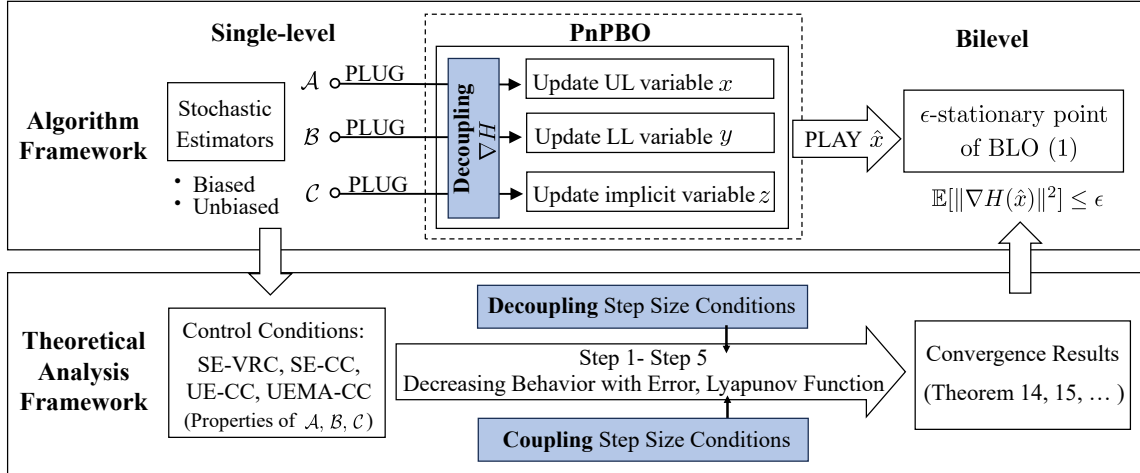


Figure 1: PnPBO and Unified Analysis Framework Schematic

4.1 On the Variance Properties of Stochastic Estimators

In this subsection, we recall a generic template for single-level stochastic estimators and their variance properties. These statements will be repeatedly invoked when we analyze the same estimators after embedding them into the bilevel framework. Specifically, we consider the single-level non-convex problem $\min_u h(u) := \frac{1}{Q} \sum_{q=1}^Q h_q(u)$.

The stochastic estimator v_k is typically constructed using the current iterate u_k , the sampled set P_k and historical information such as $\{(v_i, u_i)\}_{i=0}^{k-1}$. Some estimators additionally maintain explicit memory variables $\tilde{u}_{k-1} = (\tilde{u}_{k-1}^1, \dots, \tilde{u}_{k-1}^Q)$, which are maintained and updated along the iterations based on the iterate and sampling histories. Accordingly, the algorithm is driven by the recursion

$$v_k = V(\{\nabla h_q(u_k)\}_{q \in P_k}, \{v_i\}_{i=0}^{k-1}, \{u_i\}_{i=0}^{k-1}, \tilde{u}_{k-1}, \dots), \quad u_{k+1} = u_k - a_k v_k,$$

where a_k is the step size and other auxiliary variables are updated by prescribed rules.

We work with the following abstract mean-squared estimation error recursion for the estimator $\{v_k\}$:

$$\begin{aligned} \mathbb{E}[\|v_{k+1} - \nabla h(u_{k+1})\|^2] &\leq \theta_k \mathbb{E}[\|v_k - \nabla h(u_k)\|^2] + \eta'_k \mathbb{E}[\|\nabla h_{q_{k+1}}(u_{k+1}) - \nabla h_{q_{k+1}}(u_k)\|^2] \\ &\quad + \tau'_k \frac{1}{Q} \sum_{q=1}^Q \mathbb{E}[\|\nabla h_q(u_k) - \nabla h_q(\tilde{u}_k^q)\|^2] + \Delta_k, \end{aligned} \quad (6)$$

where $\theta_k \in [0, 1)$ and $\eta'_k \geq 0$ may depend on the batch size $p_{k+1} := |P_{k+1}|$. Let q_{k+1} be sampled uniformly from P_{k+1} (given P_{k+1}). The second term on the right-hand side of (6) captures the local change of a sampled component gradient along the update, and we keep it in a single-sample form to facilitate the subsequent bilevel analysis. Assume further that each h_q has an L_h -Lipschitz gradient. Then

$$\eta'_k \mathbb{E}[\|\nabla h_{q_{k+1}}(u_{k+1}) - \nabla h_{q_{k+1}}(u_k)\|^2] \leq L_h^2 \eta'_k a_k^2 \mathbb{E}[\|v_k\|^2] =: \eta_k a_k^2 \mathbb{E}[\|v_k\|^2].$$

If v_k depends on memory variables, define the mean-squared staleness $\delta_k := \frac{1}{Q} \sum_{q=1}^Q \mathbb{E}[\|u_k - \tilde{u}_k^q\|^2]$. By L_h -smoothness, the third term on the right-hand side of (6) (the memory term) is bounded by $\tau_k \delta_k$ with $\tau_k := \tau'_k L_h^2$. This mean-squared memory staleness δ_k typically satisfies a contraction-type recursion up to additional terms:

$$\delta_{k+1} \leq \lambda_k \delta_k + \hat{\eta}_k \mathbb{E}[\|u_{k+1} - u_k\|^2] + \hat{\eta}'_k a_k^2 \mathbb{E}[\|\nabla h(u_k)\|^2], \quad (7)$$

where $\lambda_k \in [0, 1)$ and the coefficients in the above inequality may depend on the size of the data set Q . The term Δ_k collects the method-dependent residual variance not captured by the other terms (e.g., the noise in plain SGD).

Moreover, for an unbiased estimator with $\mathbb{E}[v_k] = \nabla h(u_k)$, there exist nonnegative sequences $\{\text{Coef}_k\}_{k \geq 0}$ and $\{\text{Coef}'_k\}_{k \geq 0}$ such that

$$\text{Coef}_{k+1} \theta_k - \text{Coef}_k \leq -2a_k^2, \quad \text{Coef}_{k+1} \tau_k + \text{Coef}'_{k+1} \lambda_k - \text{Coef}'_k \leq 0. \quad (8)$$

For a biased estimator, there exist nonnegative sequences $\{\text{Coef}_k^b\}_{k \geq 0}$ and $\{\text{Coef}'_k^b\}_{k \geq 0}$ such that

$$\text{Coef}_{k+1}^b \theta_k - \text{Coef}_k^b \leq -2a_k, \quad \text{Coef}_{k+1}^b \tau_k + \text{Coef}'_{k+1}^b \lambda_k - \text{Coef}'_k^b \leq 0. \quad (9)$$

The inequalities (8)-(9) express compatibility between step sizes and Lyapunov coefficients, which will be summarized as control conditions next. Appendix B provides instantiations of (6)–(9) for several standard estimators.

4.2 Control Conditions

We summarize the single-level template in (6)–(9) into a set of abstract control conditions, which will be used as the interface between concrete stochastic estimators and our bilevel framework PnPBO. We begin by introducing the SE-VRC (Stochastic Estimator Variance Recursion Control) condition that characterizes the variance behavior of stochastic estimators. This condition applies to both biased and unbiased estimators.

Condition 1 (SE-VRC condition) *A stochastic estimator is said to satisfy the SE-VRC condition if it exhibits properties of the form (6) and (7) in the corresponding single-level stochastic optimization problem.*

Under the SE-VRC condition, we introduce coefficient control conditions, including the SE-CC condition (Stochastic Estimator Coefficient Control) for biased estimators and the UE-CC condition (Unbiased Estimator Coefficient Control) for unbiased estimators. These conditions relate the step size a_k , Lyapunov coefficients, and variance-related coefficients.

Condition 2 (SE-CC condition) *A stochastic estimator $\mathcal{A}$ is said to satisfy the SE-CC condition with $\{A, A'\}$ if there exist sequences of nonnegative coefficients $A = \{A_k\}$ and $A' = \{A'_k\}$ such that (9) hold, i.e. for $k \geq 0$:*

$$A_{k+1}\theta_k^A - A_k \leq -2a_k, \quad A_{k+1}\tau_k^A + A'_{k+1}\lambda_k^A - A'_k \leq 0.$$

Condition 3 (UE-CC condition) *A stochastic estimator $\mathcal{A}$ is said to satisfy the UE-CC condition with $\{A, A'\}$ if there exist sequences of nonnegative coefficients $A = \{A_k\}$ and $A' = \{A'_k\}$ such that (8) hold, i.e. for $k \geq 0$:*

$$A_{k+1}\theta_k^A - A_k \leq -2a_k^2, \quad A_{k+1}\tau_k^A + A'_{k+1}\lambda_k^A - A'_k \leq 0.$$

Furthermore, our framework involves unbiased estimators combined with MA, with momentum parameter $\rho_k \in (0, 1]$. In the Lyapunov analysis, the MA update introduces an additional deviation term, which is controlled by an extra Lyapunov coefficient A'' . This motivates the following UEMA-CC (Unbiased Estimator MA Coefficient Control) condition.

Condition 4 (UEMA-CC condition) *We say that a stochastic estimator $\mathcal{A}$ satisfies the UEMA-CC condition with $\{A, A', A''\}$ if there exist sequences of nonnegative coefficients $A = \{A_k\}$, $A' = \{A'_k\}$ and $A'' = \{A''_k\}$ such that for $k \geq 0$, the following inequalities hold:*

$$\begin{aligned} \rho_k^2 A''_{k+1} \theta_k^A + A_{k+1} \theta_k^A - A_k &\leq 0, \quad \frac{a_k}{2} + A''_{k+1} (1 - \rho_k) - A''_k \leq 0, \\ \rho_k^2 A''_{k+1} \tau_k^A + \tau_k^A A_{k+1} + \lambda_k^A A'_{k+1} - A'_k &\leq 0, \end{aligned}$$

where $\rho_k \in (0, 1]$ denotes the moving average weight.

Remark 3 (Discussion of the Control conditions)

(1) The SE-VRC condition characterizes the variance property that stochastic estimators should possess in single-level stochastic optimization problems. Based on this, we can establish the validity of the inequalities in the BLO setting, specifically (10)–(11) and (19)–(22).

(2) The SE-CC, UE-CC, and UEMA-CC conditions provide coefficient relations between step sizes, Lyapunov coefficients, and variance-related coefficients, ensuring that the SE-VRC recursions can be combined into a telescoping Lyapunov decrease in the bilevel analysis.

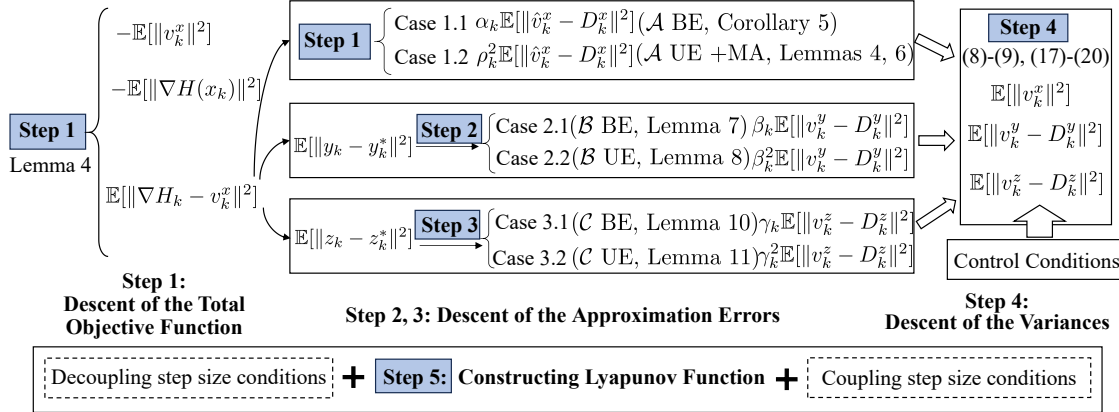


Figure 2: Roadmap of the Analysis Framework. Above, “BE” represents biased stochastic estimator and “UE” represents unbiased stochastic estimator. In the figure, we have kept only the key terms from the formulas.

4.3 Analysis Steps of Framework

This subsection provides the five-step roadmap used throughout the paper. Steps 1–4 establish monotone descent behaviors with error terms, depending on whether the stochastic estimators are biased or unbiased. These behaviors can be expressed as recursive inequalities in conditional expectation

$$\mathbb{E}[\tilde{D}_{k+1} | \mathcal{F}_k] + \Lambda_k \leq \omega_k \tilde{D}_k + \Omega_k,$$

where $\tilde{D}_k$, Λ_k , and Ω_k are nonnegative quantities, and $\omega_k \in [0, 1]$ is a contraction factor. Step 5 constructs a Lyapunov function, which integrates the lemmas from Steps 1-4 to derive the final convergence results.

Specifically, as illustrated in Figure 2: (1) Starting from the total objective function $H(x)$ in Step 1, and based on Lemma 4, the errors in the monotonic decreasing behavior of $H(x_k)$ include the gap between the UL variable update direction and the true hypergradient, i.e., $\mathbb{E}[\|\nabla H(x_k) - v_k^x\|^2]$. This gap is analyzed case by case in Corollary 5 and Lemma 6, depending on whether $\mathcal{A}$ is biased or unbiased. (2) The approximation errors $\mathbb{E}[\|y_k - y_k^*\|^2]$ and $\mathbb{E}[\|z_k - z_k^*\|^2]$ appear in the errors of the descent behavior derived in Step 1 (as shown in Corollary 5 and Lemma 6). We further analyze these errors separately in Step 2 and Step 3. (3) In Step 4, we control the variance term by examining the performance of the stochastic estimators when embedded in our bilevel setting. (4) Finally, in Step 5, we construct an appropriate Lyapunov function to integrate the above analyses and derive the convergence and complexity results. Next, we will provide the details for each step.

Step 1: Descent of the Total Objective Function $H(x)$

Under Assumptions 1 and 2, H is L^H -smooth (Lemma 20). Based on this, we can characterize the descent behavior of the overall objective function with errors, as described in the following lemma.

Lemma 4 Suppose Assumptions 1 and 2 hold, and $\alpha_k \leq 1/(2L^H)$. Then we have

$$\mathbb{E}[H(x_{k+1})] \leq \mathbb{E}[H(x_k)] - \frac{\alpha_k}{2} \mathbb{E}[\|\nabla H(x_k)\|^2] - \frac{\alpha_k}{4} \mathbb{E}[\|v_k^x\|^2] + \frac{\alpha_k}{2} \mathbb{E}[\|\nabla H(x_k) - v_k^x\|^2].$$

Notably, Lemma 4 does not depend on the type of stochastic estimator $\mathcal{A}$. To further quantify how the term $\mathbb{E}[\|\nabla H(x_k) - v_k^x\|^2]$ affects the descent of H , we establish two refined descent results, distinguishing whether $\mathcal{A}$ is biased or unbiased.

Case 1.1: $\mathcal{A}$ is a biased stochastic estimator.

We decompose the gap between the hypergradient and the iterative direction of x , $\mathbb{E}[\|\nabla H(x_k) - v_k^x\|^2]$, into two terms: $\mathbb{E}[\|\nabla H(x_k) - D_k^x\|^2]$ and $\mathbb{E}[\|v_k^x - D_k^x\|^2]$, leading to the following corollary.

Corollary 5 Suppose Assumptions 1 and 2 hold, and $\alpha_k \leq 1/(2L^H)$. Then we have

$$\begin{aligned} \mathbb{E}[H(x_{k+1})] \leq & \mathbb{E}[H(x_k)] - \frac{\alpha_k}{2} \mathbb{E}[\|\nabla H(x_k)\|^2] - \frac{\alpha_k}{4} \mathbb{E}[\|v_k^x\|^2] + \alpha_k \mathbb{E}[\|v_k^x - D_k^x\|^2] \\ & + 3c_1\alpha_k \mathbb{E}[\|y_k - y_k^*\|^2] + 3c_2\alpha_k \mathbb{E}[\|z_k - z_k^*\|^2], \end{aligned}$$

where $c_1 = (L^f)^2 + (L_2^g R)^2$, $c_2 = (L_1^g)^2$.

Case 1.2: $\mathcal{A}$ is an unbiased stochastic estimator.

When $\mathcal{A}$ is an unbiased stochastic estimator, combining it with the MA technique yields the following lemma.

Lemma 6 Supposing that Assumptions 1 and 2 hold and that $\mathcal{A}$ is an unbiased stochastic estimator. Then we have

$$\begin{aligned} & \mathbb{E}[\|v_{k+1}^x - \nabla H(x_{k+1})\|^2] - \mathbb{E}[\|v_k^x - \nabla H(x_k)\|^2] \\ \leq & -\rho_k \mathbb{E}[\|v_k^x - \nabla H(x_k)\|^2] + \rho_k^2 \mathbb{E}[\|\hat{v}_{k+1}^x - D_{k+1}^x\|^2] + \frac{3\alpha_k^2 (L^H)^2 + 12c_1\rho_k^2\alpha_k^2}{\rho_k} \mathbb{E}[\|v_k^x\|^2] \\ & + 12\rho_k(c_1\beta_k^2 \mathbb{E}[\|v_k^y\|^2] + c_2\gamma_k^2 \mathbb{E}[\|v_k^z\|^2]) + 9\rho_k(c_1 \mathbb{E}[\|y_k - y_k^*\|^2] + c_2 \mathbb{E}[\|z_k - z_k^*\|^2]). \end{aligned}$$

Step 2: Descent of the Approximation Error $\mathbb{E}[\|y_k - y_k^*\|^2]$

In this step, we evaluate the error from approximating y_k^* with y_k . We construct the corresponding lemmas separately, depending on whether $\mathcal{B}$ is a biased or unbiased estimator.

Case 2.1: $\mathcal{B}$ is a biased stochastic estimator.

Lemma 7 Suppose Assumptions 1 and 2 hold, $\mathcal{B}$ is a biased stochastic estimator and the step size satisfies $\beta_k \leq \bar{c}$. Then we have

$$\mathbb{E}[\|y_{k+1} - y_{k+1}^*\|^2] \leq (1 - c_3\beta_k) \mathbb{E}[\|y_k - y_k^*\|^2] + \frac{c_5\alpha_k^2}{\beta_k} \mathbb{E}[\|v_k^x\|^2] + c_4\beta_k \mathbb{E}[\|v_k^y - D_k^y\|^2],$$

where $\bar{c} = \min\left\{\frac{\mu+L_1^g}{\mu L_1^g}, \frac{1}{\mu+L_1^g}\right\}$, $c_3 = \frac{\mu L_1^g}{2(\mu+L_1^g)}$, $c_4 = 6\frac{\mu+L_1^g}{\mu L_1^g}$, $c_5 = \frac{2(\mu+L_1^g)L_{y^*}^2}{\mu L_1^g}$.

Case 2.2: $\mathcal{B}$ is an unbiased stochastic estimator.

Lemma 8 Suppose Assumptions 1 and 2 hold, $\mathcal{B}$ is an unbiased stochastic estimator and the step size satisfies $\beta_k \leq 1/(\mu + L_1^g)$. Then we have

$$\mathbb{E} \left[\|y_{k+1} - y_{k+1}^*\|^2 \right] \leq (1 - \beta_k \mu) \mathbb{E} \left[\|y_k - y_k^*\|^2 \right] + \frac{c_6 \alpha_k^2}{\beta_k} \mathbb{E} \left[\|v_k^x\|^2 \right] + 2\beta_k^2 \mathbb{E} \left[\|v_k^y - D_k^y\|^2 \right],$$

where $c_6 = 2L_{y^*}^2/\mu$.

Remark 9 The key difference between Lemmas 7 and 8 is the coefficient on $\mathbb{E}[\|v_k^y - D_k^y\|^2]$, which scales as β_k for biased estimators and as β_k^2 for unbiased ones; an analogous pattern holds for Lemmas 10 and 11.

Step 3: Descent of the Approximation Error $\mathbb{E}[\|z_k - z_k^*\|^2]$

Similar to Step 2, we derive the following two lemmas based on the type of $\mathcal{C}$ to measure the approximation error between z_k and z_k^* .

Case 3.1: $\mathcal{C}$ is a biased stochastic estimator.

Lemma 10 Suppose Assumptions 1 and 2 hold, $\mathcal{C}$ is a biased stochastic estimator and the step size satisfies $\gamma_k \leq \bar{c}$. Then we have

$$\begin{aligned} \mathbb{E} \left[\|z_{k+1} - z_{k+1}^*\|^2 \right] &\leq \left(1 - \frac{\mu}{4} \gamma_k \right) \mathbb{E} \left[\|z_k - z_k^*\|^2 \right] + \frac{18c_1 \gamma_k}{\mu} \mathbb{E} \left[\|y_k - y_k^*\|^2 \right] \\ &\quad + c_{10} \gamma_k \mathbb{E} \left[\|v_k^z - D_k^z\|^2 \right] + \frac{c_7 \alpha_k^2}{\gamma_k} \mathbb{E} \left[\|v_k^x\|^2 \right], \end{aligned}$$

where $c_7 = 4L_{z^*}^2/\mu$, $c_{10} = 12/\mu$.

Case 3.2: $\mathcal{C}$ is an unbiased stochastic estimator.

Lemma 11 Suppose Assumptions 1 and 2 hold, $\mathcal{C}$ is an unbiased stochastic estimator and the step size satisfies $\gamma_k \leq \min\{1/(10\mu), 1/L_1^g\}$. Then we have

$$\begin{aligned} \mathbb{E} \left[\|z_{k+1} - z_{k+1}^*\|^2 \right] &\leq (1 - \gamma_k \mu) \mathbb{E} \left[\|z_k - z_k^*\|^2 \right] + c_8 \gamma_k \mathbb{E} \left[\|y_k - y_k^*\|^2 \right] \\ &\quad + 2\gamma_k^2 \mathbb{E} \left[\|v_k^z - D_k^z\|^2 \right] + \frac{c_9 \alpha_k^2}{\gamma_k} \mathbb{E} \left[\|v_k^x\|^2 \right], \end{aligned}$$

where $c_8 = 8c_1/\mu$, $c_9 = 3L_{z^*}^2/\mu$.

Step 4: Descent of the Variances

In this section, we characterize the variance behavior of single-level stochastic estimators when they are embedded in our bilevel framework. Assuming that $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ in Algorithm 1 satisfy the SE-VRC condition in Section 4.2, we establish the corresponding bilevel variance bounds as follows. These variances include $\mathbb{E}[\|\hat{v}_k^x - D_k^x\|^2]$, $\mathbb{E}[\|v_k^y - D_k^y\|^2]$, and $\mathbb{E}[\|v_k^z - D_k^z\|^2]$.

Case 4.1: For the stochastic estimator $\mathcal{A}$ in the bilevel setting, we have

$$\begin{aligned} \mathbb{E}[\|\hat{v}_{k+1}^x - D_{k+1}^x\|^2] &\leq \theta_k^A \mathbb{E}[\|\hat{v}_k^x - D_k^x\|^2] + \eta_k^A c_1 \alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + \eta_k^A c_1 \beta_k^2 \mathbb{E}[\|v_k^y\|^2] \\ &\quad + \eta_k^A c_2 \gamma_k^2 \mathbb{E}[\|v_k^z\|^2] + \tau_k^A \delta_k^A + \Delta_k^A. \end{aligned} \tag{10}$$

Here δ_k^A denotes the mean-squared staleness induced by the memory tables maintained by $\mathcal{A}$, and it satisfies

$$\begin{aligned} \delta_{k+1}^A \leq & \lambda_k^A \delta_k^A + \hat{\eta}_k^{A,x} \alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + \hat{\eta}_k^{A,y} \beta_k^2 \mathbb{E}[\|v_k^y\|^2] + \hat{\eta}_k^{A,z} \gamma_k^2 \mathbb{E}[\|v_k^z\|^2] \\ & + \hat{\eta}_k'^{A,x} \alpha_k^2 \mathbb{E}[\|D_k^x\|^2] + \hat{\eta}_k'^{A,y} \beta_k^2 \mathbb{E}[\|D_k^y\|^2] + \hat{\eta}_k'^{A,z} \gamma_k^2 \mathbb{E}[\|D_k^z\|^2]. \end{aligned} \quad (11)$$

Detailed proofs are deferred to Appendix D, where we also provide the corresponding interface inequalities (19)–(20) for $\mathcal{B}$ and (21)–(22) for $\mathcal{C}$.

Remark 12 *In our PnPBO framework, which couples the UL and LL objectives f and g through the variables (x, y, z) , embedding a single-level estimator typically leads to recursions that are slight adaptations of the single-level ones. Here, we treat (10)–(11) and (19)–(22) as abstract interface inequalities for plug-in estimators: to embed a specific estimator, one may start from its single-level derivation and establish the corresponding bilevel counterparts term by term under Algorithm 1, thereby identifying the involved parameters; see, for example, Lemma 23.*

Step 5: Constructing Lyapunov Function

We use a Lyapunov function to combine the above inequalities. Specifically, we construct the Lyapunov function as follows

$$\begin{aligned} L_k = & \mathbb{E}[H(x_k)] + A_k \mathbb{E}[\|\hat{v}_k^x - D_k^x\|^2] + B_k \mathbb{E}[\|v_k^y - D_k^y\|^2] + C_k \mathbb{E}[\|v_k^z - D_k^z\|^2] + A'_k \delta_k^A \\ & + B'_k \delta_k^B + C'_k \delta_k^C + A''_k \mathbb{E}[\|v_k^x - \nabla H(x_k)\|^2] + C_y \mathbb{E}[\|y_k - y_k^*\|^2] + C_z \mathbb{E}[\|z_k - z_k^*\|^2]. \end{aligned}$$

Here, all the Lyapunov coefficients are nonnegative. By Assumption 1(c), $L_k \geq \mathbb{E}[H(x_k)] \geq H_*$. If $\mathcal{A}$ is a biased estimator, set $A''_k = 0$.

The construction of the above Lyapunov function is analysis-driven: we systematically decouple all error sources arising in the algorithmic recursion until no further decoupling is possible. The underlying logic is as follows. Since our stationarity criterion is based on the hypergradient $H(x_k)$, we must explicitly control both the *approximation errors* and the *stochastic errors* in hypergradient estimation. The approximation errors include $\mathbb{E}[\|y_k - y_k^*\|^2]$ and $\mathbb{E}[\|z_k - z_k^*\|^2]$, together with the additional tracking term $\mathbb{E}[\|v_k^x - \nabla H(x_k)\|^2]$, which captures the hypergradient-tracking improvement brought by the MA mechanism in the x -module. The stochastic errors consist of the variance terms $\mathbb{E}[\|\hat{v}_k^x - D_k^x\|^2]$, $\mathbb{E}[\|v_k^y - D_k^y\|^2]$, $\mathbb{E}[\|v_k^z - D_k^z\|^2]$, as well as the memory (staleness) terms δ_k^A , δ_k^B , and δ_k^C .

Combine the inequalities presented in Steps 1-4 to provide an upper bound for $L_{k+1} - L_k$. The inequalities to be selected for estimating each term can be found in Figure 2.

We choose the Lyapunov coefficients and step sizes so that the decoupling and coupling step size conditions required by our framework are satisfied (see Sections 5.1 and 5.2). Under Assumptions 1–2 and the corresponding control conditions, this yields

$$\alpha_k \mathbb{E}[\|\nabla H(x_k)\|^2] \leq L_k - L_{k+1} + \tilde{\Delta}_k,$$

where $\tilde{\Delta}_k = (A_{k+1} + A''_{k+1} \rho_k^2) \Delta_k^A + B_{k+1} \Delta_k^B + C_{k+1} \Delta_k^C$ and $\Delta_k^A, \Delta_k^B, \Delta_k^C$ are given in (10), (19) and (21), respectively.

Therefore, we can obtain the final convergence result

$$\inf_{k < K} \mathbb{E}[\|\nabla H(x_k)\|^2] \leq \frac{L_0 - H_*}{K\theta} + \frac{\sum_{k=0}^{K-1} \tilde{\Delta}_k}{K\theta},$$

where θ is a parameter that depends on α_k . This bound is an abstract template: once the specific estimators are fixed, it suffices to verify the corresponding control conditions and step size relations to obtain an explicit rate. In the next section, we instantiate our analysis framework and derive the corresponding general convergence theorems.

5. Theoretical Results

In this section, we establish general results for the cases where all estimators are biased and where all estimators are unbiased, as presented in Theorems 14 and 15, respectively.

Notably, our convergence analysis framework allows for the integration of both biased and unbiased estimators. However, due to space constraints, we omit explicit theorems for this combined case.

5.1 General Convergence Theorem 14: biased estimators

In this subsection, we present a general convergence result for the case where all stochastic estimators are biased.

Before stating the main result, we summarize the step size conditions required in Theorem 14. Specifically, we impose the following decoupling step size conditions

$$\alpha_k \leq \frac{1}{2LH}, \beta_k \leq \bar{c}, \gamma_k \leq \bar{c}, \left(B_{k+1}\eta_k^{\mathcal{B}} + B'_{k+1}\hat{\eta}_k^{\mathcal{B},y}\right)\beta_k \leq M_1, B'_{k+1}\hat{\eta}_k^{\mathcal{B},y}\beta_k \leq \frac{c_3^2}{36c_2}, \quad (12)$$

and the coupling step size conditions

$$\alpha_k \leq m_{\alpha\beta}\beta_k, \quad \alpha_k \leq m_{\alpha\gamma}\gamma_k, \quad \gamma_k \leq m_{\gamma\beta}\beta_k, \quad (13a)$$

$$\left(A_{k+1}\eta_k^{\mathcal{A}} + A'_{k+1}\hat{\eta}_k^{\mathcal{A},\ell}\right)\beta_k \leq M_2, \quad A'_{k+1}\hat{\eta}_k^{\mathcal{A},\ell}\beta_k \leq \tilde{M}_2, \quad (13b)$$

$$\left(C_{k+1}\eta_k^{\mathcal{C}} + C'_{k+1}\hat{\eta}_k^{\mathcal{C},\ell}\right)\beta_k \leq M_2, \quad C'_{k+1}\hat{\eta}_k^{\mathcal{C},\ell}\beta_k \leq \tilde{M}_2, \quad (13c)$$

$$\left(B_{k+1}\eta_k^{\mathcal{B}} + B'_{k+1}\hat{\eta}_k^{\mathcal{B},x}\right)\beta_k \leq \frac{1}{96m_{\alpha\beta}\tilde{c}_2}, \quad B'_{k+1}\hat{\eta}_k^{\mathcal{B},x}\beta_k \leq \frac{1}{196m_{\alpha\beta}}, \quad (13d)$$

where $\ell \in \{x, y, z\}$, and the constants above are determined by the constants in Assumptions 1–2. Specifically, $\tilde{c}_1 = c_1 + 1$, $\tilde{c}_2 = c_2 + 1$, $M_1 = \min\left\{\frac{1}{6\tilde{c}_2}, \frac{1}{4\tilde{c}_1}, \frac{c_3^2}{72\tilde{c}_2^2}\right\}$, $M_2 = \min\left\{\frac{1}{96m_{\alpha\beta}\tilde{c}_1}, \frac{c_3^2}{72\tilde{c}_1c_2}, \frac{1}{6\tilde{c}_1}, \frac{1}{4m_{\gamma\beta}\tilde{c}_2}, \frac{c_3^2}{72m_{\gamma\beta}\tilde{c}_2L_z^2}, \frac{\mu}{40m_{\gamma\beta}\tilde{c}_2c_{10}L_z^2}\right\}$, $m_{\alpha\beta} = \min\left\{\frac{3}{8L_{y*}^2}, \frac{c_3^2}{108c_1}\right\}$, $m_{\alpha\gamma} = \min\left\{\frac{3}{16L_{z*}^2}, \frac{\mu}{60c_2c_{10}}\right\}$, $m_{\gamma\beta} = \frac{c_3^2}{54c_1}$, $\tilde{M}_2 = \min\left\{\frac{c_3^2}{36c_2}, \frac{c_3^2}{36m_{\gamma\beta}L_z^2}, \frac{\mu}{20m_{\gamma\beta}c_{10}L_z^2}, \frac{1}{196m_{\alpha\beta}}\right\}$.

Remark 13 *Decoupling strategies bring immediate benefits in BLO, as they render each search direction linear with respect to the functions f and g , which simplifies both the algorithmic structure and the analysis. Nevertheless, the decoupling is not complete: the step sizes remain intrinsically coupled, as reflected in (13). Moreover, for ease of verification, the above step size conditions are simplified based on (13a).*

With these step size conditions in place, we are ready to state the following theorem.

Theorem 14 *Fix an iteration $K > 1$ and suppose that Assumptions 1 and 2 hold. Select stochastic estimators $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ as **biased**, and assume they satisfy the SE-CC and SE-VRC conditions with $\{A, A'\}$, $\{B, B'\}$, and $\{C, C'\}$, respectively. Assume that the sequences $\{\alpha_k\}$, $\{\beta_k\}$ and $\{\gamma_k\}$ satisfy the decoupling step size conditions in (12) and the coupling step size conditions in (13). Then we have*

$$\inf_{k < K} \mathbb{E}[\|\nabla H(x_k)\|^2] \leq \frac{2(L_0 - H_*)}{\sum_{k=0}^{K-1} \alpha_k} + \frac{2}{\sum_{k=0}^{K-1} \alpha_k} \sum_{k=0}^{K-1} (A_{k+1} \Delta_k^{\mathcal{A}} + B_{k+1} \Delta_k^{\mathcal{B}} + C_{k+1} \Delta_k^{\mathcal{C}}).$$

5.2 General Convergence Theorem 15: unbiased estimators

Subsequently, we provide a general result for the scenario in which all stochastic estimators are unbiased. Before stating the main result, we summarize the step size conditions required in Theorem 15, starting with the following decoupling conditions

$$\alpha_k \leq \frac{1}{2L^H}, \quad \beta_k \leq \min \left\{ \frac{1}{\mu + L_1^g}, \frac{\mu}{16c_2}, \frac{\mu}{16L_z^2} \right\}, \quad \gamma_k \leq \min \left\{ \frac{\mu}{10L_z^2}, \frac{1}{10\mu}, \frac{1}{L_1^g} \right\}, \quad (14a)$$

$$\rho_k \leq \min \left\{ \frac{L^H}{2\sqrt{c_1}}, 1 \right\}, \quad \left(\eta_k^{\mathcal{B}} B_{k+1} + \hat{\eta}_k^{\mathcal{B},y} B'_{k+1} \right) \leq \frac{1}{10\tilde{c}_2}, \quad \hat{\eta}_k^{\mathcal{B},y} B'_{k+1} \beta_k \leq \frac{\mu}{16c_2}, \quad (14b)$$

and the coupling step size conditions are given by

$$\alpha_k \leq m'_{\alpha\beta} \beta_k, \quad \alpha_k \leq m'_{\alpha\gamma} \gamma_k, \quad \gamma_k \leq m'_{\gamma\beta} \beta_k, \quad \rho_k \eta_k^{\mathcal{A}} \leq 4, \quad \frac{\alpha_k}{\rho_k} A''_{k+1} \leq \frac{1}{96(L^H)^2}, \quad (15a)$$

$$\rho_k A''_{k+1} \leq \min \left\{ \frac{\mu\beta_k}{144c_1}, \frac{\mu\gamma_k}{90c_2}, \frac{1}{144c_1}, \frac{1}{120c_2} \right\}, \quad \left(\eta_k^{\mathcal{B}} B_{k+1} + \hat{\eta}_k^{\mathcal{B},x} B'_{k+1} \right) \leq \frac{L^H}{16\tilde{c}_2}, \quad (15b)$$

$$\left(\eta_k^{\mathcal{A}} A_{k+1} + \hat{\eta}_k^{\mathcal{A},\ell} A'_{k+1} \right) \leq M_3, \quad \left(\eta_k^{\mathcal{C}} C_{k+1} + \hat{\eta}_k^{\mathcal{C},\ell} C'_{k+1} \right) \leq M_3, \quad (15c)$$

$$\hat{\eta}_k^{\mathcal{A},\ell} A'_{k+1} \beta_k \leq M_4, \quad \hat{\eta}_k^{\mathcal{B},x} B'_{k+1} \beta_k \leq \frac{1}{24m'_{\alpha\beta}}, \quad \hat{\eta}_k^{\mathcal{C},\ell} C'_{k+1} \beta_k \leq M_4, \quad (15d)$$

where $\ell \in \{x, y, z\}$, $M_3 = \min \left\{ \frac{1}{10\tilde{c}_2}, \frac{L^H}{16\tilde{c}_1} \right\}$, $M_4 = \min \left\{ \frac{1}{24m'_{\alpha\beta}}, \frac{\mu}{16c_2}, \frac{\mu}{10L_z^2 m'_{\gamma\beta}} \right\}$, $m'_{\alpha\beta} = \min \left\{ \frac{1}{32c_6}, \frac{\mu}{24c_1} \right\}$, $m'_{\alpha\gamma} = \min \left\{ \frac{1}{32c_9}, \frac{\mu}{15c_2} \right\}$, $m'_{\gamma\beta} = \min \left\{ \frac{\mu}{8c_8}, \frac{5}{16} \right\}$.

With these step size conditions in place, we are ready to state the following theorem.

Theorem 15 *Fix an iteration $K > 1$ and suppose that Assumptions 1 and 2 hold. Consider unbiased stochastic estimators $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. Assume that all three estimators satisfy the SE-VRC condition. In addition, $\mathcal{A}$ satisfies the UEMA-CC condition with coefficient family $\{A, A', A''\}$, and $\mathcal{B}$ and $\mathcal{C}$ satisfy the UE-CC condition with coefficient families $\{B, B'\}$ and $\{C, C'\}$, respectively. Assume that the sequences $\{\alpha_k\}$, $\{\beta_k\}$, $\{\gamma_k\}$, and $\{\rho_k\}$ satisfy the decoupling conditions in (14) and the coupling conditions in (15). Then we have*

$$\inf_{k < K} \mathbb{E}[\|\nabla H(x_k)\|^2] \leq \frac{4 \left(L_0 - H_* + \sum_{k=0}^{K-1} ((A_{k+1} + A''_{k+1} \rho_k^2) \Delta_k^{\mathcal{A}} + B_{k+1} \Delta_k^{\mathcal{B}} + C_{k+1} \Delta_k^{\mathcal{C}}) \right)}{\sum_{k=0}^{K-1} \alpha_k}.$$

6. Illustrative Examples

To illustrate the effectiveness of the proposed algorithm and its convergence analysis framework, we provide several examples in this section.

We begin by presenting the proofs of convergence and sample complexity for SFFBA and MSEBA. We then discuss how this framework can generate additional algorithms and extend to the expectation setting. It is important to note that MA-SABA and SPABA, which were introduced in the conference version, also serve as specific instances of our framework. Due to space limitations, we will not delve further into these methods here.

6.1 Convergence Rate and Sample Complexity of SFFBA

Using the convergence analysis framework, we can derive the following convergence result for SFFBA.

Theorem 16 *Fix an iteration $K > 1$ and assume that Assumptions 1 and 2 hold. Select the batch size as $b = \sqrt{N}$, and set the momentum parameter $\bar{\rho}_k \equiv \bar{\rho} = \frac{b}{2N} = \frac{1}{2\sqrt{N}}$. Then there exist positive constants $\alpha, \beta, \gamma, c_\beta, c_{\alpha\beta}$, and $c_{\gamma\beta}$ such that if*

$$\alpha_k \equiv \alpha \leq 1/(2L^H), \quad \gamma_k \equiv \gamma \leq \bar{c}, \quad \beta_k \equiv \beta \leq c_\beta, \quad \alpha_k \beta_k \leq c_{\alpha\beta}, \quad \gamma_k \beta_k \leq c_{\gamma\beta}, \quad (16)$$

and the step sizes satisfy (13a), the iterates in SFFBA using the one-time full initialization (4) satisfy

$$\inf_{k < K} \mathbb{E}[\|\nabla H(x_k)\|^2] = \mathcal{O}(K^{-1}). \quad (17)$$

To attain an ϵ -stationary point, the sample complexity is $\mathcal{O}(N + \sqrt{N}\epsilon^{-1})$.

Proof Based on the convergence rate analysis framework we provided, and in conjunction with Theorem 14, we will now give the values of the undetermined parameters, select the appropriate step sizes and Lyapunov function coefficients, and verify the conditions that need to be satisfied in Theorem 14.

- ① Select the Lyapunov function coefficients as $A_{k+1} = \frac{2\alpha}{\bar{\rho}}, B_{k+1} = \frac{2\beta}{\bar{\rho}}, C_{k+1} = \frac{2\gamma}{\bar{\rho}}$, $A'_{k+1} = 8\alpha\bar{\rho}L'', B'_{k+1} = 8\beta\bar{\rho}L'', C'_{k+1} = 8\gamma\bar{\rho}L''$. Let $c_{\gamma\beta} = c_{\alpha\beta} := \sqrt{\frac{M_2}{16c_1+12L''}}$, and $c_\beta := \min\left\{\bar{c}, \sqrt{\frac{M_1}{16+12L''}}, \sqrt{\frac{1}{(16+12L'')96m_{\alpha\beta}\bar{c}_2}}\right\}$.

- ② Validation of the SE-VRC condition:

Lemma 2 in Li et al. (2021b) verifies that the SE-VRC condition holds. More details on the properties of ZeroSARAH in the BLO setting can be found in Lemma 23.

- ③ Validation of the SE-CC Condition: It holds by

$$\begin{aligned} A_{k+1}\theta_k^A - A_k &= \frac{2\alpha}{\bar{\rho}}(1 - \bar{\rho})^2 - \frac{2\alpha}{\bar{\rho}} \leq -2\alpha, \\ A_{k+1}\tau_k^A + A'_{k+1}\lambda_k^A - A'_k &= \frac{2\alpha}{\bar{\rho}} \frac{2L''\bar{\rho}^2}{b} + 8\alpha\bar{\rho}L'' \left(1 - \frac{b}{2N}\right) - 8\alpha\bar{\rho}L'' \leq 0. \end{aligned}$$

For $\mathcal{B}$ and $\mathcal{C}$, the same reasoning applies.

- ④ Validation of the decoupling step size condition (12): It holds by (16) and

$$\left(B_{k+1} \eta_k^{\mathcal{B}} + B'_{k+1} \hat{\eta}_k^{\mathcal{B},y} \right) \beta_k = \left(\frac{2\beta}{\bar{\rho}} \frac{4}{b} + 8\beta \bar{\rho} L'' \frac{3N}{b} \right) \beta \leq (16 + 12L'') \beta^2 \leq M_1.$$

- ⑤ Validation of the coupling step size condition: We have

$$\left(A_{k+1} \eta_k^{\mathcal{A}} + A'_{k+1} \hat{\eta}_k^{\mathcal{A},\ell} \right) \beta_k = \left(\frac{8}{b\bar{\rho}} + \frac{12\bar{\rho}N}{b} L'' \right) \alpha \beta \leq M_2,$$

where $\ell \in \{x, y, z\}$. Similarly, we have the coupling conditions (13c)-(13d) are satisfied.

Thus, all conditions in Theorem 14 are satisfied, and we have $\inf_{k < K} \mathbb{E}[\|\nabla H(x_k)\|^2] \leq \frac{2(L_0 - H_*)}{\sum_{k=0}^{K-1} \alpha_k}$. Under the one-time full initialization, the initial variance and staleness terms vanish. Hence, $L_0 - H_* = \mathbb{E}[H(x_0)] + C_y \mathbb{E}[\|y_0 - y_0^*\|^2] + C_z \mathbb{E}[\|z_0 - z_0^*\|^2] - H_*$, and therefore does not depend on the N -dependent Lyapunov coefficients. Since the one-time full initialization requires $\mathcal{O}(N)$ stochastic-oracle calls and each subsequent iteration uses $\mathcal{O}(b)$ samples, the total sample complexity for attaining an ϵ -stationary point is $\mathcal{O}(N + b\epsilon^{-1}) = \mathcal{O}(N + \sqrt{N}\epsilon^{-1})$ for $b = \sqrt{N}$. $\blacksquare$

Remark 17 *This implies that there is no gap between stochastic bilevel and single-level optimization in the context of ZeroSARAH implementation. The lower bound established in (Dagr  ou et al., 2024) for BLO indicates that SFFBA achieves optimal sample complexity in the finite-sum setting when $m = \mathcal{O}(n)$ and $\epsilon = \mathcal{O}(n^{-1/2})$. However, memory remains a practical limitation of SFFBA: consistent with the ZeroSARAH estimator embedded in the framework, SFFBA maintains per-sample memory variables for variance reduction, resulting in a memory requirement of order $\mathcal{O}(n(d_x + d_y) + m(d_x + 2d_y))$.*

6.2 Convergence Rate and Sample Complexity of MSEBA

In this section, we state the convergence rate and sample complexity of MSEBA.

Theorem 18 *Fix an iteration $K > 1$ and assume that Assumptions 1 and 2 hold. Choose the batch size as $b = \sqrt{N}$, the probability $p = \frac{b}{N+b}$, and set the parameter $\bar{\rho}_k \equiv \bar{\rho} = \frac{1}{2\sqrt{N}}$. If the step sizes satisfy (13a) and $\alpha_k \equiv \alpha \leq 1/(2L^H)$, $\gamma_k \equiv \gamma \leq \bar{c}$, $\beta_k \equiv \beta \leq c_\beta$, $\alpha_k \beta_k \leq M_2/2$, $\gamma_k \beta_k \leq M_2/2$, then the iterates in MSEBA using the one-time full initialization (5) satisfy*

$$\inf_{k < K} \mathbb{E}[\|\nabla H(x_k)\|^2] = \mathcal{O}(K^{-1}),$$

where $c_\beta > 0$ is chosen as in the proof of Theorem 16.

Moreover, to achieve an ϵ -stationary point, the sample complexity is $\mathcal{O}(N + \sqrt{N}\epsilon^{-1})$.

Proof We apply Theorem 14 to prove this result.

- ① Select Lyapunov function coefficients as:

$$A_{k+1} = \frac{2\alpha}{p}, B_{k+1} = \frac{2\beta}{\bar{\rho}}, C_{k+1} = \frac{2\gamma}{p}, A'_{k+1} = C'_{k+1} = 0 \text{ and } B'_{k+1} = 8\beta \bar{\rho} L''.$$

- ② Validation of the SE-VRC condition and SE-CC Condition:

According to Lemma 3 in Li et al. (2021a) and Lemma 23, we obtain that the stochastic estimators $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ satisfy SE-VRC condition with the following values

$$\begin{aligned} \theta_k^{\mathcal{A}} &= \theta_k^{\mathcal{C}} = (1-p), \quad \theta_k^{\mathcal{B}} = (1-\bar{\rho})^2, \quad \eta_k^{\mathcal{A}} = \eta_k^{\mathcal{C}} = \frac{1-p}{b}, \quad \eta_k^{\mathcal{B}} = \frac{4}{b}, \quad \hat{\eta}_k^{\mathcal{B},\ell} = \frac{3N}{b}, \\ \tau_k^{\mathcal{A}} &= \tau_k^{\mathcal{C}} = \delta_k^{\mathcal{A}} = \delta_k^{\mathcal{C}} = \lambda_k^{\mathcal{A}} = \lambda_k^{\mathcal{C}} = \hat{\eta}_k^{\mathcal{A},\ell} = \hat{\eta}_k^{\mathcal{C},\ell} = 0, \quad \tau_k^{\mathcal{B}} = \frac{2L''\bar{\rho}^2}{b}, \quad \delta_k^{\mathcal{B}} = \delta_k, \quad \lambda_k^{\mathcal{B}} = 1 - \frac{b}{2N}, \\ \hat{\eta}_k^{\prime\mathcal{A},\ell} &= \hat{\eta}_k^{\prime\mathcal{B},\ell} = \hat{\eta}_k^{\prime\mathcal{C},\ell} = \Delta_k^{\mathcal{A}} = \Delta_k^{\mathcal{B}} = \Delta_k^{\mathcal{C}} = 0. \end{aligned}$$

The stochastic estimator $\mathcal{B}$ satisfies the SE-CC condition, as verified in ③ in the proof of Theorem 16. For $\mathcal{A}$, we have

$$A_{k+1}\theta_k^{\mathcal{A}} - A_k = (1-p)\frac{2\alpha}{p} - \frac{2\alpha}{p} + 0 = -2\alpha, \quad A_{k+1}\tau_k^{\mathcal{A}} + A'_{k+1}\lambda_k^{\mathcal{A}} - A'_k = 0.$$

The same applies to $\mathcal{C}$.

③ Validation of the decoupling step size condition: see ④ in the proof of Theorem 16.

④ Validation of the coupling step size condition: for conditions (13b)-(13c), we have

$$\left(A_{k+1}\eta_k^{\mathcal{A}} + A'_{k+1}\hat{\eta}_k^{\mathcal{A},\ell}\right)\beta_k = 2\alpha\beta\frac{1-p}{pb} = 2\alpha\beta \leq M_2, \quad \left(C_{k+1}\eta_k^{\mathcal{C}} + C'_{k+1}\hat{\eta}_k^{\mathcal{C},\ell}\right)\beta_k \leq M_2.$$

Thus, all the conditions in Theorem 14 are satisfied, and we finally obtain the convergence result, which implies that, to achieve an ϵ -stationary point, the sample complexity is $\mathcal{O}(N + b\epsilon^{-1} + (pN + (1-p)b)\epsilon^{-1}) = \mathcal{O}(N + \sqrt{N}\epsilon^{-1})$. $\blacksquare$

6.3 Extension to the Expectation Setting

It is worth noting that our algorithm framework, PnPBO, and its convergence analysis framework can be directly applied to the expectation setting, where the objective functions are

$$f(x, y) = \mathbb{E}_{\xi} [F(x, y; \xi)], \quad g(x, y) = \mathbb{E}_{\zeta} [G(x, y; \zeta)].$$

This extends problem (2) and captures scenarios where the data comes from a stream or online setting with a substantial or potentially infinite number of samples.

In this expectation setting, we propose two illustrative algorithms as examples. First, we introduce SRMBA, which directly integrates the stochastic estimator STORM into PnPBO. Secondly, by adjusting the batch size, we propose the generalized SPABA, which is the application of PAGE within PnPBO. These algorithms are described in Section 2.4 and 2.3 of the conference version Chu et al. (2024).

The analysis framework remains the same as described in Section 4, with the difference being that its variance reduction in Step 4 may depend on the addition of the following variance bounding assumption:

Assumption 3 *There exist positive constants σ_f , $\sigma_{g,1}$, and $\sigma_{g,2}$ such that*

$$\begin{aligned} \mathbb{E}[\|\nabla F(x, y; \xi) - \nabla f(x, y)\|^2] &\leq \sigma_f^2, \\ \mathbb{E}[\|\nabla G(x, y; \zeta) - \nabla g(x, y)\|^2] &\leq \sigma_{g,1}^2, \quad \mathbb{E}[\|\nabla^2 G(x, y; \zeta) - \nabla^2 g(x, y)\|^2] \leq \sigma_{g,2}^2. \end{aligned}$$

Under Assumptions 1-3, the sample complexity of SRMBA is $\tilde{\mathcal{O}}(\epsilon^{-1.5})$, with an improved batch size of $\mathcal{O}(1)$. For the generalized SPABA, it achieves optimal sample complexity $\mathcal{O}(\epsilon^{-1.5})$. For more details, please refer to Theorem 3.11 and Theorem 3.9 in the conference version. We compare SRMBA and the generalized SPABA with other methods in the expectation setting in Table 2.

Algorithm	Stochastic Estimator	Sample Complexity	Gap	Batch Size	Reference
stocBiO	SGD	$\tilde{\mathcal{O}}(\epsilon^{-2})$	✓	$\tilde{\mathcal{O}}(\epsilon^{-1})$	Ji et al. (2021)
MRBO	STORM	$\tilde{\mathcal{O}}(\epsilon^{-1.5})$	✓	$\tilde{\mathcal{O}}(1)$	Yang et al. (2021)
SUSTAIN	STORM	$\tilde{\mathcal{O}}(\epsilon^{-1.5})$	✓	$\tilde{\mathcal{O}}(1)$	Khanduri et al. (2021)
VRBO	SARAH	$\tilde{\mathcal{O}}(\epsilon^{-1.5})$	✓	$\tilde{\mathcal{O}}(\epsilon^{-0.5})$	Yang et al. (2021)
SRMBA	STORM	$\tilde{\mathcal{O}}(\epsilon^{-1.5})$	✗	$\mathcal{O}(1)$	Theorem 3.11*
SPABA	PAGE	$\mathcal{O}(\epsilon^{-1.5})$	✗	$\mathcal{O}(1)$	Theorem 3.9*
Lower Bound: $\Omega(\epsilon^{-1.5})$ (Arjevani et al., 2023)					

Table 2: Comparison of Stochastic Algorithms for BLO in the Expectation Setting.

* indicates that this result has been published in our conference paper (Chu et al., 2024). We omit comparisons with BSA (Ghadimi and Wang, 2018), TTSA (Hong et al., 2023), SOBA (Dagr  ou et al., 2022), AmIGO (Arbel and Mairal, 2022), and MA-SOBA (Chen et al., 2024), which rely on smoothness assumptions for f and g (instead of the mean-squared smoothness condition), with a best-case sample complexity of $\mathcal{O}(\epsilon^{-2})$.

7. Numerical Experiments

In this section, we conduct numerical experiments grounded in our theoretical results to validate the effectiveness of the proposed framework and associated algorithms.

We first compare the performance of our algorithms—SPABA, SFFBA, and MSEBA—with several benchmark methods on two widely studied tasks: data hyper-cleaning on the corrupted MNIST² data set, and hyperparameter optimization for ℓ^2 -penalized logistic regression on the IJCNN³ and covtype⁴ benchmark data sets. In addition, to validate the effectiveness of the PnPBO framework, we conduct ablation studies on both the moving average technique and the variable clipping technique. All experiments were repeated five times, and the average results were reported.

Setting. We strictly follow the experimental settings provided in benchmark_bilevel in Dagr  ou et al. (2022). The original results and configurations are also publicly available at https://benchopt.github.io/results/benchmark_bilevel.html.

In all experiments, we use a batch size of 64 across all methods. The step sizes and momentum parameters for the benchmark algorithms are adopted directly from the fine-tuned values provided by Dagr  ou et al. (2022). For our methods—SPABA, SFFBA, and MSEBA—we select the best constant step sizes based on a grid search.

2. <http://yann.lecun.com/exdb/mnist/>

3. <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html>

4. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_covtype.html

In this experiment, the optimal pair (α, β, γ) is selected via grid search. The parameter α is chosen from a set of 11 values logarithmically spaced between 10^{-3} and 10^0 . For β and γ , we first select ϕ and κ from 11 values logarithmically spaced between 10^{-5} and 10^0 , and then set $\beta = \frac{\alpha}{\phi}$ and $\gamma = \frac{\alpha}{\kappa}$. In SPABA, the probability parameter $1 - p$ is selected from the set $\{0.001, 0.002, \dots, 1\}$. For SFFBA and MSEBA, the momentum parameter $\bar{\rho}$ is chosen from the set $\{0.01, 0.02, \dots, 1\}$. The clipping radius is set to $R = 1$.

The benchmark methods we compared include SGD-based algorithms: stocBio (Ji et al., 2021), SOBA (Dagr  ou et al., 2022), AmIGO (Arbel and Mairal, 2022), TTSA (Hong et al., 2023); variance-reduction algorithms: SUSTAIN (Khanduri et al., 2021), MRBO (Yang et al., 2021); algorithms proposed for finite-sum setting: SABA (Dagr  ou et al., 2022) and SRBA (Dagr  ou et al., 2024).

7.1 Data Hyper-Cleaning

The first learning task we address is data hyper-cleaning on the MNIST data set. The data set is partitioned into a training set $(d_i^{\text{train}}, y_i^{\text{train}})$ with 20,000 samples, a validation set $(d_j^{\text{val}}, y_j^{\text{val}})$ with 5,000 samples, and a test set with 10,000 samples. The input samples d are 784-dimensional, with target labels y ranging from 0 to 9. In the training set, each sample is independently corrupted with probability $\tilde{p}$ by randomly replacing its label y_i with one from $\{0, \dots, 9\}$. The validation and test sets remain uncorrupted.

The objective of data hyper-cleaning is to train a multinomial logistic regression model on the corrupted training set, assigning a weight to each training sample to ideally reduce the influence of corrupted samples by assigning them weights close to zero. This is formulated as a BLO problem, with the UL objective function $f(\lambda, \theta) = \frac{1}{m} \sum_{j=1}^m \ell(\theta d_j^{\text{val}}, y_j^{\text{val}})$, and the LL objective function $g(\lambda, \theta) = \frac{1}{n} \sum_{i=1}^n \sigma(\lambda_i) \ell(\theta d_i^{\text{train}}, y_i^{\text{train}}) + C_r \|\theta\|^2$, where ℓ denotes the cross-entropy loss, σ is the sigmoid function, and we choose $C_r = 0.2$ after a manual search to achieve the best final test accuracy. In this experiment, the LL variable θ is a matrix of size 10×784 , and the UL variable λ is a vector in dimension $n_{\text{train}} = 20,000$.

Figure 3(a), we report the test error—defined as the percentage of incorrect predictions on the test set—comparing SPABA, SFFBA, and MSEBA with variance-reduction-based methods, including those designed for the finite-sum setting, under a corruption probability of $\tilde{p} = 0.9$. To supplement this comparison, we conducted additional experiments using corruption probabilities $\tilde{p} \in \{0.5, 0.7, 0.9\}$ with a broader set of comparison methods. The corresponding results are presented in Figures 4(a)–4(c).

We observe that our method, SPABA, consistently outperforms other methods across all three values of $\tilde{p}$, particularly on the more challenging settings with $\tilde{p} = 0.7$ and $\tilde{p} = 0.9$, achieving lowest test errors at a faster rate. Our second algorithm, MSEBA, which employs a mixed strategy by combining the stochastic estimators from both SPABA and SFFBA, performs between SPABA and SFFBA. It surpasses most other baseline algorithms and demonstrates strong performance on the data hyper-cleaning task.

7.2 Hyperparameter Selection on the covtype Data Set

In the second task, we address the problem of hyperparameter selection for determining regularization parameters in ℓ^2 -penalized logistic regression, evaluated on the `covtype` data

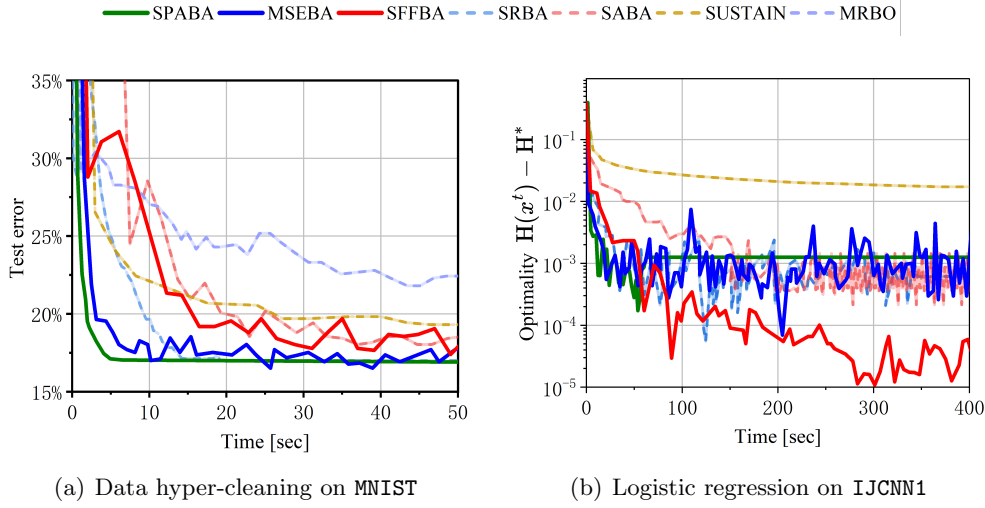


Figure 3: Comparison of SPABA, SFFBA and MSEBA with other variance-reduction-based stochastic BLO methods. **Left:** Test error for data hyper-cleaning on MNIST with $\tilde{p} = 0.9$ corruption rate, **Right:** Suboptimality gap for hyperparameter optimization for l^2 penalized logistic regression on IJCNN1 data set.

set. This data set comprises 581,012 samples, each with $\hat{p} = 54$ features, and includes $C = 7$ classes. We use $n = 371,847$ training samples, $m = 92,962$ validation samples, and $n_{\text{test}} = 116,203$ test samples.

Let $\{(d_i^{\text{train}}, y_i^{\text{train}})\}_{1 \leq i \leq n}$ and $\{(d_j^{\text{val}}, y_j^{\text{val}})\}_{1 \leq j \leq m}$ denote the training and validation sets, respectively. The bilevel objective functions are defined as $f(\theta, \lambda) = \frac{1}{m} \sum_{j=1}^m \ell(\theta d_j^{\text{val}}, y_j^{\text{val}})$, $g(\theta, \lambda) = \frac{1}{n} \sum_{i=1}^n \ell(\theta d_i^{\text{train}}, y_i^{\text{train}}) + \sum_{c=1}^C e^{\lambda_c} \sum_{i=1}^{\hat{p}} \theta_{i,c}^2$, where the LL variable $\theta \in \mathbb{R}^{54 \times 7}$ represents the model parameters, and the UL variable $\lambda \in \mathbb{R}^7$ denotes the class-wise regularization coefficients.

Figure 4(d) shows comparisons with all benchmark methods on the `covtype` data set, focusing on the test error. All algorithms exhibit a rapid decrease in test error, with the performance of SRBA and our method, SPABA, closely aligned. Both methods achieve the lowest test error value and do so at the fastest rate.

7.3 Hyperparameter Selection on the IJCNN1 Data Set

In the third task, we fit a logistic regression model (for binary classification), and select the regularization parameters (one hyperparameter per feature) on the IJCNN1 data set. Let $\{(d_i^{\text{train}}, y_i^{\text{train}})\}_{1 \leq i \leq n}$ and $\{(d_j^{\text{val}}, y_j^{\text{val}})\}_{1 \leq j \leq m}$ denote the training and validation sets, respectively.

In this context, the LL variable θ corresponds to the model parameters, while the UL variable λ represents the regularization parameters. The UL and LL objective functions are defined as $f(\lambda, \theta) = \frac{1}{m} \sum_{j=1}^m \varphi(y_j^{\text{val}} \langle d_j^{\text{val}}, \theta \rangle)$ and $g(\lambda, \theta) = \frac{1}{n} \sum_{i=1}^n \varphi(y_i^{\text{train}} \langle d_i^{\text{train}}, \theta \rangle) + \frac{1}{2} \sum_{k=1}^p e^{\lambda_k} \theta_k^2$. In this case, 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 $|\mathcal{D}_{\text{train}}| = 49,990$, and $|\mathcal{D}_{\text{val}}| = 91,701$. The LL variable $\theta \in \mathbb{R}^{22}$ is the regression coefficient, while the UL variable $\lambda \in \mathbb{R}^{22}$ is a vector of regulariza-

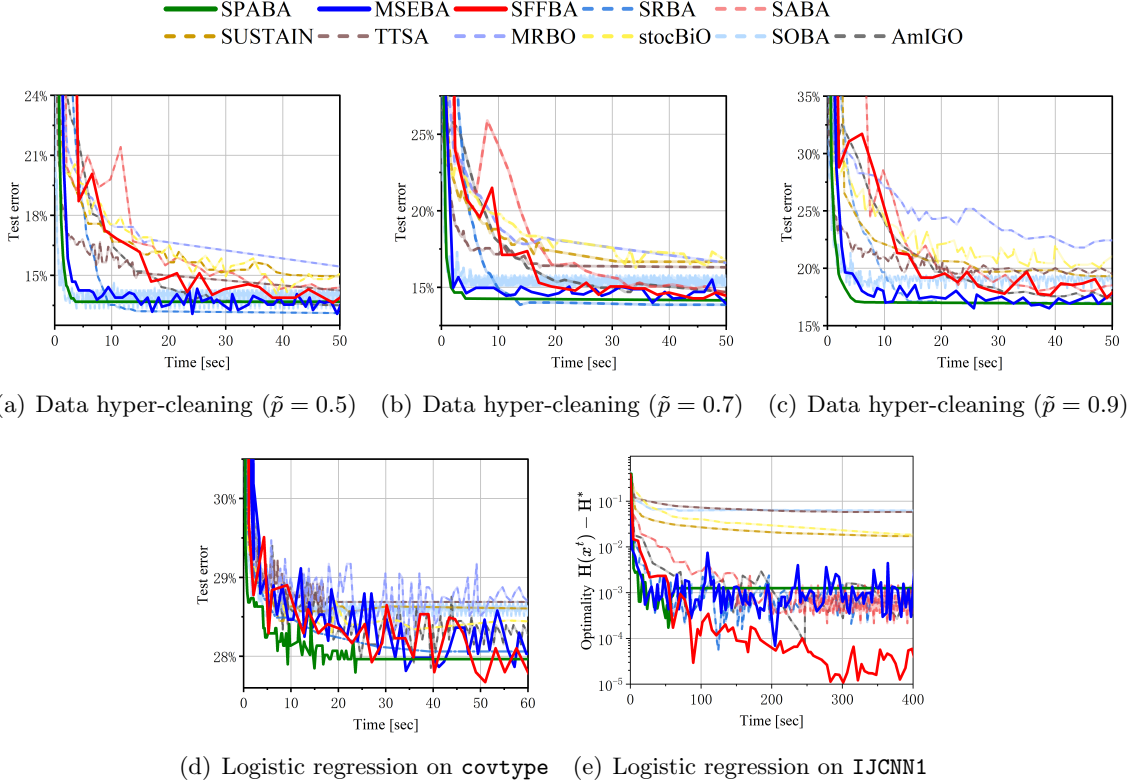


Figure 4: **Top:** Compare SPABA, SFFBA, and MSEBA with other benchmark algorithms for the data hyper-cleaning problem under different $\tilde{p}$. **Bottom:** Compare SPABA, SFFBA, and MSEBA with other benchmark algorithms for hyperparameter selection experiments on `covtype` and `IJCNN1`.

tion parameters. Unfortunately, the previously reported results for MRBO on the `IJCNN1` data set and SABA on the `covtype` data set are currently not reproducible due to conflicts with the latest developer version of `Benchopt`.

Figure 3(b) compares the suboptimality gap of our methods—SPABA, SFFBA, and MSEBA—with that of other variance-reduction-based algorithms for hyperparameter optimization in ℓ^2 -penalized logistic regression on the `IJCNN1` data set. Figure 4(e) further compares our methods with all other benchmark algorithms. We observe that SFFBA achieves the lowest values, clearly outperforming the other algorithms. Notably, SFFBA does not require computing full gradients, and its advantage over the other algorithms becomes more pronounced. Specifically, we observe that SFFBA is the only algorithm that consistently reaches an objective value below 10^{-4} . Both MSEBA and SPABA also achieve objective values close to 10^{-4} , outperforming most baseline algorithms, including SRBA.

7.4 Ablation Study

Our PnPBO framework introduces clipping for implicit variable updates and combines MA with the unbiased estimator $\mathcal{A}$. In this section, we conduct ablation studies to further demonstrate the effects of these two techniques.

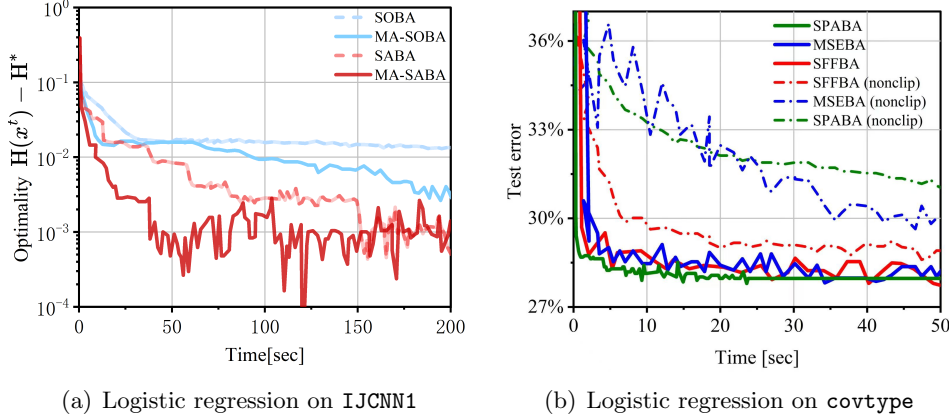


Figure 5: **Left:** Compare the performance of SOBA and MA-SOBA, and SABA and MA-SABA in the problem of hyperparameter selection experiment on the IJCNN1 data set. **Right:** Compare the performance of SPABA, SFFBA, and MSEBA with their versions without the clipping technique in the problem of hyperparameter selection experiment on the covtype data set.

First, we perform an ablation study to assess the impact of the MA technique on updating UL variables when using unbiased estimators. Figure 5(a) presents the performance comparison of SOBA and SABA with their MA-enhanced counterparts, MA-SOBA and MA-SABA, in the l^2 -penalized logistic regression task on the IJCNN1 data set. We observe that the algorithms using the MA technique reach the minimum faster, demonstrating a potential acceleration effect.

Next, we compare the use of the clipping technique in the implicit variable update for SPABA, SFFBA and MSEBA. Omitting the clipping step may degrade convergence speed, as shown in Figure 5(b).

8. Conclusion

In this work, we propose the plug-and-play stochastic BLO algorithm framework, PnPBO, which enhances the framework in (Dagr  ou et al., 2022; Arbel and Mairal, 2022) and introduces a more general convergence analysis framework. Based on this, we apply stochastic estimators such as SAGA, STORM, PAGE, ZeroSARAH, and their mixed strategies in the bilevel setting with convergence guarantees. We resolve the open question of whether the optimal complexity bounds for bilevel optimization are the same as those for single-level optimization. It’s worth mentioning that the proposed algorithm framework still relies on second-order information. An interesting and promising direction is to use value function approaches (Kwon et al., 2024; Yao et al., 2025) to develop single-loop, Hessian-free stochastic bilevel algorithms.

Acknowledgments

The authors are listed in alphabetical order. We sincerely thank the reviewers for their constructive comments, which have significantly improved this paper. We also thank the Action Editor for valuable guidance throughout the review process. This work is supported

by National Key R & D Program of China (2023YFA1011400), National Natural Science Foundation of China (12571321, 12131003), Southern University of Science and Technology High-level Special Funds (G03050K008).

Appendix A. General Lemmas

In this section, we present general results that will be used. For proof details of Lemma 19-21, see Appendix D.2. in the conference version (Chu et al., 2024).

Lemma 19 (*Lipschitz continuity of $y^*(x)$ and $z^*(x)$, boundedness of $z^*(x)$*)

(1) Under the Assumptions 2, $y^*(x)$ is L_{y^*} -Lipschitz continuous, where $L_{y^*} = L_1^g/\mu$.

(2) Under the Assumptions 1 and 2, $z^*(x)$ is L_{z^*} -Lipschitz continuous, and for each x , $\|z^*(x)\| \leq R$, where

$$L_{z^*} = \left(L^f/\mu + C^f L_2^g/\mu^2 \right) (1 + L_1^g/\mu), \quad R = C^f/\mu.$$

Lemma 20 (*smoothness of function H*)

Suppose Assumptions 1 and 2 hold, the function $H(x)$ is L^H -smooth, where

$$L^H = L^f + \frac{2L^f L_2^g + (C^f)^2 L_2^g}{\mu} + \frac{L^f (L_1^g)^2 + 2C^f L_1^g L_2^g}{\mu^2} + \frac{C^f (L_1^g)^2 L_2^g}{\mu^3}.$$

Lemma 21 Suppose Assumptions 1 and 2 hold. Then the following inequalities hold

$$\begin{aligned} \mathbb{E} \left[\|D_k^y\|^2 \right] &\leq (L_1^g)^2 \mathbb{E} [\|y_k - y_k^*\|^2], \quad \mathbb{E} \left[\|D_k^z\|^2 \right] \leq L_z^2 \mathbb{E} [\|z_k - z_k^*\|^2] + L_z^2 \mathbb{E} [\|y_k - y_k^*\|^2], \\ \mathbb{E} \left[\|D_k^x - \nabla H(x_k)\|^2 \right] &\leq 3c_1 \mathbb{E} [\|y_k - y_k^*\|^2] + 3c_2 \mathbb{E} [\|z_k - z_k^*\|^2], \end{aligned} \quad (18)$$

where $L_z^2 = \max\{3(L_1^g)^2, 3R^2(L_2^g)^2 + 3(L^f)^2\}$.

Lemma 22 Under the Assumptions 1 and 2, we have

$$\|D_{k+1}^x - D_k^x\|^2 \leq 4c_1 \left(\alpha_k^2 \|v_k^x\|^2 + \beta_k^2 \|v_k^y\|^2 \right) + 4c_2 \gamma_k^2 \|v_k^z\|^2.$$

Proof Using Assumptions 1 and 2, we have

$$\begin{aligned} \|D_{k+1}^x - D_k^x\|^2 &\leq 2 \|\nabla_1 f(x_{k+1}, y_{k+1}) - \nabla_1 f(x_k, y_k)\|^2 + 4 \|\nabla_{12}^2 g(x_k, y_k)(z_k - z_{k+1})\|^2 \\ &\quad + 4 \|\nabla_{12}^2 g(x_k, y_k)z_{k+1} - \nabla_{12}^2 g(x_{k+1}, y_{k+1})z_{k+1}\|^2 \\ &\leq \left[2(L^f)^2 + 4R^2(L_2^g)^2 \right] (\alpha_k^2 \|v_k^x\|^2 + \beta_k^2 \|v_k^y\|^2) + 4(L_1^g)^2 \gamma_k^2 \|v_k^z\|^2. \end{aligned}$$

The second inequality also uses the boundedness of z_{k+1} , which arises from the clipping. ■

Appendix B. Single-level Instantiations

In Table 3, we list representative single-level instantiations of the abstract interface inequalities from Section 4.1.

Method	Active terms in (6)–(7)	Coefficients in (8) or (9)
SGD (Unbiased) $v_k = \nabla h_{P_k}(u_k);$ step size $a_k = \mathcal{O}(K^{-1/2}),$ batch size $p_k = 1$	$\Delta_k = \sigma_h^2$ (Assumption: there exist positive constant σ_h such that $\mathbb{E}[\ \nabla h(u) - \nabla h_q(u)\ ^2] \leq \sigma_h^2$)	$\text{Coef}_k = \mathcal{O}(a_k^2)$ $\text{Coef}'_k = 0$
SAGA (Unbiased) $v_k = \nabla h_{P_k}(u_k) - \nabla h(\tilde{u}_k; P_k)$ $\quad + \nabla h(\tilde{u}_k; [Q]),$ where $[Q] = \{1, \dots, Q\};$ update $\tilde{u}_{k+1}^q = \begin{cases} u_k, & q \in P_k, \\ \tilde{u}_k^q, & q \notin P_k; \end{cases}$ step size $a_k = \mathcal{O}(Q^{-\frac{2}{3}}),$ batch size $p_k = 1$	$\theta_k = (4 + 4Q)a_k^2 L_h^2$ $\eta_k = (5 + 4Q)L_h^2$ $\tau_k = (1 - 1/(4Q))L_h^2$ $\delta_k = \frac{1}{Q} \sum_{q=1}^Q \mathbb{E}[\ u_k - \tilde{u}_k^q\ ^2]$ $\lambda_k = 1 - 1/(4Q)$ $\hat{\eta}_k = 1, \hat{\eta}'_k = 2 + 2Q$	$\text{Coef}_k = \mathcal{O}(Q^{-\frac{4}{3}})$ $\text{Coef}'_k = \mathcal{O}(Q^{-\frac{1}{3}})$
ZeroSARAH (Biased) $v_k = (1 - \hat{\rho}_k)(v_{k-1} - \nabla h_{P_k}(u_{k-1}))$ $\quad + \hat{\rho}_k(\nabla h(\tilde{u}_{k-1}; [Q]) - \nabla h(\tilde{u}_{k-1}; P_k))$ $\quad + \nabla h_{P_k}(u_k);$ update $\tilde{u}_k^q = \begin{cases} u_k, & q \in P_k, \\ \tilde{u}_{k-1}^q, & q \notin P_k; \end{cases}$ step size $a_k = \mathcal{O}(1),$ parameter $\hat{\rho}_k = \mathcal{O}(1/\sqrt{Q}),$ batch size $p = \mathcal{O}(\sqrt{Q})$	$\theta_k = (1 - \hat{\rho}_{k+1})^2$ $\eta_k = 2L_h^2/p$ $\tau_k = 2L_h^2\hat{\rho}_k^2/p$ $\delta_k = \frac{1}{Q} \sum_{q=1}^Q \mathbb{E}[\ u_k - \tilde{u}_k^q\ ^2]$ $\lambda_k = 1 - p_k/2Q$ $\hat{\eta}_k = 3Q/p$	$\text{Coef}_k = \mathcal{O}\left(\frac{\alpha_k}{\hat{\rho}_k}\right)$ $\text{Coef}'_k = \mathcal{O}(\alpha_k \hat{\rho}_k)$
PAGE (Biased) with probability $\hat{\rho},$ set $v_k = \nabla h(u_k);$ otherwise, set $v_k = v_{k-1} + \nabla h_{P_k}(u_k) - \nabla h_{P_k}(u_{k-1})$ step size $a_k = \mathcal{O}(1),$ batch size $p = \sqrt{Q},$ probability $\hat{\rho} = p_k/(Q + p_k)$	$\theta_k = (1 - \hat{\rho})$ $\eta_k = (1 - \hat{\rho})L^2/p$	$\text{Coef}_k = \mathcal{O}(\alpha_k/\hat{\rho})$ $\text{Coef}'_k = 0$

Table 3: Single-level instantiations of the abstract inequalities in Section 4.1.

In the *Active terms* column, we report only nonzero coefficients.

The SAGA row is based on Lemma C.5 in (Dagr  ou et al., 2022); the coefficients reported here are obtained by specializing that lemma to the single-level finite-sum problem and applying the bound $\mathbb{E}[\|v_k - \nabla h(u_k)\|^2] \leq L_h^2 \delta_k$ together with simple constant relaxations. The ZeroSARAH row follows from Lemmas 2–3 in (Li et al., 2021b). The PAGE row follows from Lemma 3 in (Li et al., 2021a).

Appendix C. The Missing Proof in Section 4

Proof (Proof of Lemma 6) When $\mathcal{A}$ is an unbiased estimator, it is necessary to combine it with the MA technique, i.e., $v_{k+1}^x = (1 - \rho_k)v_k^x + \rho_k \hat{v}_{k+1}^x$. Using the unbiasedness of $\hat{v}_{k+1}^x$,

we obtain

$$\begin{aligned} \mathbb{E}[\|v_{k+1}^x - \nabla H(x_{k+1})\|^2] &= \mathbb{E}\left[\left\| (1 - \rho_k)(v_k^x - \nabla H(x_k)) + \nabla H(x_k) - \nabla H(x_{k+1}) + \rho_k D_k^x \right.\right. \\ &\quad \left.\left. - \rho_k \nabla H(x_k) + \rho_k (D_{k+1}^x - D_k^x) \right\|^2\right] + \rho_k^2 \mathbb{E}[\|\hat{v}_{k+1}^x - D_{k+1}^x\|^2]. \end{aligned}$$

Based on the convexity of $\|\cdot\|^2$, we can derive the following

$$\begin{aligned} \mathbb{E}[\|v_{k+1}^x - \nabla H(x_{k+1})\|^2] &\leq (1 - \rho_k) \mathbb{E}[\|v_k^x - \nabla H(x_k)\|^2] + \rho_k^2 \mathbb{E}[\|\hat{v}_{k+1}^x - D_{k+1}^x\|^2] \\ &\quad + \rho_k \mathbb{E}\left[\left\| D_{k+1}^x - D_k^x + D_k^x - \nabla H(x_k) + \frac{\nabla H(x_k) - \nabla H(x_{k+1})}{\rho_k} \right\|^2\right] \\ &\leq (1 - \rho_k) \mathbb{E}[\|v_k^x - \nabla H(x_k)\|^2] + \rho_k^2 \mathbb{E}[\|\hat{v}_{k+1}^x - D_{k+1}^x\|^2] + 3\rho_k \mathbb{E}[\|D_{k+1}^x - D_k^x\|^2] \\ &\quad + 3\rho_k \mathbb{E}[\|D_k^x - \nabla H(x_k)\|^2] + \frac{3}{\rho_k} \mathbb{E}[\|\nabla H(x_k) - \nabla H(x_{k+1})\|^2]. \end{aligned}$$

Using (18), Lemma 22 and the L^H -smoothness of H , we conclude the proof. $\blacksquare$

Proof (Proof of Lemma 10)

Based on the definition of z_{k+1} and utilizing Young's inequality twice, we obtain

$$\|z_{k+1} - z^*(x_{k+1})\|^2 \leq (1 + \delta'_k) \|z_{k+1} - z^*(x_k)\|^2 + \left(1 + \frac{1}{\delta'_k}\right) L_{z^*}^2 \alpha_k^2 \|v_k^x\|^2.$$

Since $\|z^*(x_k)\| \leq R$, the clipping operator $\text{Clip}(\cdot; R)$ coincides with the Euclidean projection onto $B(0, R)$, and hence it is nonexpansive. Using this nonexpansiveness and applying Young's inequality, we have

$$\begin{aligned} \|z_{k+1} - z^*(x_k)\|^2 &= \|\text{Clip}(z_k - \gamma_k v_k^z; R) - z^*(x_k)\|^2 \\ &\leq \|z_k - \gamma_k v_k^z - z^*(x_k)\|^2 \\ &\leq \left(1 + \frac{\mu\gamma_k}{4}\right) \|z_k - \gamma_k D_k^z - z^*(x_k)\|^2 + \left(1 + \frac{4}{\mu\gamma_k}\right) \|\gamma_k (v_k^z - D_k^z)\|^2. \end{aligned}$$

For the first term, we use the identity $\nabla_{22}^2 g(x_k, y^*(x_k)) z^*(x_k) - \nabla_2 f(x_k, y^*(x_k)) = 0$ to obtain the decomposition

$$\begin{aligned} \|z_k - \gamma_k D_k^z - z^*(x_k)\|^2 &= \left\| (I - \gamma_k \nabla_{22}^2 g(x_k, y_k))(z_k - z^*(x_k)) \right. \\ &\quad \left. + \gamma_k ((\nabla_{22}^2 g(x_k, y^*(x_k)) - \nabla_{22}^2 g(x_k, y_k)) z^*(x_k) + \nabla_2 f(x_k, y_k) - \nabla_2 f(x_k, y^*(x_k))) \right\|^2. \end{aligned}$$

Applying Young's inequality yields

$$\begin{aligned} \|z_k - \gamma_k D_k^z - z^*(x_k)\|^2 &\leq \left(1 + \frac{\gamma_k \mu}{2}\right) \|(I - \gamma_k \nabla_{22}^2 g(x_k, y_k))(z_k - z^*(x_k))\|^2 \\ &\quad + \left(2 + \frac{4}{\gamma_k \mu}\right) \gamma_k^2 \|\nabla_{22}^2 g(x_k, y^*(x_k)) - \nabla_{22}^2 g(x_k, y_k)\|^2 \|z^*(x_k)\|^2 \\ &\quad + \left(2 + \frac{4}{\gamma_k \mu}\right) \gamma_k^2 \|\nabla_2 f(x_k, y_k) - \nabla_2 f(x_k, y^*(x_k))\|^2. \end{aligned}$$

Using Assumption 2(a) and the step size condition $\gamma_k \leq 1/L_1^g$, we have

$$\|I - \gamma_k \nabla_{22}^2 g(x_k, y_k)\|^2 \leq (1 - \gamma_k \mu)^2.$$

Moreover, by Assumptions 1(b) and 2(b), together with $\|z^*(x_k)\| \leq R$, we obtain

$$\begin{aligned} & \|z_k - \gamma_k D_k^z - z^*(x_k)\|^2 \\ & \leq \left(1 + \frac{\gamma_k \mu}{2}\right) (1 - \gamma_k \mu)^2 \|z_k - z^*(x_k)\|^2 + \left(2 + \frac{4}{\gamma_k \mu}\right) \gamma_k^2 c_1 \|y_k - y^*(x_k)\|^2, \end{aligned}$$

where $c_1 = (L_2^g R)^2 + (L^f)^2$.

Rearranging the above inequalities and taking the total expectation yields

$$\begin{aligned} \mathbb{E} [\|z_{k+1} - z^*(x_{k+1})\|^2] & \leq \left(1 + \frac{\mu \gamma_k}{4}\right) \left(1 + \frac{\gamma_k \mu}{2}\right)^2 (1 - \gamma_k \mu)^2 \mathbb{E} [\|z_k - z^*(x_k)\|^2] \\ & \quad + \left(1 + \frac{\mu \gamma_k}{4}\right) \left(1 + \frac{\gamma_k \mu}{2}\right) \left(2 + \frac{4}{\gamma_k \mu}\right) \gamma_k^2 c_1 \mathbb{E} [\|y_k - y^*(x_k)\|^2] \\ & \quad + \left(1 + \frac{\mu \gamma_k}{2}\right) \left(1 + \frac{4}{\mu \gamma_k}\right) \gamma_k^2 \mathbb{E} [\|v_k^z - D_k^z\|^2] \\ & \quad + \left(1 + \frac{2}{\mu \gamma_k}\right) L_{z^*}^2 \alpha_k^2 \mathbb{E} [\|v_k^x\|^2]. \end{aligned}$$

Using the step size condition $\gamma_k \leq \frac{1}{L_1^g + \mu}$, we have

$$\begin{aligned} \mathbb{E} [\|z_{k+1} - z^*(x_{k+1})\|^2] & \leq \left(1 - \frac{\gamma_k \mu}{4}\right) \mathbb{E} [\|z_k - z^*(x_k)\|^2] + \frac{18c_1 \gamma_k}{\mu} \mathbb{E} [\|y_k - y^*(x_k)\|^2] \\ & \quad + \frac{12\gamma_k}{\mu} \mathbb{E} [\|v_k^z - D_k^z\|^2] + \frac{4L_{z^*}^2 \alpha_k^2}{\mu \gamma_k} \mathbb{E} [\|v_k^x\|^2]. \end{aligned}$$

■

Proof (Proofs of the Remaining Lemmas and Corollary):

They have already been proved in our conference version. Specifically, Lemma 4 can be found in Lemma D.7 of the conference version, Corollary 5 in Corollary F.1, Lemma 7 in Lemma F.3, Lemmas 8 and 11 in Lemma E.2. ■

Appendix D. Variance Results of Stochastic Estimators in the Bilevel Setting

For the stochastic estimator $\mathcal{B}$, we have

$$\begin{aligned} \mathbb{E} [\|D_{k+1}^y - v_{k+1}^y\|^2] & \leq \theta_k^{\mathcal{B}} \mathbb{E} [\|D_k^y - v_k^y\|^2] + \eta_k^{\mathcal{B}} c_2 \left(\alpha_k^2 \mathbb{E} [\|v_k^x\|^2] + \beta_k^2 \mathbb{E} [\|v_k^y\|^2] \right) \\ & \quad + \tau_k^{\mathcal{B}} \delta_k^{\mathcal{B}} + \Delta_k^{\mathcal{B}}, \end{aligned} \tag{19}$$

and

$$\delta_{k+1}^{\mathcal{B}} \leq \lambda_k^{\mathcal{B}} \delta_k^{\mathcal{B}} + \hat{\eta}_k^{\mathcal{B},x} \alpha_k^2 \mathbb{E} [\|v_k^x\|^2] + \hat{\eta}_k^{\mathcal{B},y} \beta_k^2 \mathbb{E} [\|v_k^y\|^2] + \hat{\eta}_k^{\mathcal{B},x} \alpha_k^2 \mathbb{E} [\|D_k^x\|^2] + \hat{\eta}_k^{\mathcal{B},y} \beta_k^2 \mathbb{E} [\|D_k^y\|^2]. \tag{20}$$

For the stochastic estimator $\mathcal{C}$, we have

$$\begin{aligned} \mathbb{E} \left[\|D_{k+1}^z - v_{k+1}^z\|^2 \right] &\leq \theta_k^{\mathcal{C}} \mathbb{E} \left[\|D_k^z - v_k^z\|^2 \right] + \eta_k^{\mathcal{C}} c_1 \alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + \eta_k^{\mathcal{C}} c_1 \beta_k^2 \mathbb{E}[\|v_k^y\|^2] \\ &\quad + \eta_k^{\mathcal{C}} c_2 \gamma_k^2 \mathbb{E}[\|v_k^z\|^2] + \tau_k^{\mathcal{C}} \delta_k^{\mathcal{C}} + \Delta_k^{\mathcal{C}}, \end{aligned} \quad (21)$$

and

$$\begin{aligned} \delta_{k+1}^{\mathcal{C}} &\leq \lambda_k^{\mathcal{C}} \delta_k^{\mathcal{C}} + \hat{\eta}_k^{\mathcal{C},x} \alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + \hat{\eta}_k^{\mathcal{C},y} \beta_k^2 \mathbb{E}[\|v_k^y\|^2] + \hat{\eta}_k^{\mathcal{C},z} \gamma_k^2 \mathbb{E}[\|v_k^z\|^2] \\ &\quad + \hat{\eta}_k^{\mathcal{C},x} \alpha_k^2 \mathbb{E}[\|D_k^x\|^2] + \hat{\eta}_k^{\mathcal{C},y} \beta_k^2 \mathbb{E}[\|D_k^y\|^2] + \hat{\eta}_k^{\mathcal{C},z} \gamma_k^2 \mathbb{E}[\|D_k^z\|^2]. \end{aligned} \quad (22)$$

Proof (Proof of Inequalities (10), (11), and (19)-(22))

We take $\mathcal{A}$ as an illustrative example. Since $\mathcal{A}$ satisfies the SE-VRC condition, we have

$$\begin{aligned} &\mathbb{E} \left[\|D_{k+1}^x - \hat{v}_{k+1}^x\|^2 \right] \\ &\leq \theta_k^{\mathcal{A}} \mathbb{E} \left[\|D_k^x - \hat{v}_k^x\|^2 \right] + \eta_k^{\mathcal{A}} \mathbb{E} \left[\|\nabla_1 F_i(x_{k+1}, y_{k+1}) - \nabla_1 F_i(x_k, y_k)\|^2 \right] \\ &\quad + \eta_k^{\mathcal{A}} \mathbb{E} \left[\|\nabla_{12}^2 G_j(x_{k+1}, y_{k+1}) z_{k+1} - \nabla_{12}^2 G_j(x_k, y_k) z_k\|^2 \right] \\ &\quad + \tau_k^{\mathcal{A}} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\nabla_1 F_i(x_k, y_k) - \nabla_1 F_i(w_{k,i}^x, w_{k,i}^y)\|^2] \\ &\quad + \tau_k^{\mathcal{A}} \frac{1}{m} \sum_{j=1}^m \mathbb{E}[\|\nabla_{12}^2 G_j(x_k, y_k) z_k - \nabla_{12}^2 G_j(\tilde{w}_{k,j}^x, \tilde{w}_{k,j}^y) \tilde{w}_{k,j}^z\|^2] + \Delta_k^{\mathcal{A}}, \end{aligned} \quad (23)$$

where, conditional on I_{k+1} and J_{k+1} , the indices i and j are sampled independently and uniformly from I_{k+1} and J_{k+1} , respectively. Here $w_{k,i} = (w_{k,i}^x, w_{k,i}^y)$ for $i \in [n]$ and $\tilde{w}_{k,j} = (\tilde{w}_{k,j}^x, \tilde{w}_{k,j}^y, \tilde{w}_{k,j}^z)$ for $j \in [m]$ are the memory variables maintained by $\mathcal{A}$ (when memory is required), corresponding to the oracle calls for f and g , respectively. We define

$$\delta_k^{f,x} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|x_k - w_{k,i}^x\|^2], \quad \delta_k^{g,x} = \frac{1}{m} \sum_{j=1}^m \mathbb{E}[\|x_k - \tilde{w}_{k,j}^x\|^2]$$

and analogously define $\delta_k^{f,y}$, $\delta_k^{g,y}$, $\delta_k^{g,z}$.

For the second and third terms on the right-hand side of (23), we can bound them as follows:

$$\eta_k^{\mathcal{A}} \mathbb{E} \left[\|\nabla_1 F_i(x_{k+1}, y_{k+1}) - \nabla_1 F_i(x_k, y_k)\|^2 \right] \leq \eta_k^{\mathcal{A}} (L^f)^2 (\alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + \beta_k^2 \mathbb{E}[\|v_k^y\|^2]), \quad (24)$$

$$\begin{aligned} &\eta_k^{\mathcal{A}} \mathbb{E} \left[\|\nabla_{12}^2 G_j(x_{k+1}, y_{k+1}) z_{k+1} - \nabla_{12}^2 G_j(x_k, y_k) z_k\|^2 \right] \\ &\leq 2\eta_k^{\mathcal{A}} \mathbb{E} \left[\|\nabla_{12}^2 G_j(x_k, y_k)(z_{k+1} - z_k)\|^2 \right] \\ &\quad + 2\eta_k^{\mathcal{A}} \mathbb{E}[\|(\nabla_{12}^2 G_j(x_{k+1}, y_{k+1}) - \nabla_{12}^2 G_j(x_k, y_k))\|^2 \|z_{k+1}\|^2] \\ &\leq 2R^2 (L_2^g)^2 \eta_k^{\mathcal{A}} \alpha_k^2 (\mathbb{E}[\|v_k^x\|^2] + \beta_k^2 \mathbb{E}[\|v_k^y\|^2]) + 2(L_1^g)^2 \eta_k^{\mathcal{A}} \gamma_k^2 \mathbb{E}[\|v_k^z\|^2]. \end{aligned} \quad (25)$$

For the fourth term on the right-hand side of (23), by Assumption 1, we have

$$\tau_k^{\mathcal{A}} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\nabla F_i(x_k, y_k) - \nabla F_i(w_{k,i}^x, w_{k,i}^y)\|^2] \leq \tau_k^{\mathcal{A}} (L^f)^2 \delta_k^{f,x} + \tau_k^{\mathcal{A}} (L^f)^2 \delta_k^{f,y}. \quad (26)$$

For the fifth term on the right-hand side of (23), invoking Assumption 2 and Lemma 19, and adding and subtracting $\nabla_{12}^2 G_j(\tilde{w}_{k,j}^x, \tilde{w}_{k,j}^y)z_k$ yield

$$\begin{aligned} & \tau_k'^A \frac{1}{m} \sum_{j=1}^m \mathbb{E}[\|\nabla_{12}^2 G_j(x_k, y_k)z_k - \nabla_{12}^2 G_j(\tilde{w}_{k,j}^x, \tilde{w}_{k,j}^y)\tilde{w}_{k,j}^z\|^2] \\ & \leq 2\tau_k'^A (L_2^g)^2 R^2 (\delta_k^{g,x} + \delta_k^{g,y}) + 2\tau_k'^A (L_1^g)^2 \delta_k^{g,z}. \end{aligned} \quad (27)$$

Substituting the above inequalities into (23) yields

$$\begin{aligned} \mathbb{E}[\|D_{k+1}^x - \hat{v}_{k+1}^x\|^2] & \leq \theta_k^A \mathbb{E}[\|D_k^x - \hat{v}_k^x\|^2] + 2c_1 \eta_k'^A \alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + 2c_1 \eta_k'^A \beta_k^2 \mathbb{E}[\|v_k^y\|^2] \\ & \quad + 2c_2 \eta_k'^A \gamma_k^2 \mathbb{E}[\|v_k^z\|^2] + \tau_k'^A (L^f)^2 (\delta_k^{f,x} + \delta_k^{f,y}) \\ & \quad + 2\tau_k'^A (L_2^g)^2 R^2 (\delta_k^{g,x} + \delta_k^{g,y}) + 2\tau_k'^A (L_1^g)^2 \delta_k^{g,z} + \Delta_k^A \\ & \leq \theta_k^A \mathbb{E}[\|D_k^x - \hat{v}_k^x\|^2] + \tau_k^A \delta_k^A + \Delta_k^A \\ & \quad + \eta_k^A (c_1 \alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + c_1 \beta_k^2 \mathbb{E}[\|v_k^y\|^2] + c_2 \gamma_k^2 \mathbb{E}[\|v_k^z\|^2]), \end{aligned}$$

where we set $\eta_k^A = 2\eta_k'^A$, $\tau_k^A = \tau_k'^A L''$, $L'' = \max\{(L^f)^2, 2(L_2^g R)^2, 2(L_1^g)^2\}$, and $\delta_k^A := \delta_k^{f,x} + \delta_k^{f,y} + \delta_k^{g,x} + \delta_k^{g,y} + \delta_k^{g,z}$.

Based on the memory-update rule of $\mathcal{A}$, which satisfies the SE-VRC condition, the mean-squared staleness admits the recursions

$$\delta_{k+1}^{f,x} \leq \lambda_k^{A,n} \delta_k^{f,x} + \hat{\eta}_k^{A,n,x} \alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + \hat{\eta}_k^{A,n,x} \alpha_k^2 \mathbb{E}[\|D_k^x\|^2], \quad (28a)$$

$$\delta_{k+1}^{g,x} \leq \lambda_k^{A,m} \delta_k^{g,x} + \hat{\eta}_k^{A,m,x} \alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + \hat{\eta}_k^{A,m,x} \alpha_k^2 \mathbb{E}[\|D_k^x\|^2], \quad (28b)$$

$$\delta_{k+1}^{f,y} \leq \lambda_k^{A,n} \delta_k^{f,y} + \hat{\eta}_k^{A,n,y} \beta_k^2 \mathbb{E}[\|v_k^y\|^2] + \hat{\eta}_k^{A,n,y} \beta_k^2 \mathbb{E}[\|D_k^y\|^2], \quad (28c)$$

$$\delta_{k+1}^{g,y} \leq \lambda_k^{A,m} \delta_k^{g,y} + \hat{\eta}_k^{A,m,y} \beta_k^2 \mathbb{E}[\|v_k^y\|^2] + \hat{\eta}_k^{A,m,y} \beta_k^2 \mathbb{E}[\|D_k^y\|^2], \quad (28d)$$

$$\delta_{k+1}^{g,z} \leq \lambda_k^{A,m} \delta_k^{g,z} + \hat{\eta}_k^{A,m,z} \gamma_k^2 \mathbb{E}[\|v_k^z\|^2] + \hat{\eta}_k^{A,m,z} \gamma_k^2 \mathbb{E}[\|D_k^z\|^2]. \quad (28e)$$

Note that the staleness terms are induced by the memory tables maintained by $\mathcal{A}$. The coefficients in the recursion may depend on the data set sizes n and m through the sampling and refresh mechanisms. In addition, $\hat{\eta}_k^{A,\cdot,\cdot}$ and $\hat{\eta}_k^{A,\cdot,\cdot}$ may vary with the properties of the update directions (v_k^x, v_k^y, v_k^z) (e.g., biased vs. unbiased), generated by $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, respectively. We therefore distinguish these coefficients using the superscripts in (28). Define $\hat{\eta}_k^{A,x} := \hat{\eta}_k^{A,n,x} + \hat{\eta}_k^{A,m,x}$, $\hat{\eta}_k^{A,y} := \hat{\eta}_k^{A,n,y} + \hat{\eta}_k^{A,m,y}$, with $\hat{\eta}_k^{A,y}$, $\hat{\eta}_k^{A,z}$, $\hat{\eta}_k^{A,y}$, and $\hat{\eta}_k^{A,z}$ defined analogously. Summing (28) yields

$$\begin{aligned} \delta_{k+1}^A & \leq \lambda_k^A \delta_k^A + \hat{\eta}_k^{A,x} \alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + \hat{\eta}_k^{A,y} \beta_k^2 \mathbb{E}[\|v_k^y\|^2] + \hat{\eta}_k^{A,z} \gamma_k^2 \mathbb{E}[\|v_k^z\|^2] \\ & \quad + \hat{\eta}_k^{A,x} \alpha_k^2 \mathbb{E}[\|D_k^x\|^2] + \hat{\eta}_k^{A,y} \beta_k^2 \mathbb{E}[\|D_k^y\|^2] + \hat{\eta}_k^{A,z} \gamma_k^2 \mathbb{E}[\|D_k^z\|^2]. \end{aligned}$$

The proofs of (19)–(22) follow analogously and are omitted. ■

Appendix E. Verification of SFFBA Satisfying the SE-VRC condition.

Lemma 23 *Under the conditions stated in Assumptions 1 and 2, SFFBA satisfies the SE-VRC condition and we have*

$$\begin{aligned}\theta_k^A &= \theta_k^B = \theta_k^C = (1 - \bar{\rho}_{k+1})^2, \quad \eta_k^A = \eta_k^B = \eta_k^C = \frac{4}{b}, \quad \tau_k^A = \tau_k^B = \tau_k^C = \frac{2\bar{\rho}_{k+1}^2 L''}{b}, \\ \delta_k^A &= \delta_k^C = \delta_k^B = \delta_k, \quad \lambda_k^A = \lambda_k^B = \lambda_k^C = 1 - \frac{b}{2N}, \quad \Delta_k^A = \Delta_k^B = \Delta_k^C = 0, \\ \hat{\eta}_k^{A,\ell} &= \hat{\eta}_k^{B,\ell} = \hat{\eta}_k^{C,\ell} = \frac{3N}{b}, \quad \hat{\eta}_k^{A,\ell} = \hat{\eta}_k^{B,\ell} = \hat{\eta}_k^{C,\ell} = 0, \quad \forall \ell \in \{x, y, z\},\end{aligned}$$

where $\delta_k = \delta_k^{f,x} + \delta_k^{g,x} + \delta_k^{f,y} + \delta_k^{g,y} + \delta_k^{g,z}$, $L'' = \max\{(L^f)^2, 2(L_2^g R)^2, 2(L_1^g)^2\}$.

Proof According to Lemma 2 in Li et al. (2021b), we can directly obtain that the SE-VRC condition holds and the selection of the parameters. For the sake of completeness in the proof, we provide a detailed proof in BLO setting for $\mathbb{E}[\|\hat{v}_k^x - D_k^x\|^2]$ and $E_{k,f}^x$ as an example.

First, for the descent property of $\mathbb{E}[\|\hat{v}_k^x - D_k^x\|^2]$, from the update rule of $\hat{v}_k^x$, we have

$$\begin{aligned}\mathbb{E}[\|v_{k+1}^x - D_{k+1}^x\|^2] &= \mathbb{E}\left[\left\|(1 - \bar{\rho}_{k+1})(v_k^x - D_k^x) + D_{k+1;I,J}^x - D_{k;I,J}^x + D_k^x - D_{k+1}^x\right.\right. \\ &\quad \left.\left.+ \bar{\rho}_{k+1}\left(D_{k;I,J}^x - D_k^x - \hat{D}_{k;I,J}^x(w, \tilde{w}) + \hat{D}_{k;[n],[m]}^x(w, \tilde{w})\right)\right\|^2\right] \\ &\leq (1 - \bar{\rho}_{k+1})^2 \mathbb{E}[\|v_k^x - D_k^x\|^2] + 2\mathbb{E}[\|D_{k+1;I,J}^x - D_{k;I,J}^x + D_k^x - D_{k+1}^x\|^2] \\ &\quad + 2\bar{\rho}_{k+1}^2 \mathbb{E}[\|D_{k;I,J}^x - D_k^x - \hat{D}_{k;I,J}^x(w, \tilde{w}) + \hat{D}_{k;[n],[m]}^x(w, \tilde{w})\|^2] \\ &\leq (1 - \bar{\rho}_{k+1})^2 \mathbb{E}[\|v_k^x - D_k^x\|^2] + \frac{2}{b^2} \sum_{i \in I} \mathbb{E}[\|\nabla_1 F_i(x_{k+1}, y_{k+1}) - \nabla_1 F_i(x_k, y_k)\|^2] \\ &\quad + \frac{2}{b^2} \sum_{j \in J} \mathbb{E}[\|\nabla_{12}^2 G_j(x_{k+1}, y_{k+1}) z_{k+1} - \nabla_{12}^2 G_j(x_k, y_k) z_k\|^2] \\ &\quad + \frac{2\bar{\rho}_{k+1}^2}{b} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\nabla_1 F_i(x_k, y_k) - \nabla_1 F_i(w_{k,i}^x, w_{k,i}^y)\|^2] \\ &\quad + \frac{2\bar{\rho}_{k+1}^2}{b} \frac{1}{m} \sum_{j=1}^m \mathbb{E}[\|\nabla_{12}^2 G_j(x_k, y_k) z_k - \nabla_{12}^2 G_j(\tilde{w}_{k,j}^x, \tilde{w}_{k,j}^y) \tilde{w}_{k,j}^z\|^2],\end{aligned}$$

where $\hat{v}_k^x = v_k^x$, the first inequality uses that $D_{k+1;I,J}^x - D_{k;I,J}^x$ is an unbiased estimator of $D_{k+1}^x - D_k^x$ and that $D_{k;I,J}^x - \hat{D}_{k;I,J}^x(w, \tilde{w})$ is an unbiased estimator of $D_k^x - \hat{D}_{k;[n],[m]}^x(w, \tilde{w})$, so the cross terms vanish; we then apply $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. The second inequality follows from using the independence between the i - and j -samples, and applying $\mathbb{E}[\|X - \mathbb{E}X\|^2] \leq \mathbb{E}[\|X\|^2]$.

Following the derivations in (24)–(27), we further obtain

$$\begin{aligned}\mathbb{E}[\|v_{k+1}^x - D_{k+1}^x\|^2] &\leq (1 - \bar{\rho}_{k+1})^2 \mathbb{E}[\|v_k^x - D_k^x\|^2] + \frac{2L''\bar{\rho}_{k+1}^2}{b} \delta_k \\ &\quad + \frac{4}{b} \left(c_1 \alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + c_1 \beta_k^2 \mathbb{E}[\|v_k^y\|^2] + c_2 \gamma_k^2 \mathbb{E}[\|v_k^z\|^2] \right),\end{aligned}$$

where $\delta_k = \delta_k^{f,x} + \delta_k^{g,x} + \delta_k^{f,y} + \delta_k^{g,y} + \delta_k^{g,z}$.

Next, for the descent property of $\delta_k^{f,x}$, we have

$$\begin{aligned} E_k[\|x_{k+1} - w_{k+1,i}^x\|^2] &= \left(1 - \frac{b}{n}\right) E_k[\|x_{k+1} - w_{k,i}^x\|^2] \\ &\leq \left(1 - \frac{b}{n}\right) \left(1 + \frac{b}{2n}\right) E_k[\|x_k - w_{k,i}^x\|^2] + \left(1 - \frac{b}{n}\right) \left(1 + \frac{2n}{b}\right) E_k[\|x_{k+1} - x_k\|^2] \\ &\leq \left(1 - \frac{b}{2n}\right) E_k[\|x_k - w_{k,i}^x\|^2] + \frac{3n\alpha_k^2}{b} E_k[\|v_k^x\|^2]. \end{aligned}$$

Taking the full expectation yields

$$\delta_{k+1}^{f,x} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|x_{k+1} - w_{k+1,i}^x\|^2] \leq \left(1 - \frac{b}{2n}\right) \delta_k^{f,x} + \frac{3n\alpha_k^2}{b} \mathbb{E}[\|v_k^x\|^2].$$

Similarly, the same reasoning applies to the other terms in δ_k . Then we have

$$\delta_{k+1} \leq \left(1 - \frac{b}{2N}\right) \delta_k + \frac{3N\alpha_k^2}{b} \mathbb{E}[\|v_k^x\|^2] + \frac{3N\beta_k^2}{b} \mathbb{E}[\|v_k^y\|^2] + \frac{3N\gamma_k^2}{b} \mathbb{E}[\|v_k^z\|^2].$$

■

Appendix F. Proof of Theorem 14

Proof Since the stochastic estimator $\mathcal{A}$ is biased, it follows that $v_k^x = \hat{v}_k^x$, and we set $A_k'' = 0$. Under the conditions of Assumptions 1, 2, and the SE-VRC condition, we have that Corollary 5, Lemmas 7 and 10, and the equations (10)-(11) and (19)-(22) hold. Based on these results, and by combining Lemma 21 with the following inequalities

$$\mathbb{E}[\|v_k^y\|^2] \leq 2\mathbb{E}[\|v_k^y - D_k^y\|^2] + 2\mathbb{E}[\|D_k^y\|^2], \quad \mathbb{E}[\|v_k^z\|^2] \leq 2\mathbb{E}[\|v_k^z - D_k^z\|^2] + 2\mathbb{E}[\|D_k^z\|^2],$$

we obtain

$$\begin{aligned} L_{k+1} - L_k &\leq -\frac{\alpha_k}{2} \mathbb{E}[\|\nabla H(x_k)\|^2] + \text{Part}_1 \mathbb{E}[\|v_k^x\|^2] + \text{Part}_2 \mathbb{E}[\|y_k - y_k^*\|^2] \\ &\quad + \text{Part}_3 \mathbb{E}[\|z_k - z_k^*\|^2] + \text{Part}_4 \mathbb{E}[\|D_k^x - v_k^x\|^2] + \text{Part}_5 \mathbb{E}[\|v_k^y - D_k^y\|^2] \\ &\quad + \text{Part}_6 \mathbb{E}[\|D_k^z - v_k^z\|^2] + \text{Part}_7 \delta_k^{\mathcal{A}} + \text{Part}_8 \delta_k^{\mathcal{B}} + \text{Part}_9 \delta_k^{\mathcal{C}} + A_{k+1} \Delta_k^{\mathcal{A}} + B_{k+1} \Delta_k^{\mathcal{B}} + C_{k+1} \Delta_k^{\mathcal{C}}, \end{aligned}$$

where

$$\begin{aligned} \text{Part}_1 &= -\frac{\alpha_k}{4} + c_5 C_y \frac{\alpha_k^2}{\beta_k} + c_7 C_z \frac{\alpha_k^2}{\gamma_k} + 2(A'_{k+1} \hat{\eta}'^{\mathcal{A},x} + B'_{k+1} \hat{\eta}'^{\mathcal{B},x} + C'_{k+1} \hat{\eta}'^{\mathcal{C},x}) \alpha_k^2, \\ &\quad + \left(A_{k+1} \eta_k^{\mathcal{A}} c_1 + A'_{k+1} \hat{\eta}_k^{\mathcal{A},x} + B_{k+1} c_2 \eta_k^{\mathcal{B}} + B'_{k+1} \hat{\eta}_k^{\mathcal{B},x} + C_{k+1} c_1 \eta_k^{\mathcal{C}} + C'_{k+1} \hat{\eta}_k^{\mathcal{C},x}\right) \alpha_k^2, \\ \text{Part}_2 &= 3c_1 \alpha_k - c_3 C_y \beta_k + \frac{18C_z c_1}{\mu} \gamma_k + c_2 \left(2B_{k+1} c_2 \eta_k^{\mathcal{B}} + B'_{k+1} (2\hat{\eta}_k^{\mathcal{B},y} + \hat{\eta}_k'^{\mathcal{B},y})\right) \beta_k^2 \\ &\quad + c_2 \left(2A_{k+1} c_1 \eta_k^{\mathcal{A}} + A'_{k+1} (2\hat{\eta}_k^{\mathcal{A},y} + \hat{\eta}_k'^{\mathcal{A},y}) + 2C_{k+1} c_1 \eta_k^{\mathcal{C}} + C'_{k+1} (2\hat{\eta}_k^{\mathcal{C},y} + \hat{\eta}_k'^{\mathcal{C},y})\right) \beta_k^2 \\ &\quad + L_z^2 \left(2A_{k+1} c_2 \eta_k^{\mathcal{A}} + A'_{k+1} (2\hat{\eta}_k^{\mathcal{A},z} + \hat{\eta}_k'^{\mathcal{A},z}) + 2C_{k+1} c_2 \eta_k^{\mathcal{C}} + C'_{k+1} (2\hat{\eta}_k^{\mathcal{C},z} + \hat{\eta}_k'^{\mathcal{C},z})\right) \gamma_k^2, \end{aligned}$$

$$\begin{aligned}
 \text{Part}_3 &= 3c_2\alpha_k - \frac{C_z\mu}{4}\gamma_k + L_z^2 \left(2A_{k+1}c_2\eta_k^A + 2A'_{k+1}\hat{\eta}_k^{A,z} + A'_{k+1}\hat{\eta}_k^{A,z} \right) \gamma_k^2 \\
 &\quad + L_z^2 \left(2C_{k+1}c_2\eta_k^C + C'_{k+1}(2\hat{\eta}_k^{C,z} + \hat{\eta}_k^{C,z}) \right) \gamma_k^2, \\
 \text{Part}_4 &= \alpha_k + A_{k+1}\theta_k^A - A_k + 2A'_{k+1}\hat{\eta}_k^{A,x}\alpha_k^2 + 2B'_{k+1}\hat{\eta}_k^{B,x}\alpha_k^2 + 2C'_{k+1}\hat{\eta}_k^{C,x}\alpha_k^2, \\
 \text{Part}_5 &= c_4\beta_k C_y + B_{k+1}\theta_k^B - B_k + 2 \left(A_{k+1}c_1\eta_k^A + A'_{k+1}\hat{\eta}_k^{A,y} \right) \beta_k^2 \\
 &\quad + 2 \left(B_{k+1}c_2\eta_k^B + B'_{k+1}\hat{\eta}_k^{B,y} \right) \beta_k^2 + 2 \left(C_{k+1}c_1\eta_k^C + C'_{k+1}\hat{\eta}_k^{C,y} \right) \beta_k^2, \\
 \text{Part}_6 &= c_{10}\gamma_k C_z + C_{k+1}\theta_k^C - C_k + 2 \left(A_{k+1}c_2\eta_k^A + A'_{k+1}\hat{\eta}_k^{A,z} \right) \gamma_k^2 \\
 &\quad + 2 \left(C_{k+1}c_2\eta_k^C + C'_{k+1}\hat{\eta}_k^{C,z} \right) \gamma_k^2, \\
 \text{Part}_7 &= A_{k+1}\tau_k^A + A'_{k+1}\lambda_k^A - A'_k, \\
 \text{Part}_8 &= B_{k+1}\tau_k^B + B'_{k+1}\lambda_k^B - B'_k, \\
 \text{Part}_9 &= C_{k+1}\tau_k^C + C'_{k+1}\lambda_k^C - C'_k.
 \end{aligned}$$

It remains to verify that, under the conditions stated in Theorem 14, $\text{Part}_i \leq 0$, $i = 1, 2, \dots, 9$. Granting this claim for the moment, we obtain

$$L_{k+1} - L_k \leq -\frac{\alpha_k}{2} \mathbb{E} \left[\|\nabla H(x_k)\|^2 \right] + A_{k+1}\Delta_k^A + B_{k+1}\Delta_k^B + C_{k+1}\Delta_k^C.$$

By rearranging and summing up, we obtain

$$\sum_{k=0}^{K-1} \alpha_k \mathbb{E} \left[\|\nabla H(x_k)\|^2 \right] \leq 2L_0 - 2L_K + 2 \sum_{k=0}^{K-1} (A_{k+1}\Delta_k^A + B_{k+1}\Delta_k^B + C_{k+1}\Delta_k^C),$$

which leads to

$$\inf_{k < K} \mathbb{E}[\|\nabla H(x_k)\|^2] \leq \frac{2(L_0 - H_*)}{\sum_{k=0}^{K-1} \alpha_k} + \frac{2}{\sum_{k=0}^{K-1} \alpha_k} \sum_{k=0}^{K-1} (A_{k+1}\Delta_k^A + B_{k+1}\Delta_k^B + C_{k+1}\Delta_k^C).$$

For completeness, we will now verify $\text{Part}_i \leq 0$ for $i = 1, 2, \dots, 9$ individually.

① First, using (13a), we have the following implication:

$$\left(B_{k+1}\eta_k^B + B'_{k+1}\hat{\eta}_k^{B,y} \right) \alpha_k \leq \left(B_{k+1}\eta_k^B + B'_{k+1}\hat{\eta}_k^{B,y} \right) m_{\alpha\beta}\beta_k \leq \frac{1}{64\tilde{c}_2}.$$

Similarly, we can derive that

$$\begin{aligned}
 \left(A_{k+1}\eta_k^A + A'_{k+1}\hat{\eta}_k^{A,x} \right) \alpha_k &\leq \frac{1}{96\tilde{c}_1}, \quad \left(A'_{k+1}\hat{\eta}_k^{A,x} + B'_{k+1}\hat{\eta}_k^{B,x} + C'_{k+1}\hat{\eta}_k^{C,x} \right) \alpha_k \leq \frac{1}{64}, \\
 \left(B_{k+1}\eta_k^B + B'_{k+1}\hat{\eta}_k^{B,x} \right) \alpha_k &\leq \frac{1}{96\tilde{c}_2}, \quad \left(C_{k+1}\eta_k^C + C'_{k+1}\hat{\eta}_k^{C,x} \right) \alpha_k \leq \frac{1}{96\tilde{c}_1}, \\
 \left(A_{k+1}\eta_k^A + A'_{k+1}\hat{\eta}_k^{A,z} \right) \gamma_k &\leq \min \left\{ \frac{1}{4\tilde{c}_2}, \frac{c_3^2}{72\tilde{c}_2 L_z^2}, \frac{\mu}{40\tilde{c}_2 c_{10} L_z^2} \right\},
 \end{aligned}$$

$$\begin{aligned} (C_{k+1}\eta_k^{\mathcal{C}} + C'_{k+1}\hat{\eta}_k^{\mathcal{C},z})\gamma_k &\leq \min\left\{\frac{1}{4\tilde{c}_2}, \frac{c_3^2}{72\tilde{c}_2L_z^2}, \frac{\mu}{40\tilde{c}_2c_{10}L_z^2}\right\}, \\ A'_{k+1}\hat{\eta}_k^{\mathcal{A},z}\gamma_k &\leq \min\left\{\frac{c_3^2}{36L_z^2}, \frac{\mu}{20c_{10}L_z^2}\right\}, \quad C'_{k+1}\hat{\eta}_k^{\mathcal{C},z}\gamma_k \leq \min\left\{\frac{c_3^2}{36L_z^2}, \frac{\mu}{20c_{10}L_z^2}\right\}. \end{aligned}$$

- ② For Part_{*i*} ($i = 7, 8, 9$), by applying the SE-CC Condition for $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, we have Part_{*i*} ≤ 0 for $i = 7, 8, 9$. Moreover, by setting $C_y = 1/c_4$, $C_z = 1/c_{10}$, we obtain

$$c_4\beta_k C_y + B_{k+1}\theta_k^{\mathcal{B}} - B_k \leq -\beta_k, \quad c_{10}\gamma_k C_z + C_{k+1}\theta_k^{\mathcal{C}} - C_k \leq -\gamma_k. \quad (29)$$

- ③ For Part₁, since

$$\begin{aligned} \alpha_k &\leq \min\left\{\frac{3}{8L_{y^*}^2}\beta_k, \frac{3}{16L_{z^*}^2}\gamma_k\right\}, \quad (A_{k+1}\eta_k^{\mathcal{A}} + A'_{k+1}\hat{\eta}_k^{\mathcal{A},x})\alpha_k \leq \frac{1}{96\tilde{c}_1}, \\ (A'_{k+1}\hat{\eta}_k^{\mathcal{A},x} + B'_{k+1}\hat{\eta}_k^{\mathcal{B},x} + C'_{k+1}\hat{\eta}_k^{\mathcal{C},x})\alpha_k &\leq \frac{1}{64}, \\ (B_{k+1}\eta_k^{\mathcal{B}} + B'_{k+1}\hat{\eta}_k^{\mathcal{B},x})\alpha_k &\leq \frac{1}{96\tilde{c}_2}, \quad (C_{k+1}\eta_k^{\mathcal{C}} + C'_{k+1}\hat{\eta}_k^{\mathcal{C},x})\alpha_k \leq \frac{1}{96\tilde{c}_1}, \end{aligned}$$

it follows that

$$\begin{aligned} \text{Part}_1 &\leq -\frac{\alpha_k}{16} + 2(A'_{k+1}\hat{\eta}_k^{\mathcal{A},x} + B'_{k+1}\hat{\eta}_k^{\mathcal{B},x} + C'_{k+1}\hat{\eta}_k^{\mathcal{C},x})\alpha_k^2 \\ &\quad + \tilde{c}_2(B_{k+1}\eta_k^{\mathcal{B}} + B'_{k+1}\hat{\eta}_k^{\mathcal{B},x})\alpha_k^2 + \tilde{c}_1(A_{k+1}\eta_k^{\mathcal{A}} + A'_{k+1}\hat{\eta}_k^{\mathcal{A},x} + C_{k+1}\eta_k^{\mathcal{C}} + C'_{k+1}\hat{\eta}_k^{\mathcal{C},x})\alpha_k^2 \\ &\leq -\frac{\alpha_k}{32} + \tilde{c}_2(B_{k+1}\eta_k^{\mathcal{B}} + B'_{k+1}\hat{\eta}_k^{\mathcal{B},x})\alpha_k^2 + \tilde{c}_1(C_{k+1}\eta_k^{\mathcal{C}} + C'_{k+1}\hat{\eta}_k^{\mathcal{C},x})\alpha_k^2 \\ &\quad + \tilde{c}_1(A_{k+1}\eta_k^{\mathcal{A}} + A'_{k+1}\hat{\eta}_k^{\mathcal{A},x})\alpha_k^2 \leq 0. \end{aligned}$$

- ④ For Part₂, under the parameter choices and bounds

$$\begin{aligned} C_y = \frac{1}{c_4} \left(\Rightarrow c_3 C_y = \frac{c_3^2}{3} \right), \quad \alpha_k &\leq \frac{c_3^2}{108c_1}\beta_k, \quad \gamma_k \leq \frac{c_3^2}{54c_1}\beta_k, \quad B'_{k+1}\hat{\eta}_k^{\mathcal{B},y}\beta_k \leq \frac{c_3^2}{36c_2}, \\ C'_{k+1}\hat{\eta}_k^{\mathcal{C},y}\beta_k &\leq \frac{c_3^2}{36c_2}, \quad C'_{k+1}\hat{\eta}_k^{\mathcal{C},z}\gamma_k \leq \frac{c_3^2}{36L_z^2}, \quad A'_{k+1}\hat{\eta}_k^{\mathcal{A},z}\gamma_k \leq \frac{c_3^2}{36L_z^2}, \\ A'_{k+1}\hat{\eta}_k^{\mathcal{A},y}\beta_k &\leq \frac{c_3^2}{36c_2}, \quad (A_{k+1}\eta_k^{\mathcal{A}} + A'_{k+1}\hat{\eta}_k^{\mathcal{A},y})\beta_k \leq \frac{c_3^2}{72\tilde{c}_1c_2}, \\ (B_{k+1}\eta_k^{\mathcal{B}} + B'_{k+1}\hat{\eta}_k^{\mathcal{B},y})\beta_k &\leq \frac{c_3^2}{72\tilde{c}_2^2}, \quad (C_{k+1}\eta_k^{\mathcal{C}} + C'_{k+1}\hat{\eta}_k^{\mathcal{C},y})\beta_k \leq \frac{c_3^2}{72\tilde{c}_1c_2}, \\ (C_{k+1}\eta_k^{\mathcal{C}} + C'_{k+1}\hat{\eta}_k^{\mathcal{C},z})\gamma_k &\leq \frac{c_3^2}{72\tilde{c}_2L_z^2}, \quad (A_{k+1}\eta_k^{\mathcal{A}} + A'_{k+1}\hat{\eta}_k^{\mathcal{A},z})\gamma_k \leq \frac{c_3^2}{72\tilde{c}_2L_z^2}. \end{aligned}$$

Therefore, the sum of all nonnegative terms in Part₂ is at most $c_3C_y\beta_k$, and hence Part₂ ≤ 0 .

- ⑤ For Part₃, under the parameter choices and bounds

$$\begin{aligned}\alpha_k &\leq \frac{\mu}{60c_2c_{10}}\gamma_k, \quad A'_{k+1}\hat{\eta}_k'^{A,z}\gamma_k \leq \frac{\mu}{20c_{10}L_z^2}, \quad C'_{k+1}\hat{\eta}_k'^{C,z}\gamma_k \leq \frac{\mu}{20c_{10}L_z^2}, \\ \left(A_{k+1}\eta_k^A + A'_{k+1}\hat{\eta}_k^{A,z}\right)\gamma_k &\leq \frac{\mu}{40\tilde{c}_2c_{10}L_z^2}, \quad \left(C_{k+1}\eta_k^C + C'_{k+1}\hat{\eta}_k^{C,z}\right)\gamma_k \leq \frac{\mu}{40\tilde{c}_2c_{10}L_z^2}.\end{aligned}$$

By substituting the above bounds into the definition of Part_3 and summing the resulting termwise estimates, we conclude that the total contribution of all nonnegative terms is dominated by the negative term $-(C_z\mu/4)\gamma_k$. Hence, $\text{Part}_3 \leq 0$.

⑥ For Part_4 : by the SE-CC condition and $(A'_{k+1}\hat{\eta}_k'^{A,x} + B'_{k+1}\hat{\eta}_k'^{B,x} + C'_{k+1}\hat{\eta}_k'^{C,x})\alpha_k \leq 1/2$, we have $\text{Part}_4 \leq 0$.

⑦ For Part_5 : By combining (29) with the following conditions

$$\begin{aligned}\left(A_{k+1}\eta_k^A + A'_{k+1}\hat{\eta}_k^{A,y}\right)\beta_k &\leq 1/(6\tilde{c}_1), \quad \left(B_{k+1}\eta_k^B + B'_{k+1}\hat{\eta}_k^{B,y}\right)\beta_k \leq 1/(6\tilde{c}_2), \\ \left(C_{k+1}\eta_k^C + C'_{k+1}\hat{\eta}_k^{C,y}\right)\beta_k &\leq 1/(6\tilde{c}_1).\end{aligned}$$

Plugging these bounds into Part_5 shows that the sum of all nonnegative terms is at most $\beta_k/2$. Therefore, $\text{Part}_5 \leq 0$.

⑧ For Part_6 : Similarly, the validity of inequality $\text{Part}_6 \leq 0$ relies on inequality (29) and

$$\left(A_{k+1}\eta_k^A + A'_{k+1}\hat{\eta}_k^{A,z}\right)\gamma_k \leq 1/(4\tilde{c}_2), \quad \left(C_{k+1}\eta_k^C + C'_{k+1}\hat{\eta}_k^{C,z}\right)\gamma_k \leq 1/(4\tilde{c}_2).$$

Thus, the proof of the theorem is completed. ■

Appendix G. Proof of Theorem 15

Proof Let $C_y = C_z = 1/2$. Then, applying Lemmas 4, 6, 8, 11 and combining with (10)-(11) and (19)-(22), we obtain

$$\begin{aligned}L_{k+1} - L_k &\leq \text{Part}'_0 \mathbb{E} \left[\|\nabla H(x_k)\|^2 \right] + \text{Part}'_1 \mathbb{E} \left[\|\nabla H(x_k) - v_k^x\|^2 \right] + \text{Part}'_2 \mathbb{E} \left[\|v_k^x\|^2 \right] \\ &\quad + \text{Part}'_3 \mathbb{E} \left[\|y_k - y_k^*\|^2 \right] + \text{Part}'_4 \mathbb{E} \left[\|z_k - z_k^*\|^2 \right] + \text{Part}'_5 \mathbb{E} \left[\|\hat{v}_k^x - D_k^x\|^2 \right] \\ &\quad + \text{Part}'_6 \mathbb{E} \left[\|D_k^y - v_k^y\|^2 \right] + \text{Part}'_7 \mathbb{E} \left[\|D_k^z - v_k^z\|^2 \right] + \text{Part}'_8 \delta_k^A + \text{Part}'_9 \delta_k^B \\ &\quad + \text{Part}'_{10} \delta_k^C + A''_{k+1} \rho_k^2 \Delta_k^A + A_{k+1} \Delta_k^A + B_{k+1} \Delta_k^B + C_{k+1} \Delta_k^C,\end{aligned}$$

where

$$\begin{aligned}\text{Part}'_0 &= -\alpha_k/2 + 2(A'_{k+1}\hat{\eta}_k'^{A,x} + B'_{k+1}\hat{\eta}_k'^{B,x} + C'_{k+1}\hat{\eta}_k'^{C,x})\alpha_k^2, \\ \text{Part}'_1 &= \alpha_k/2 + A''_{k+1}(1 - \rho_k) - A''_k, \\ \text{Part}'_2 &= -\frac{\alpha_k}{4} + \frac{c_6\alpha_k^2}{2\beta_k} + \frac{c_9\alpha_k^2}{2\gamma_k} + \frac{3\alpha_k^2(L^H)^2}{\rho_k}A''_{k+1} + 12c_1\rho_k\alpha_k^2A''_{k+1} + c_1\rho_k^2A''_{k+1}\eta_k^A\alpha_k^2 \\ &\quad + \left(\eta_k^A A_{k+1}c_1 + \hat{\eta}_k^{A,x}A'_{k+1} + \eta_k^C C_{k+1}c_1 + \hat{\eta}_k^{C,x}C'_{k+1} + \eta_k^B B_{k+1}c_2 + \hat{\eta}_k^{B,x}B'_{k+1}\right)\alpha_k^2,\end{aligned}$$

$$\begin{aligned}
 \text{Part}'_3 &= -\mu\beta_k/2 + c_8\gamma_k/2 + c_2 \left(2B_{k+1}\eta_k^{\mathcal{B}}c_2 + B'_{k+1}(2\hat{\eta}_k^{\mathcal{B},y} + \hat{\eta}'^{\mathcal{B},y}_k) \right) \beta_k^2 \\
 &\quad + c_2 \left(2A_{k+1}c_1\eta_k^{\mathcal{A}} + A'_{k+1}(2\hat{\eta}_k^{\mathcal{A},y} + \hat{\eta}'^{\mathcal{A},y}_k) + 2C_{k+1}\eta_k^{\mathcal{C}}c_1 + C'_{k+1}(2\hat{\eta}_k^{\mathcal{C},y} + \hat{\eta}'^{\mathcal{C},y}_k) \right) \beta_k^2 \\
 &\quad + L_z^2 \left(2\eta_k^{\mathcal{C}}C_{k+1}c_2 + C'_{k+1}(2\hat{\eta}_k^{\mathcal{C},z} + \hat{\eta}'^{\mathcal{C},z}_k) + 2\eta_k^{\mathcal{A}}A_{k+1}c_2 + A'_{k+1}(2\hat{\eta}_k^{\mathcal{A},z} + \hat{\eta}'^{\mathcal{A},z}_k) \right) \gamma_k^2 \\
 &\quad + 24c_1c_2\rho_k\beta_k^2A''_{k+1} + 2c_1c_2\rho_k^2A''_{k+1}\eta_k^{\mathcal{A}}\beta_k^2 + 2c_2L_z^2\rho_k^2A''_{k+1}\eta_k^{\mathcal{A}}\gamma_k^2 + 24c_2L_z^2\rho_k\gamma_k^2A''_{k+1} \\
 &\quad + 9c_1\rho_kA''_{k+1} + 6c_1 \left(\hat{\eta}_k^{\mathcal{A},x}A'_{k+1} + \hat{\eta}_k^{\mathcal{B},x}B'_{k+1} + \hat{\eta}_k^{\mathcal{C},x}C'_{k+1} \right) \alpha_k^2, \\
 \text{Part}'_4 &= 9c_2\rho_kA''_{k+1} - \mu\gamma_k/2 + 24c_2L_z^2\rho_k\gamma_k^2A''_{k+1} + 2c_2L_z^2\rho_k^2A''_{k+1}\eta_k^{\mathcal{A}}\gamma_k^2 \\
 &\quad + L_z^2 \left(2c_2\eta_k^{\mathcal{A}}A_{k+1} + A'_{k+1}(2\hat{\eta}_k^{\mathcal{A},z} + \hat{\eta}'^{\mathcal{A},z}_k) + 2c_2\eta_k^{\mathcal{C}}C_{k+1} + C'_{k+1}(2\hat{\eta}_k^{\mathcal{C},z} + \hat{\eta}'^{\mathcal{C},z}_k) \right) \gamma_k^2 \\
 &\quad + 6c_2 \left(\hat{\eta}_k^{\mathcal{A},x}A'_{k+1} + \hat{\eta}_k^{\mathcal{B},x}B'_{k+1} + \hat{\eta}_k^{\mathcal{C},x}C'_{k+1} \right) \alpha_k^2, \\
 \text{Part}'_5 &= \rho_k^2A''_{k+1}\theta_k^{\mathcal{A}} + A_{k+1}\theta_k^{\mathcal{A}} - A_k, \\
 \text{Part}'_6 &= \beta_k^2 + \theta_k^{\mathcal{B}}B_{k+1} - B_k + 24c_1\rho_k\beta_k^2A''_{k+1} + 2c_1\rho_k^2A''_{k+1}\eta_k^{\mathcal{A}}\beta_k^2 \\
 &\quad + 2 \left(\eta_k^{\mathcal{B}}B_{k+1}c_2 + \hat{\eta}_k^{\mathcal{B},y}B'_{k+1} + \eta_k^{\mathcal{A}}A_{k+1}c_1 + \hat{\eta}_k^{\mathcal{A},y}A'_{k+1} + \eta_k^{\mathcal{C}}C_{k+1}c_1 + \hat{\eta}_k^{\mathcal{C},y}C'_{k+1} \right) \beta_k^2, \\
 \text{Part}'_7 &= \gamma_k^2 + \theta_k^{\mathcal{C}}C_{k+1} - C_k + 2c_2\rho_k^2A''_{k+1}\eta_k^{\mathcal{A}}\gamma_k^2 + 24c_2\rho_k\gamma_k^2A''_{k+1} \\
 &\quad + 2 \left(A_{k+1}\eta_k^{\mathcal{A}}c_2 + \hat{\eta}_k^{\mathcal{A},z}A'_{k+1} + \eta_k^{\mathcal{C}}C_{k+1}c_2 + \hat{\eta}_k^{\mathcal{C},z}C'_{k+1} \right) \gamma_k^2, \\
 \text{Part}'_8 &= \rho_k^2A''_{k+1}\tau_k^{\mathcal{A}} + \tau_k^{\mathcal{A}}A_{k+1} + \lambda_k^{\mathcal{A}}A'_{k+1} - A'_k, \\
 \text{Part}'_9 &= \tau_k^{\mathcal{B}}B_{k+1} + \lambda_k^{\mathcal{B}}B'_{k+1} - B'_k, \\
 \text{Part}'_{10} &= \tau_k^{\mathcal{C}}C_{k+1} + \lambda_k^{\mathcal{C}}C'_{k+1} - C'_k.
 \end{aligned}$$

Following a similar approach to the proof of Theorem 14, we have $\text{Part}'_0 \leq -\frac{\alpha_k}{4}$ and $\text{Part}'_i \leq 0$ for $i = 1, 2, \dots, 10$. Specifically, the UE-CC and UEMA-CC conditions ensure that $\text{Part}'_i \leq 0$ holds for $i = 1, 5, 8, 9, 10$. For the remaining i , following a similar process to ① in the proof of Theorem 14, we can derive the following conditions:

$$\begin{aligned}
 \hat{\eta}_k^{\mathcal{A},x}A'_{k+1}\alpha_k &\leq \frac{1}{24}, \quad \hat{\eta}_k^{\mathcal{B},x}B'_{k+1}\alpha_k \leq \frac{1}{24}, \quad \hat{\eta}_k^{\mathcal{C},x}C'_{k+1}\alpha_k \leq \frac{1}{24}, \quad \hat{\eta}_k^{\mathcal{B},y}B'_{k+1}\beta_k \leq \frac{\mu}{16c_2}, \\
 \left(\eta_k^{\mathcal{A}}A_{k+1} + \hat{\eta}_k^{\mathcal{A},x}A'_{k+1} \right) \alpha_k &\leq \frac{1}{32\tilde{c}_1}, \quad \left(\eta_k^{\mathcal{A}}A_{k+1} + \hat{\eta}_k^{\mathcal{A},y}A'_{k+1} \right) \leq \frac{1}{10\tilde{c}_2}, \\
 \left(\eta_k^{\mathcal{A}}A_{k+1} + \hat{\eta}_k^{\mathcal{A},z}A'_{k+1} \right) &\leq \frac{1}{8\tilde{c}_2}, \quad \hat{\eta}_k^{\mathcal{A},y}A'_{k+1}\beta_k \leq \frac{\mu}{16c_2}, \quad \hat{\eta}_k^{\mathcal{A},z}A'_{k+1}\gamma_k \leq \frac{\mu}{10L_z^2}, \\
 \left(\eta_k^{\mathcal{B}}B_{k+1} + \hat{\eta}_k^{\mathcal{B},x}B'_{k+1} \right) \alpha_k &\leq \frac{1}{32\tilde{c}_2}, \quad \left(\eta_k^{\mathcal{B}}B_{k+1} + \hat{\eta}_k^{\mathcal{B},y}B'_{k+1} \right) \leq \frac{1}{10\tilde{c}_2}, \\
 \left(\eta_k^{\mathcal{C}}C_{k+1} + \hat{\eta}_k^{\mathcal{C},x}C'_{k+1} \right) \alpha_k &\leq \frac{1}{32\tilde{c}_1}, \quad \left(\eta_k^{\mathcal{C}}C_{k+1} + \hat{\eta}_k^{\mathcal{C},y}C'_{k+1} \right) \leq \frac{1}{10\tilde{c}_2}, \\
 \left(\eta_k^{\mathcal{C}}C_{k+1} + \hat{\eta}_k^{\mathcal{C},z}C'_{k+1} \right) &\leq \frac{1}{8\tilde{c}_2}, \quad \hat{\eta}_k^{\mathcal{C},y}C'_{k+1}\beta_k \leq \frac{\mu}{16c_2}, \quad \hat{\eta}_k^{\mathcal{C},z}C'_{k+1}\gamma_k \leq \frac{\mu}{10L_z^2}, \\
 \rho_k\alpha_kA''_{k+1} &\leq \frac{1}{384c_1}, \quad \rho_k^2\alpha_kA''_{k+1}\eta_k^{\mathcal{A}} \leq \frac{1}{32c_1}, \quad \rho_k^2A''_{k+1}\eta_k^{\mathcal{A}} \leq \min \left\{ \frac{1}{10c_1}, \frac{1}{8c_2} \right\},
 \end{aligned}$$

which guarantee that $\text{Part}'_i \leq 0$ holds for $i = 2, 3, 4, 6, 7$ and $\text{Part}'_0 \leq -\frac{\alpha_k}{4}$. Similar to the proof of Theorem 14, we obtain the final conclusion. $\blacksquare$

References

- Michael Arbel and Julien Mairal. Amortized implicit differentiation for stochastic bilevel optimization. In *International Conference on Learning Representations*, 2022.
- 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):165–214, 2023.
- Abhiroop Bhattacharjee, Abhishek Moitra, and Priyadarshini Panda. Clipformer: Key-value clipping of transformers on memristive crossbars for write noise mitigation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 44(2):592–601, 2025.
- Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *International Conference on Computational Statistics*, pages 177–186, 2010.
- Lesi Chen, Yaohua Ma, and Jingzhao Zhang. Near-optimal nonconvex-strongly-convex bilevel optimization with fully first-order oracles. *Journal of Machine Learning Research*, 26(109):1–56, 2025.
- Xuxing Chen, Tesi Xiao, and Krishnakumar Balasubramanian. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *Journal of Machine Learning Research*, 25(151):1–51, 2024.
- Tianshu Chu, Dachuan Xu, Wei Yao, and Jin Zhang. SPABA: A single-loop and probabilistic stochastic bilevel algorithm achieving optimal sample complexity. In *International Conference on Machine Learning*, volume 235, pages 8848–8903. PMLR, 2024.
- Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, volume 32, pages 15236 – 15245, 2019.
- Mathieu Dagr  ou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *Advances in Neural Information Processing Systems*, volume 35, pages 26698–26710, 2022.
- Mathieu Dagr  ou, Thomas Moreau, Samuel Vaiter, and Pierre Ablin. A lower bound and a near-optimal algorithm for bilevel empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics*, volume 238, pages 82–90. PMLR, 2024.
- Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, volume 27, pages 1646 – 1654, 2014.

- Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31, pages 687 – 697, 2018.
- Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, volume 80, pages 1568–1577. PMLR, 2018.
- Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.
- Bin Gu, Guodong Liu, Yanfu Zhang, Xiang Geng, and Heng Huang. Optimizing large-scale hyperparameters via automated learning algorithm. *arXiv preprint arXiv:2102.09026*, 2021.
- Jie Hao, Kaiyi Ji, and Mingrui Liu. Bilevel coreset selection in continual learning: A new formulation and algorithm. In *Advances in Neural Information Processing Systems*, volume 36, pages 51026–51049, 2023.
- 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, Zi-Hao Qiu, Zhishuai Guo, Lijun Zhang, and Tianbao Yang. Blockwise stochastic variance-reduced methods with parallel speedup for multi-block bilevel optimization. In *International Conference on Machine Learning*, volume 202, pages 13550–13583. PMLR, 2023.
- Feihu Huang. On momentum-based gradient methods for bilevel optimization with nonconvex lower-level. *arXiv preprint arXiv:2303.03944*, 2023.
- Kaiyi Ji, Jason D Lee, Yingbin Liang, and H. Vincent Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. In *Advances in Neural Information Processing Systems*, volume 33, pages 11490–11500, 2020.
- Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, volume 139, pages 4882–4892. PMLR, 2021.
- Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26, pages 315 – 323, 2013.
- 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. In *Advances in Neural Information Processing Systems*, volume 34, pages 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*, volume 202, pages 18083–18113. PMLR, 2023.
- Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *International Conference on Learning Representations*, 2024.
- Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.
- Junyi Li and Heng Huang. Provably faster algorithms for bilevel optimization via without-replacement sampling. In *Advances in Neural Information Processing Systems*, volume 37, pages 70520–70556, 2024.
- 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.
- Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, volume 139, pages 6286–6295. PMLR, 2021a.
- Zhize Li, Slavomír Hanzely, and Peter Richtárik. ZeroSARAH: Efficient nonconvex finite-sum optimization with zero full gradient computation. *arXiv preprint arXiv:2103.01447*, 2021b.
- Renjie Liao, Yuwen Xiong, Ethan Fetaya, Lisa Zhang, KiJung Yoon, Xaq Pitkow, Raquel Urtasun, and Richard Zemel. Reviving and improving recurrent back-propagation. In *International Conference on Machine Learning*, volume 80, pages 3082–3091. PMLR, 2018.
- Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pages 1540–1552. PMLR, 2020.
- Kazusato Oko, Shunta Akiyama, Denny Wu, Tomoya Murata, and Taiji Suzuki. SILVER: Single-loop variance reduction and application to federated learning. In *International Conference on Machine Learning*, volume 235, pages 38683–38739. PMLR, 2024.
- Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning*, volume 48, pages 737–746. PMLR, 2016.
- Nhan H. Pham, Lam M. Nguyen, Dzung T. Phan, and Quoc Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.
- Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, volume 25, pages 2663 – 2671, 2012.

- Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems*, volume 29, pages 46 – 54, 2016.
- Han Shen, Zhuoran Yang, and Tianyi Chen. Principled penalty-based methods for bilevel reinforcement learning and RLHF. In *International Conference on Machine Learning*, volume 235, pages 44774–44799. PMLR, 2024.
- Han Shen, Quan Xiao, and Tianyi Chen. On penalty-based bilevel gradient descent method. *Mathematical Programming*, 199:1–51, 2025.
- Naresh K. Sinha and Michael P. Griscik. A stochastic approximation method. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-1(4):338–344, 1971.
- Daouda Sow, Kaiyi Ji, and Yingbin Liang. On the convergence theory for hessian-free bilevel algorithms. In *Advances in Neural Information Processing Systems*, volume 35, pages 4136–4149, 2022.
- Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, volume 32, pages 2406 – 2416, 2019.
- Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 34, pages 13670–13682, 2021.
- Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Achieving $O(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 36, pages 39491–39503, 2023.
- Wei Yao, Haian Yin, Shangzhi Zeng, and Jin Zhang. Overcoming lower-level constraints in bilevel optimization: A novel approach with regularized gap functions. In *International Conference on Learning Representations*, 2025.
- Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.
- Dongruo Zhou and Quanquan Gu. Lower bounds for smooth nonconvex finite-sum optimization. In *International Conference on Machine Learning*, volume 97, pages 7574–7583. PMLR, 2019.
- Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for non-convex optimization. *Journal of Machine Learning Research*, 21(103):1–63, 2020.