A SINGLE-LOOP, CONSTANT-BATCH FIRST-ORDER PENALTY METHOD FOR STOCHASTIC BILEVEL OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in penalty-based methods for stochastic bilevel optimization (SBO) have eliminated the need for second-order derivative oracles. However, for stochastic nonconvex–strongly convex bilevel problems, existing first-order methods typically rely on nested loops and/or large batch sizes for attaining $O(\epsilon^{-6})$ or $O(\epsilon^{-4})$ sample complexity under standard bounded-variance assumption or mean-square smoothness assumption. Achieving these rates with a single-loop penalty method and a constant batch size remains challenging due to a large penalty value needed for an accurate approximation. To address this challenge, we develop a stochastic SIngle-loop COnstant-Batch first-order penalty method (SICO) that combines two complementary ingredients. First, it performs one stochastic-gradient update per-iteration for both the original lower-level and penalized problems, with a projection that controls the separation between their iterates. Second, it applies an exponential moving average to stabilize the upper-level gradient estimator. We show that this combination achieves $O(\epsilon^{-6})$ sample complexity using only $O(1)$ stochastic-gradient samples per iteration under unbiased, bounded-variance stochastic gradients. Under the additional mean-square smoothness assumption on the lower-level stochastic gradients, the same algorithm improves the complexity to $O(\epsilon^{-4})$ also with $O(1)$ batch size. To the best of our knowledge, this is the first work to match the best-known convergence rate for fully first-order SBO methods using a single loop and a constant batch size. This result addresses an open problem posed in the literature.

1 INTRODUCTION

In this paper we study *nonconvex–strongly convex (NC–SC) stochastic bilevel optimization (SBO)*:

$$\min_{\mathbf{w}\in\mathbb{R}^{d_w}} F(\mathbf{w}) := f(\mathbf{w}, \mathbf{y}^*(\mathbf{w})), \qquad \mathbf{y}^*(\mathbf{w}) := \arg\min_{\mathbf{y}\in\mathbb{R}^{d_y}} g(\mathbf{w}, \mathbf{y}), \tag{1}$$

where the lower level (LL) function $g(\mathbf{w}, \mathbf{y}) := \mathbf{E}_{\xi_g\sim\mathcal{D}_g} g(\mathbf{w}, \mathbf{y}, \xi_g)$ is jointly smooth and strongly convex in $\mathbf{y}$ for every $\mathbf{w}$, while the upper level (UL) function $f(\mathbf{w}, \mathbf{y}) := \mathbf{E}_{\xi_f\sim\mathcal{D}_f} f(\mathbf{w}, \mathbf{y}, \xi_f)$ is smooth but potentially nonconvex. Both f and g are accessed through stochastic gradients. This formulation captures learning problems in which one optimization task depends on the solution of another, including hyperparameter optimization (Franceschi et al., 2018; Bao et al., 2021), data hyper-cleaning and reweighting (Gao et al., 2022), and meta-learning with strongly convex inner problems (Ji et al., 2021; Kwon et al., 2023).

A substantial body of work has developed algorithms and convergence guarantees for this problem class. Representative contributions include early approximation methods (Ghadimi & Wang, 2018), two-timescale stochastic approximation (Hong et al., 2020), improved hypergradient estimation and alternating stochastic updates (Ji et al., 2021; Chen et al., 2021), momentum-based single-loop methods (Guo et al., 2021b), and variance-reduced algorithms (Yang et al., 2021; Khanduri et al., 2021; Guo et al., 2021a). These approaches estimate the *hypergradient* $\nabla F(\mathbf{w})$ by differentiating through the lower-level solution map. This typically requires second-order information, accessed through Hessian- or Jacobian-vector products (HVPs/JVPs) or approximate inverse-Hessian computations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To reduce the computational and memory costs associated with second-order derivative computations, recent work has developed *fully first-order* methods, which use only ordinary stochastic gradients of f and g and avoid second-order oracle access altogether. A prominent approach is to construct a penalty-based reformulation of the bilevel problem that can be optimized using only first-order information (Liu et al., 2022; Kwon et al., 2023; Shen & Chen, 2023). In particular, problem (1) admits the equivalent constraint reformulation:

$$\min_{\mathbf{w},\mathbf{y}} f(\mathbf{w},\mathbf{y}) \qquad \text{subject to} \qquad g(\mathbf{w},\mathbf{y}) - \min_{\mathbf{z}} g(\mathbf{w},\mathbf{z}) \leq 0. \tag{2}$$

Penalizing the constraint violation yields a surrogate problem:

$$\min_{\mathbf{w}} \mathcal{L}_\lambda^*(\mathbf{w}), \qquad \mathcal{L}_\lambda^*(\mathbf{w}) := \min_{\mathbf{y}} \left\{ f(\mathbf{w},\mathbf{y}) + \lambda \left[g(\mathbf{w},\mathbf{y}) - \min_{\mathbf{z}} g(\mathbf{w},\mathbf{z}) \right] \right\}, \quad \lambda > 0. \tag{3}$$

Under standard regularity conditions and for sufficiently large λ, the gradient of the surrogate objective approximates the bilevel hypergradient with error $\|\nabla \mathcal{L}_\lambda^*(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 = O(\lambda^{-1})$ (Kwon et al., 2023). Consequently, choosing $\lambda = \Theta(\epsilon^{-1})$ ensures that an $O(\epsilon)$-stationary point of $\mathcal{L}_\lambda^*$ is also an $O(\epsilon)$-stationary point of F.

Despite this first-order representation, attaining sharp sample complexity with simple updates remains challenging. Existing fully first-order methods achieve $O(\epsilon^{-6})$ sample complexity under bounded-variance stochastic gradients and $O(\epsilon^{-4})$ under additional mean-square smoothness. However, the methods attaining these rates rely on nested lower-level solvers (Kwon et al., 2024; Chen et al., 2025; Shen & Chen, 2023), ϵ-dependent large batch sizes (Kwon et al., 2024; Chen et al., 2025; Gu et al., 2026; Shen & Chen, 2023), or both. Meanwhile, existing methods that combine a single loop with constant-size batches have weaker sample complexity guarantees (Kwon et al., 2023; Yang, 2026). This gap motivates the following open question (Wright, 2025; Yang, 2026):

Can we design a fully first-order method for NC-SC SBO that uses a single loop and constant-size batches while matching the best-known sample complexity under standard stochastic assumptions?

We answer this question affirmatively with a simple algorithmic design supported by refined convergence analyses. Our method maintains $\mathbf{y}_t$ and $\mathbf{z}_t$ to track the minimizers of the penalized problem over $\mathbf{y}$ and original lower-level problem over $\mathbf{z}$ in (3), respectively, taking one stochastic-gradient step for each at every iteration given $\mathbf{w}_t$. Inspired by the projected-coupling idea of Kwon et al. (2024), we project the provisional $\mathbf{z}$-iterate onto a ball of radius $O(\lambda^{-1})$ centered at the updated $\mathbf{y}$-iterate. This projection constrains the separation between the two iterates rather than their individual locations. We combine it with a standard exponential moving average of the stochastic gradient estimator for the upper-level variable $\mathbf{w}$. The projection controls the relative lower-level tracking error, while the moving average controls the upper-level gradient estimation error with a constant batch size. We refer to the proposed algorithm as SICO.

By establishing refined contraction and sensitivity bounds for coupled lower-level tracking, we prove that SICO achieves $O(\epsilon^{-6})$ sample complexity under unbiased, bounded-variance stochastic gradients and $O(\epsilon^{-4})$ under the additional mean-square smoothness assumption on the lower-level stochastic gradients, using only $O(1)$ stochastic-gradient samples per iteration in both settings. In addition, our results also achieve the best known dependence on the condition number (cf. Table 1).

We evaluate the proposed method on bilevel problems arising in data hyper-cleaning and regularization selection, demonstrating its effectiveness compared with existing fully first-order baselines.

2 Related Work

Penalty methods for bilevel optimization build on classical value-function reformulations and associated exact-penalty analyses (Ye & Zhu, 1995). These formulations express lower-level optimality through a constraint on the lower-level optimality gap. More recent developments include (Liu et al., 2022; Shen & Chen, 2023; Lu & Mei, 2024). Building on this broader literature, recent research on NC–SC SBO has focused on controlling the approximation error and stochastic noise introduced by penalization.

Kwon et al. (2023) introduced F2SA, a fully first-order penalty method that avoids Jacobian and Hessian evaluations. It uses simple stochastic gradient updates for all variables. Its established sample complexity is $O(\epsilon^{-7})$ in the fully stochastic setting. Their F3SA variant incorporates STORM-

Table 1: Comparison of fully first-order methods for bilevel optimization. The second column indicates the mean-square smoothness (MSS) of the stochastic gradients (cf. Assumption 3).

Algorithm	MSS	Loop	Batch size	Sample complexity
F2SA (Kwon et al., 2023)	No	Single	$O(1)$	$O(\mathrm{poly}(\kappa)\epsilon^{-7})$
Yang (2026)	No	Single	$O(1)$	$O(\kappa^{15}\epsilon^{-7})$
Kwon et al. (2024)	No	Double	$O(\epsilon^{-4})$	$O(\mathrm{poly}(\kappa)\epsilon^{-6})$
F2BSA (Chen et al., 2025)	No	Double	$O(\epsilon^{-4})$	$\widetilde{O}(\kappa^{12}\epsilon^{-6})$
F3BSA (Chen et al., 2025)	No	Double	$O(\kappa)$	$\widetilde{O}(\kappa^{11}\epsilon^{-6})$
Stoc-SGHA (Gu et al., 2026)	No	Single	$O(\epsilon^{-4})$	$O(\kappa^{17}\epsilon^{-6})$
SICO (Theorem 1)	**No**	**Single**	$\boldsymbol{O(1)}$	$\boldsymbol{O(\kappa^{11}\epsilon^{-6})}$
F3SA (Kwon et al., 2023)	Yes	Single	$O(1)$	$O(\mathrm{poly}(\kappa)\epsilon^{-5})$
Kwon et al. (2024)	Yes	Double	$O(\epsilon^{-2})$	$O(\mathrm{poly}(\kappa)\epsilon^{-4})$
Stoc-SGHA (Gu et al., 2026)	Yes	Single	$O(\epsilon^{-2})$	$O(\kappa^{11}\epsilon^{-4})$
SICO (Theorem 2)	**Yes**	**Single**	$\boldsymbol{O(1)}$	$\boldsymbol{O(\kappa^{9}\epsilon^{-4})}$

style gradient estimators for both the upper- and lower-level variables, improving the complexity to $O(\epsilon^{-5})$ under additional mean-square smoothness for both the upper and lower level stochastic gradients while using a single loop and constant-size batches.

Kwon et al. (2024) established two improved complexity bounds: $O(\epsilon^{-6})$ under bounded-variance stochastic gradients and $O(\epsilon^{-4})$ under additional mean-square smoothness. Their method uses a double-loop structure with batch sizes of order ϵ^{-4} and ϵ^{-2} in the two settings, respectively. The analysis exploits the proximity of the original and penalized lower-level solutions through projection and, under mean-square smoothness. These ideas motivate our algorithm, while our refined tracking analysis enables single-loop updates with constant-size batches.

Chen et al. (2025) showed that the smoothness constant of the reduced penalty objective $\mathcal{L}_\lambda^*$ can be bounded independently of λ. Their F2BSA method combines inner solvers with large batch sizes to attain $\widetilde{O}(\kappa^{12}\epsilon^{-6})$ sample complexity in the fully stochastic setting. Their F3BSA variant uses multilevel Monte Carlo and single-sample stochastic updates to attain $\widetilde{O}(\kappa^{11}\epsilon^{-6})$ complexity. However, constructing its gradient estimator still requires nested stochastic solvers and a batch size in the order of $O(\kappa)$.

Using an equivalent min–max reformulation of the penalty problem (3), Yang (2026) developed a single-loop stochastic momentum gradient descent–ascent method with constant-size batches. Its established sample complexity under bounded variance is $O(\epsilon^{-7})$, leaving a gap relative to the $O(\epsilon^{-6})$ dependence achieved by nested-loop methods.

In concurrent work, Gu et al. (2026) proposed SGHA and its stochastic counterpart, Stoc-SGHA, based on a regularized Lagrangian formulation of the lower-level stationarity constraints. These methods approximate Hessian–vector products using finite differences of first-order gradients. Stoc-SGHA achieves expected oracle complexities of $O(\kappa^{17}\epsilon^{-6})$ under bounded variance and $O(\kappa^{11}\epsilon^{-4})$ under lower-level mean-square smoothness. However, its stated in-expectation guarantees additionally assume that the iterates are almost surely bounded and use batch sizes of order ϵ^{-4} and ϵ^{-2}, respectively. Thus, its single-loop structure still entails an ϵ-dependent per-iteration sampling cost. In contrast, SICO operates directly on stochastic gradients and does not require finite-difference approximations or an additional almost-sure bounded-iterate assumption and has sharper dependence on the condition number. Table 1 summarizes these comparisons.

3 Preliminaries

Notation. We use $\nabla_1 f(\mathbf{w},\mathbf{y})$ and $\nabla_2 f(\mathbf{w},\mathbf{y})$ to denote gradients with respect to the first and second arguments, respectively, and $\|\cdot\|$ to denote the Euclidean norm. Let $\Pi_{\mathbb{B}(\mathbf{c},r)}(\mathbf{u})$ denote the Euclidean projection of $\mathbf{u}$ onto a closed Euclidean ball $\mathbb{B}(\mathbf{c},r)=\{\mathbf{u}:\|\mathbf{u}-\mathbf{c}\|\le r\}$. We write $\mathbf{E}_t[\cdot]$ for the conditional expectation given the history prior to drawing the random samples at iteration t. A possibly random point $\mathbf{w}$ is called ϵ-stationary if $\mathbf{E}[\|\nabla F(\mathbf{w})\|^2]\le\epsilon^2$. A function f is C-Lipschitz continuous if $|f(\mathbf{x})-f(\mathbf{y})|\le C\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x},\mathbf{y}$. It is L-smooth if it is differentiable and $\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|\le L\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x},\mathbf{y}$. A differentiable function g is μ_g-strongly convex

if $g(\mathbf{y}) \geq g(\mathbf{x}) + \nabla g(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu_g}{2}\|\mathbf{y}-\mathbf{x}\|^2$ for all $\mathbf{x}, \mathbf{y}$. A mapping $H(\mathbf{s}) : \mathbb{R}^d \to \mathbb{R}^d$ is μ-strongly monotone if $(H(\mathbf{s}) - H(\mathbf{s}'))^\top(\mathbf{s}-\mathbf{s}') \geq \mu\|\mathbf{s}-\mathbf{s}'\|^2$.

3.1 Assumption

We make the following standard assumptions regarding the problem (1).

Assumption 1. *Regarding f, g and F, we assume:*
- *$f(\mathbf{w}, \mathbf{y})$ is C_f-Lipschitz in $\mathbf{y}$ and jointly L_f-smooth in terms of $(\mathbf{w}, \mathbf{y})$*
- *$g(\mathbf{w}, \mathbf{y})$ is jointly L_g-smooth and μ_g-strongly convex in terms of $\mathbf{y}$ for any $\mathbf{w}$*
- *f is twice continuously differentiable, and $\nabla^2 g(\mathbf{w}, \mathbf{y})$ is jointly Q_g-Lipschitz continuous.*
- *$F(\mathbf{w})$ is lower bounded, i.e., $\inf_{\mathbf{w}\in\mathbb{R}^{d_w}} F(\mathbf{w}) > -\infty$*

Assumption 2. *There exist stochastic gradients of f and g that are unbiased and have bounded variance σ_f^2, σ_g^2, respectively, i.e.*

$$\begin{aligned} \mathbf{E}_{\xi_f}[\nabla f(\mathbf{w}, \mathbf{y}, \xi_f)] = \nabla f(\mathbf{w}, \mathbf{y}), \quad \mathbf{E}_{\xi_g}[\nabla g(\mathbf{w}, \mathbf{y}, \xi_g)] = \nabla g(\mathbf{w}, \mathbf{y}) \\ \mathbf{E}_{\xi_f}\|\nabla f(\mathbf{w}, \mathbf{y}, \xi_f) - \nabla f(\mathbf{w}, \mathbf{y})\|^2 \leq \sigma_f^2, \quad \mathbf{E}_{\xi_g}\|\nabla g(\mathbf{w}, \mathbf{y}, \xi_g) - \nabla g(\mathbf{w}, \mathbf{y})\|^2 \leq \sigma_g^2. \end{aligned} \tag{4}$$

For deriving improved convergence rate, we additionally consider the following mean-square smoothness (MSS) assumption regarding stochastic gradient of g.

Assumption 3. *g is mean-square smooth in $\mathbf{y}$, i.e.*

$$\mathbf{E}_{\xi_g}\|\nabla g(\mathbf{w}, \mathbf{y}_1, \xi_g) - \nabla g(\mathbf{w}, \mathbf{y}_2, \xi_g)\|^2 \leq \tilde{L}_g^2\|\mathbf{y}_1 - \mathbf{y}_2\|^2. \tag{5}$$

We follow the tradition from (Chen et al., 2025) and define $\kappa := \bar{\ell}/\mu_g$ where $\bar{\ell} := \max\{C_f, L_f, L_g, Q_g\}$. When lower-level objective is also $\tilde{L}_g$-MSS, we slightly abuse the notation and define $\bar{\ell} := \max\{C_f, L_f, L_g, \tilde{L}_g, Q_g\}$.

3.2 Penalty Surrogate

Recall the penalized problem in (3). For any $\mathbf{w}$, we denote the unique minimizer of the strongly convex lower-level problem by

$$\mathbf{y}^*(\mathbf{w}) := \arg\min_{\mathbf{z}} g(\mathbf{w}, \mathbf{z}). \tag{6}$$

For a penalty parameter $\lambda > 0$, define

$$\mathcal{L}_\lambda(\mathbf{w}, \mathbf{y}) := f(\mathbf{w}, \mathbf{y}) + \lambda\left(g(\mathbf{w}, \mathbf{y}) - \min_{\mathbf{z}} g(\mathbf{w}, \mathbf{z})\right). \tag{7}$$

When $\lambda \geq 2L_f/\mu_g$, $\lambda^{-1}f(\mathbf{w}, \cdot) + g(\mathbf{w}, \cdot)$ is $(\mu_g - \frac{L_f}{\lambda})$-strongly convex and we define $\mathbf{y}_\lambda^*(\mathbf{w})$ as the unique minimizer of the penalized function $\mathcal{L}_\lambda(\mathbf{w}, \mathbf{y})$ with respect to the lower-level variable given $\mathbf{w}$, i.e.,

$$\mathbf{y}_\lambda^*(\mathbf{w}) := \arg\min_{\mathbf{y}} \mathcal{L}_\lambda(\mathbf{w}, \mathbf{y}) = \arg\min_{\mathbf{y}} \left\{\lambda^{-1}f(\mathbf{w}, \mathbf{y}) + g(\mathbf{w}, \mathbf{y})\right\}. \tag{8}$$

The penalized objective introduced in (3) can then be written as

$$\mathcal{L}_\lambda^*(\mathbf{w}) := \mathcal{L}_\lambda(\mathbf{w}, \mathbf{y}_\lambda^*(\mathbf{w})) = f(\mathbf{w}, \mathbf{y}_\lambda^*(\mathbf{w})) + \lambda\left(g(\mathbf{w}, \mathbf{y}_\lambda^*(\mathbf{w})) - g(\mathbf{w}, \mathbf{y}^*(\mathbf{w}))\right). \tag{9}$$

Moreover,

$$\begin{aligned} \nabla\mathcal{L}_\lambda^*(\mathbf{w}) &= \nabla_1\mathcal{L}_\lambda(\mathbf{w}, \mathbf{y}_\lambda^*(\mathbf{w})) + \nabla\mathbf{y}_\lambda^*(\mathbf{w})\underbrace{\nabla_2\mathcal{L}_\lambda(\mathbf{w}, \mathbf{y}_\lambda^*(\mathbf{w}))}_{=\mathbf{0}} \\ &= \nabla_1 f\big(\mathbf{w}, \mathbf{y}_\lambda^*(\mathbf{w})\big) + \lambda\left[\nabla_1 g\big(\mathbf{w}, \mathbf{y}_\lambda^*(\mathbf{w})\big) - \nabla_1 g\big(\mathbf{w}, \mathbf{y}^*(\mathbf{w})\big)\right], \end{aligned} \tag{10}$$

which involves only first-order derivatives once $\mathbf{y}_\lambda^*(\mathbf{w})$ and $\mathbf{y}^*(\mathbf{w})$ are available as described in the previous section.

We next present the approximation and smoothness properties of the penalty surrogate as established in previous work.

Lemma 1 (Lemma 4.1 of Chen et al. (2025)). *Under Assumption 1, for $\lambda \geq 2L_f/\mu_g$, it holds for $\forall \mathbf{w} \in \mathbb{R}^{d_w}$ that*

$$|\mathcal{L}_\lambda^*(\mathbf{w}) - F(\mathbf{w})| = O\left(\frac{\bar{\ell}\kappa^2}{\lambda}\right), \quad \|\nabla \mathcal{L}_\lambda^*(\mathbf{w}) - \nabla F(\mathbf{w})\| = O\left(\frac{\bar{\ell}\kappa^3}{\lambda}\right), \tag{11}$$

$$\|\nabla \mathcal{L}_\lambda^*(\mathbf{w}_1) - \nabla \mathcal{L}_\lambda^*(\mathbf{w}_2)\| \leq L_* \|\mathbf{w}_1 - \mathbf{w}_2\|, \text{ where } L_* = O(\bar{\ell}\kappa^3). \tag{12}$$

We make two remarks regarding this lemma. First, choosing $\lambda = \Omega(\bar{\ell}\kappa^3/\epsilon)$ ensures that the gradient of the penalty objective approximates the true hypergradient to within $O(\epsilon)$. Consequently, any $O(\epsilon)$-stationary point of $\mathcal{L}_\lambda^*$ is also an $O(\epsilon)$-stationary point of F. Second, although λ must increase as ϵ decreases, the smoothness constant L_* of the reduced penalty objective does *not* grow with λ. Thus, increasing the penalty parameter does not, by itself, require a smaller outer step size on smoothness grounds. The main challenge posed by a large λ lies in stochastic gradient estimation, where the penalty factor can amplify the estimation error.

4 A PENALTY-BASED FIRST-ORDER SINGLE LOOP METHOD

In this section, we present our stochastic single-loop constant-batch first-order method. We denote by $\mathcal{B}_{fw}^t, \mathcal{B}_{fy}^t$ independent batches for evaluating stochastic gradients of $f(\mathbf{w}, \mathbf{y})$ at the t-th iteration. Similarly, denote by $\mathcal{B}_{gw}^t, \mathcal{B}_{gy}^t$ independent batches for evaluating stochastic gradients of $g(\mathbf{w}, \mathbf{y})$ at the t-th iteration. We assume $|\mathcal{B}_{fw}^t| = |\mathcal{B}_{fy}^t| = |\mathcal{B}_{gw}^t| = |\mathcal{B}_{gy}^t| = B = O(1)$.

4.1 THE ALGORITHM

We begin by motivating the proposed algorithm. By equation (10), the gradient of $\mathcal{L}_\lambda^*(\mathbf{w})$ has the compositional form $\mathcal{M}(\mathbf{w}, \mathbf{y}^*(\mathbf{w}), \mathbf{y}_\lambda^*(\mathbf{w}))$, where $\mathcal{M}$ is Lipschitz continuous and $\mathbf{y}^*(\mathbf{w})$ and $\mathbf{y}_\lambda^*(\mathbf{w})$ are solutions to strongly convex optimization problems. This structure fits the compositional gradient framework discussed in (Yang, 2026, Section 4.5), which provides a general scheme for estimating such gradients.

To estimate the gradient in equation (10), we maintain two sequences, $\mathbf{y}_{t+1}$ and $\mathbf{z}_{t+1}$, to track $\mathbf{y}_\lambda^*(\mathbf{w}_t)$ and $\mathbf{y}^*(\mathbf{w}_t)$, respectively. Substituting these iterates for the corresponding optimal solutions and replacing the full gradients with stochastic estimates yields

$$\hat{G}_t^w \coloneqq \nabla_1 f(\mathbf{w}_t, \mathbf{y}_{t+1}, \mathcal{B}_{fw}^t) + \lambda \left(\nabla_1 g(\mathbf{w}_t, \mathbf{y}_{t+1}, \mathcal{B}_{gw}^t) - \nabla_1 g(\mathbf{w}_t, \mathbf{z}_{t+1}, \mathcal{B}_{gw}^t)\right). \tag{13}$$

Because the iterates $\mathbf{y}_{t+1}, \mathbf{z}_{t+1}$ only approximate their respective targets, this estimator is generally biased: $G_t^w \coloneqq \mathbf{E}_t[\hat{G}_t^w] \neq \nabla \mathcal{L}_\lambda^*(\mathbf{w}_t)$. To reduce stochastic fluctuations, we use an exponential moving average with $\beta \in (0, 1)$,

$$\mathbf{v}_{t+1} \coloneqq (1 - \beta)\mathbf{v}_t + \beta \hat{G}_t^w, \tag{14}$$

and update the primal variable by

$$\mathbf{w}_{t+1} \coloneqq \mathbf{w}_t - \eta \mathbf{v}_{t+1}. \tag{15}$$

Controlling the **tracking-induced squared bias** $\|G_t^w - \nabla \mathcal{L}_\lambda^*(\mathbf{w}_t)\|^2$, however, requires suitable updates for the inner iterates. A natural choice, following (Yang, 2026), is to perform a single SGD step for each inner problem at every iteration. However, when λ scales as $1/\epsilon$, this approach yields a complexity of $O(\epsilon^{-7})$. The bottleneck is controlling the squared bias $\|G_t^w - \nabla \mathcal{L}_\lambda^*(\mathbf{w}_t)\|^2$ despite the large penalty parameter. Specifically, for $\lambda \geq 1$, a direct Lipschitz bound gives

$$\|G_t^w - \nabla \mathcal{L}_\lambda^*(\mathbf{w}_t)\|^2 \leq O(\lambda^2) \left(\|\mathbf{y}_{t+1} - \mathbf{y}_\lambda^*(\mathbf{w}_t)\|^2 + \|\mathbf{z}_{t+1} - \mathbf{y}^*(\mathbf{w}_t)\|^2\right).$$

This bound amplifies both tracking errors by a factor of $O(\lambda^2)$, requiring increasingly accurate inner solutions as ϵ decreases and thereby worsening the overall complexity.

Projected lower-level tracking. Inspired by Kwon et al. (2024), our algorithm exploits the proximity of $\mathbf{y}_\lambda^*(\mathbf{w})$ and $\mathbf{y}^*(\mathbf{w})$. Specifically, for every $\mathbf{w}$, we have (see Appendix A)

$$\|\mathbf{y}_\lambda^*(\mathbf{w}) - \mathbf{y}^*(\mathbf{w})\| \leq r_\lambda \coloneqq \frac{C_f}{\mu_g \lambda}. \tag{16}$$

Algorithm 1 Stochastic Single-loop Fully First-order Penalty Method (SICO) for SBO

Input: T, λ, $\mathbf{w}_0 = \mathbf{w}_{-1}$, $\mathbf{y}_0 = \mathbf{y}_{-1}$, $\mathbf{z}_0 = \Pi_{\mathbb{B}(\mathbf{y}_0, r_\lambda)}(\mathbf{z}_{-1})$, $\mathbf{v}_0 = \mathbf{0}$, η, γ, β, $r_\lambda = C_f/(\mu_g \lambda)$.
for $t = 0, 1, \cdots, T-1$ **do**
  Sample independent constant-size batches $\mathcal{B}_{fy}^t, \mathcal{B}_{gy}^t, \mathcal{B}_{fw}^t, \mathcal{B}_{gw}^t$
  Compute $\hat{G}_t^y = \lambda^{-1}\nabla_2 f(\mathbf{w}_t, \mathbf{y}_t, \mathcal{B}_{fy}^t) + \nabla_2 g(\mathbf{w}_t, \mathbf{y}_t, \mathcal{B}_{gy}^t)$ and $\hat{G}_t^z = \nabla_2 g(\mathbf{w}_t, \mathbf{z}_t, \mathcal{B}_{gy}^t)$.
  Update $\mathbf{y}_{t+1} = \mathbf{y}_t - \gamma \hat{G}_t^y$ and $\bar{\mathbf{z}}_t = \mathbf{z}_t - \gamma \hat{G}_t^z$.
  Project $\mathbf{z}_{t+1} = \Pi_{\mathbb{B}(\mathbf{y}_{t+1}, r_\lambda)}(\bar{\mathbf{z}}_t)$.
  Compute $\hat{G}_t^w = \nabla_1 f(\mathbf{w}_t, \mathbf{y}_{t+1}, \mathcal{B}_{fw}^t) + \lambda \left(\nabla_1 g(\mathbf{w}_t, \mathbf{y}_{t+1}, \mathcal{B}_{gw}^t) - \nabla_1 g(\mathbf{w}_t, \mathbf{z}_{t+1}, \mathcal{B}_{gw}^t)\right)$.
  Update $\mathbf{v}_{t+1} = (1-\beta)\mathbf{v}_t + \beta \hat{G}_t^w$ and $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{v}_{t+1}$.
end for
Output: $\mathbf{w}_\tau$ with τ uniformly sampled from $\{0, 1, \cdots, T-1\}$.

We use this bound to couple the updates of $\mathbf{y}$ and $\mathbf{z}$. After taking an SGD step for each variable, we project the updated $\mathbf{z}$ onto the ball of radius r_λ centered at the updated $\mathbf{y}$. The resulting updates are

$$\hat{G}_t^y = \lambda^{-1}\nabla_2 f(\mathbf{w}_t, \mathbf{y}_t, \mathcal{B}_{fy}^t) + \nabla_2 g(\mathbf{w}_t, \mathbf{y}_t, \mathcal{B}_{gy}^t), \qquad \hat{G}_t^z = \nabla_2 g(\mathbf{w}_t, \mathbf{z}_t, \mathcal{B}_{gy}^t), \tag{17}$$

$$\mathbf{y}_{t+1} = \mathbf{y}_t - \gamma \hat{G}_t^y, \qquad \bar{\mathbf{z}}_t = \mathbf{z}_t - \gamma \hat{G}_t^z, \tag{18}$$

$$\mathbf{z}_{t+1} = \Pi_{\mathbb{B}(\mathbf{y}_{t+1}, r_\lambda)}(\bar{\mathbf{z}}_t). \tag{19}$$

Combining these projected lower-level updates with the momentum update of $\mathbf{w}$ yields our algorithm, SICO, presented in Algorithm 1. Compared with (Kwon et al., 2024), our main algorithmic modification is to apply this projection at every iteration and combine it with momentum-based gradient estimation. This yields a single-loop method that requires neither inner loops nor large batches supported by our novel analysis.

We next present convergence guarantees for SICO with and without the mean-square smoothness assumption on g, and then highlight the novelty of our analysis.

4.2 Convergence under standard bounded variance

We first consider Assumptions 1 and 2, without imposing mean-square smoothness on the stochastic gradient of g. The following theorem shows that Algorithm 1 matches the $O(\epsilon^{-6})$ dependence of the best known first order methods while using a single loop and only a constant batch size.

Theorem 1. *Under Assumptions 1 and 2, let the batch size be $B = O(1)$. With a warm-up stage for $\mathbf{y}$ such that the initialization satisfies the condition that $\mathbf{E}[\|\mathbf{y}_0 - \mathbf{y}_\lambda^*(\mathbf{w}_0)\|^2] \le O(1/\kappa)$, by setting*

$$\lambda = \Theta\left(\frac{\kappa^3}{\epsilon}\right), \qquad \beta = \Theta\left(\frac{\epsilon^4}{\kappa^6}\right), \qquad \gamma = \Theta\left(\frac{\epsilon^4}{\kappa^7}\right), \qquad \eta = \Theta\left(\frac{\epsilon^4}{\kappa^{11}}\right),$$

Algorithm 1 returns $\mathbf{w}_\tau$ satisfying

$$\mathbf{E}\left\|\nabla F(\mathbf{w}_\tau)\right\|^2 \le \epsilon^2$$

within $T = O(\kappa^{11}\epsilon^{-6})$ stochastic first-order oracle calls, where τ is uniformly sampled from $\{0, 1, \cdots, T-1\}$.

Remark: Note that the involved optimization problem in (8) that define $\mathbf{y}_\lambda^*(\mathbf{w}_0)$ is strongly convex, hence the warm-up stage is cheap. For example, running a simple SGD for $T_0 = O(\kappa^3)$-iterations for solving $\min_{\mathbf{y}} \mathcal{L}_\lambda(\mathbf{w}_0, \mathbf{y})$ will find a solution $\mathbf{y}_0$ such that $\mathbf{E}[\|\mathbf{y}_0 - \mathbf{y}_\lambda^*(\mathbf{w}_0)\|^2] \le O(1/(\mu_g^2 T_0)) = O(1/\kappa)$.

4.3 Improved rate under mean-square smoothness

We next establish an improved rate assuming the mean-square smoothness of g as in Assumption 3. The intuition behind the improvement is that under mean-square smoothness of g, **the variance of the estimator** $\hat{G}_t^w$ will become independent of λ. As a result, the same algorithm therefore provides a stronger benefit under Assumption 3, leading to the improved rate below.

Theorem 2. *Under Assumptions 1, 2 and 3, let the batch size be $B = O(1)$. With the same warm-up initialization and by setting*

$$\lambda = \Theta\left(\frac{\kappa^3}{\epsilon}\right), \qquad \beta = \Theta\left(\frac{\epsilon^2}{\kappa^2}\right), \qquad \gamma = \Theta\left(\frac{\epsilon^2}{\kappa^5}\right), \qquad \eta = \Theta\left(\frac{\epsilon^2}{\kappa^9}\right),$$

Algorithm 1 returns $\mathbf{w}_\tau$ satisfying

$$\mathbf{E}\left\|\nabla F(\mathbf{w}_\tau)\right\|^2 \le \epsilon^2$$

within $T = O(\kappa^9\epsilon^{-4})$ stochastic first-order oracle calls, where τ is uniformly sampled from $\{0, 1, \cdots, T-1\}$.

Remark: To the best of our knowledge, this rate also has the best dependence on κ compared with that of other fully first-order methods.

4.4 Sketch of Analysis

We outline the convergence analysis and highlight its novel elements. Let $\mathbf{E}_t$ denote expectation over the randomness at iteration t, conditional on $\mathbf{w}_t, \mathbf{y}_t, \mathbf{z}_t$. Define

$$\mathbf{s}_t \coloneqq \mathbf{y}_t - \mathbf{z}_t,\ Y_t \coloneqq \left\|\mathbf{y}_{t+1} - \mathbf{y}^*_\lambda(\mathbf{w}_t)\right\|^2,\ S_t \coloneqq \left\|\mathbf{s}_{t+1} - \mathbf{s}^*(\mathbf{w}_t)\right\|^2,\ V_t \coloneqq \left\|\mathbf{v}_{t+1} - \nabla\mathcal{L}^*_\lambda(\mathbf{w}_t)\right\|^2.$$

The smoothness of $\mathcal{L}^*_\lambda$ in (12) yields the following descent inequality for $\eta \le 1/(2L_*)$ (see Lemma 5):

$$\mathbf{E}_t\mathcal{L}^*_\lambda(\mathbf{w}_{t+1}) \le \mathcal{L}^*_\lambda(\mathbf{w}_t) + \frac{\eta}{2}\mathbf{E}_t V_t - \frac{\eta}{4}\mathbf{E}_t\left\|\mathbf{v}_{t+1}\right\|^2 - \frac{\eta}{2}\left\|\nabla\mathcal{L}^*_\lambda(\mathbf{w}_t)\right\|^2. \tag{20}$$

The moving-average update further gives the following recursion for the gradient estimation error (see Lemma 7):

$$\mathbf{E}_t V_{t+1} \le (1-\beta)V_t + \frac{2L_*^2\eta^2}{\beta}\left\|\mathbf{v}_{t+1}\right\|^2 + 4\beta\left\|G^w_{t+1} - \nabla\mathcal{L}^*_\lambda(\mathbf{w}_{t+1})\right\|^2 + \beta^2\mathbf{E}_t\left\|\hat{G}^w_{t+1} - G^w_{t+1}\right\|^2.$$

The next step is to control the squared bias $\left\|G^w_{t+1} - \nabla\mathcal{L}^*_\lambda(\mathbf{w}_{t+1})\right\|^2$ and the variance $\mathbf{E}_t\left\|\hat{G}^w_{t+1} - G^w_{t+1}\right\|^2$. These terms are weighted by β and β^2, respectively. Thus, the moving average suppresses the variance more strongly than the squared bias, making tight control of the bias particularly important. For the variance, we leverage the following bounds (see Lemma 6):

$$\text{Under Assumption 2:}\quad \mathbf{E}_t\left\|\hat{G}^w_{t+1} - G^w_{t+1}\right\|^2 \le \frac{1}{B}\left(2\sigma_f^2 + 8\lambda^2\sigma_g^2\right), \tag{21}$$

$$\text{Under Assumptions 2 and 3:}\quad \mathbf{E}_t\left\|\hat{G}^w_{t+1} - G^w_{t+1}\right\|^2 \le \frac{1}{B}\left(2\sigma_f^2 + \frac{8\tilde{L}_g^2C_f^2}{\mu_g^2}\right). \tag{22}$$

In particular, mean-square smoothness removes the dependence on λ from the variance bound.

For the squared bias, our analysis relies on the following decomposition (see Lemma 8):

$$\left\|G^w_t - \nabla\mathcal{L}^*_\lambda(\mathbf{w}_t)\right\|^2 \le O(1)Y_t + O(\lambda^2)S_t + \frac{O(\kappa^4)}{\lambda^2}. \tag{23}$$

The arguments up to this point largely follow those developed in prior work (Yang, 2026; Kwon et al., 2024).

It remains to establish recursions for the tracking errors Y_t and S_t. The recursion for Y_t follows standard SGD analysis for strongly convex objectives. The main departure from prior analyses lies in controlling S_t. Let $G^s_t \coloneqq \mathbf{E}_t[\hat{G}^y_t - \hat{G}^z_t]$ and $\mathbf{s}^*_t \coloneqq \mathbf{s}^*(\mathbf{w}_t)$. The analysis of Kwon et al. (2024) relies on an inequality of the form

$$\langle G^s_t,\ \mathbf{s}_t - \mathbf{s}^*_t\rangle \ge c\left\|\mathbf{s}_t - \mathbf{s}^*_t\right\|^2 \tag{24}$$

for some $c > 0$. However, this inequality does not hold in general, as demonstrated by the counterexample in Appendix D.

To establish a valid contraction bound, we introduce the mapping

$$H(\mathbf{w}, \mathbf{y}, \mathbf{s}) \coloneqq \lambda^{-1}\nabla_2 f(\mathbf{w}, \mathbf{y}) + \nabla_2 g(\mathbf{w}, \mathbf{y}) - \nabla_2 g(\mathbf{w}, \mathbf{y} - \mathbf{s}). \tag{25}$$

For fixed $(\mathbf{w}, \mathbf{y})$, the mapping $H(\mathbf{w}, \mathbf{y}, \cdot)$ is μ_g-strongly monotone. Using this property, we establish the following lower bound (see Lemma 10):

$$\langle G_t^s,\, \mathbf{s}_t - \mathbf{s}_t^* \rangle \geq \frac{\mu_g}{2} \left\| \mathbf{s}_t - \mathbf{s}_t^* \right\|^2 - \frac{1}{\mu_g} \left(\frac{L_f^2}{\lambda^2} + Q_g^2 r_\lambda^2 \right) \left\| \mathbf{y}_t - \mathbf{y}_\lambda^*(\mathbf{w}_t) \right\|^2 . \tag{26}$$

This bound accounts explicitly for the $\mathbf{y}$-tracking error. Since $r_\lambda = C_f/(\mu_g \lambda)$, the coefficient of this additional error term is $O(\lambda^{-2})$. This bound is central to deriving the one-step recursion for the $\mathbf{s}$-tracking error in Lemma 11.

Combining these bounds yields Theorems 1 and 2. The complete proofs are provided in Appendices B and C.

5 Numerical experiments

We evaluate SICO on two stochastic bilevel learning tasks: data hyper-cleaning and feature-wise regularization selection. We compare against F2SA and F3SA (Kwon et al., 2023), PMLLC[1] (Kwon et al., 2024), F2BSA (Chen et al., 2025), and Stoc-SGHA (Gu et al., 2026).

Implementation and hyperparameter tuning. All methods use logistic regression with an intercept, mini-batches of size 64, and a budget of 10,000 inner iterations, with evaluation every 100 iterations. Training and validation batches are sampled independently. Double loop method F2BSA and PMLLC perform five inner updates per outer update. We tune upper level learning rate η in $\{10^{-2}, 10^{-1}, 0.5, 1, 3, 10, 10^2, 10^3\}$ and lower level learning rate γ in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$; r_λ in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 2\}$; λ in $\{10, 20, 30, 100, 300, 1000\}$. For more tuning detail please refer to Appendix E. Experiments are conducted on an NVIDIA A30 GPU.

Data Hyper-Cleaning. In data hyper-cleaning, the training set contains corrupted labels, while a small clean validation set is available. An upper-level variable $\mathbf{w}_i$ is introduced to control the importance of each training example. Let $\mathcal{D}_{\mathrm{tr}} = \{(a_i, b_i)\}_{i=1}^n, \mathcal{D}_{\mathrm{val}} = \{(a_j^{\mathrm{val}}, b_j^{\mathrm{val}})\}_{j=1}^m$. We consider

$$\min_{\mathbf{w} \in \mathbb{R}^n} \frac{1}{m} \sum_{j=1}^m \ell\left(\mathbf{y}^*(\mathbf{w}); a_j^{\mathrm{val}}, b_j^{\mathrm{val}}\right)$$

$$\mathbf{y}^*(\mathbf{w}) = \arg\min_{\mathbf{y}} \frac{1}{n} \sum_{i=1}^n \sigma(\mathbf{w}_i) \ell(\mathbf{y}; a_i, b_i) + \frac{\mu}{2} \left\| \mathbf{y} \right\|^2 ,$$

where $\sigma(\cdot)$ is the sigmoid function and ℓ is the cross-entropy loss. We use a multiclass logistic regression classifier, and the quadratic regularizer guarantees strong convexity of the lower-level problem. Stochastic upper- and lower-level gradients are computed using independently sampled validation and training mini-batches.

We use MNIST and FASHION-MNIST to verify our algorithm on this task. MNIST and Fashion-MNIST (LeCun et al., 1998; Xiao et al., 2017) contain 28×28 grayscale images of handwritten digits and clothing items, respectively, with ten classes each. For each dataset, we randomly partition the official 60,000 training images into 50,000 training and 10,000 validation examples and retain the official 10,000 test images. Each training label is independently replaced by a uniformly sampled incorrect class with probability $p \in \{0.5, 0.9, 0.99\}$; validation and test labels remain clean. We fix $\mu = 0.1$. Following previous studies Chen et al. (2025); Gu et al. (2026), we report the average learned weight on corrupted examples, $|G|^{-1} \sum_{i \in G} \sigma(w_i)$, where G is the corruption set. Figure 1 shows that our algorithm learns a smallest weight for garbage data and convergence is the fastest.

Regularization Selection. We next consider learning feature-wise regularization parameters for logistic regression. Let $\mathbf{y}$ denote the classifier parameters and let $\mathbf{w}_k$ control the regularization

[1]The full name is *Penalty Methods with Lower-Level Coupling*. Since the original authors did not provide an official abbreviation, we introduce one here for ease of presentation.

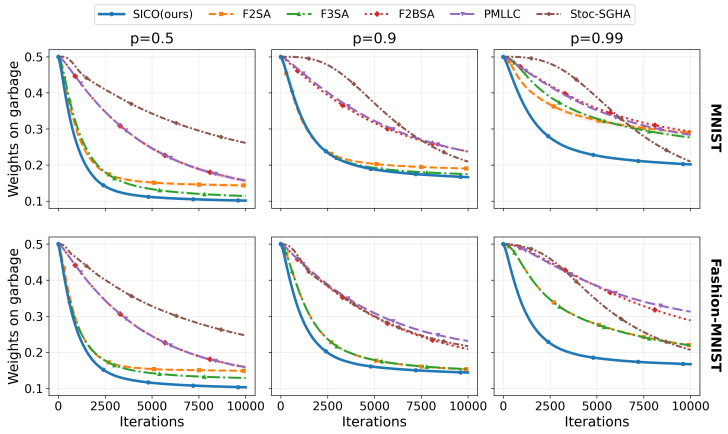


Figure 1: Comparison of first order method of NC-SC SBO problem for data hyper-cleaning task. Upper row: MNIST dataset; Lower row: Fashion-MNIST dataset.

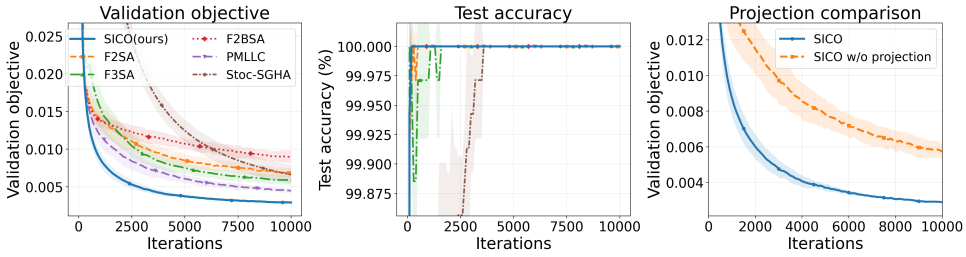


Figure 2: Left two: Comparison of first order method of NC-SC SBO problem for regularization selection task on the Mushroom dataset. Right: Ablation study on SICO vs w/o projection.

strength of the k-th feature. We solve

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{m}\sum_{j=1}^{m} \ell\left(\mathbf{y}^*(\mathbf{w}); a_j^{\text{val}}, b_j^{\text{val}}\right)$$

$$\mathbf{y}^*(\mathbf{w}) = \arg\min_{\mathbf{y}} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{y}; a_i, b_i) + \frac{1}{2d}\sum_{k=1}^{d} \exp(w_k)\left\|\mathbf{y}_{:,k}\right\|^2 + \frac{\mu_0}{2d}\left\|\mathbf{y}\right\|_F^2 .$$

We use the LIBSVM version of Mushroom dataset (Chang & Lin, 2011), which contains 8124 examples from two classes and 112 binary features encoding categorical attributes. A seeded random partition yields $5802/1161/1161$ training/validation/test examples. Features are standardized using training-set means and standard deviations, with unit scale for constant features; labels are mapped to $\{0, 1\}$. We fix $\mu_0 = 0.1$ and divide both quadratic penalties by the number of feature weights ($d = 112$). The intercept receives only the fixed penalty. Figure 2 (left two) reports validation cross-entropy and test accuracy and shows that our method reach the lowest validation error.

Finally, we conduct an ablation study on the regularization selection task comparing SICO with the variant without the projection step on $\mathbf{z}$. The result shown in Figure 2 shows that projection improves the convergence of SICO.

6 Conclusions

We have proposed a fully first-order method for non-convex-strongly-convex stochastic bilevel optimization. Building on the penalty framework and projected lower-level coupling, we combine projected tracking of lower level variables with a momentum update of the upper level variable to obtain a single-loop algorithm using only $O(1)$ batch size. Our refined analysis establishes $O(\kappa^{11}\epsilon^{-6})$ complexity under standard bounded variance and $O(\kappa^{9}\epsilon^{-4})$ complexity under the additional mean-square smoothness assumption, matching the best-known rates of existing fully first-order methods without requiring large batches or nested loops.

7 STATEMENTS

AI USE STATEMENT

In this work, we used generative AI tools to polish the main paper writing. We also use them to help build the codebase for numerical experiments presented in section 5. We have reviewed all the AI-assisted work. Specifically, 2 of the authors reviewed and verified the training and evaluation pipeline and algorithms implementation(including our method and baseline methods) to make sure they are correct. Besides, we use AI tools to help construct the counter example in appendix D. All the remaining required disclosure tasks are either inapplicable to this work or are done by ourselves. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

For theoretical results, we provide formal assumption and theorem and necessary explanation in the main paper, and technical lemma with their proof in appendix A, and proof of theorems in appendix B and C.

For experiment results, we manually set the random seed for each experiment and provide the detail information about experiment setting in 5. All the datasets used in our experiment are open source. All the aforementioned efforts help make sure that all of presented result are reproducible.

REFERENCES

Fan Bao, Guoqiang Wu, Chongxuan Li, Jun Zhu, and Bo Zhang. Stability and generalization of bilevel programming in hyperparameter optimization. *Advances in neural information processing systems*, 34:4529–4541, 2021.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

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. URL http://jmlr.org/papers/v26/23-1104.html.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=r6cNUjS8cm0.

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

Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. Self-guided noise-free data generation for efficient zero-shot learning. *arXiv preprint arXiv:2205.12679*, 2022.

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

Zhihao Gu, Qilong Wu, and Junchi Yang. Sgha: A single-loop fully first-order algorithm for nonconvex-strongly-convex bilevel optimization. *arXiv preprint arXiv:2608.23211*, 2026.

Zhishuai Guo, Quanqi Hu, Lijun Zhang, and Tianbao Yang. Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *CoRR*, abs/2105.02266, 2021a. URL https://arxiv.org/abs/2105.02266.

Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. Unified convergence analysis for adaptive optimization with moving average estimator. *Mach. Learn.*, 114(4), February 2021b. ISSN 0885-6125. doi: 10.1007/s10994-024-06650-8. URL https://doi.org/10.1007/s10994-024-06650-8.

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, 2020.

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

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 *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

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*, pp. 18083–18113. PMLR, 2023.

Jeongyeol Kwon, Dohyun Kwon, and Hanbaek Lyu. On the complexity of first-order methods in stochastic bilevel optimization. *arXiv preprint arXiv:2402.07101*, 2024.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. Bome! bilevel optimization made easy: a simple first-order approach. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Zhaosong Lu and Sanyou Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024. doi: 10.1137/23M1566753.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30992–31015. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/shen23c.html.

Stephen J. Wright. Gradient-based methods for large-scale bilevel optimization. Plenary talk at the 8th International Conference on Continuous Optimization (ICCOPT 2025), July 2025. URL https://sites.google.com/view/iccopt2025/home/speakers/plenary-speakers. University of Southern California, Los Angeles, CA, USA.

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

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

T. Yang. *Compositional Optimization for Advanced Machine Learning*. Springer Optimization and Its Applications. Springer Nature Switzerland, 2026. ISBN 9783032343574. URL https://books.google.com/books?id=0E4h0gEACAAJ.

Jane J. Ye and Daoli Zhu. Optimality conditions for bilevel programming problems. *Optimization*, 33(1):9–27, 1995. doi: 10.1080/02331939508844060.

A TECHNICAL LEMMA

A.1 PRELIMINARY

Notation For convenience, let us first introduce the following notations:

$$\mathbf{y}_{\lambda,t}^* \coloneqq \mathbf{y}_\lambda^*(\mathbf{w}_t), \mathbf{s}_t^* \coloneqq \mathbf{s}^*(\mathbf{w}_t),$$

and denote $\mathbf{I}$ as identity matrix.

Below we will introduce several useful lemma.

Lemma 2. *Under assumption 1, for $\lambda \geq \frac{2L_f}{\mu_g}$, it holds that*

$$\|\mathbf{y}^*(\mathbf{w}) - \mathbf{y}_\lambda^*(\mathbf{w})\| \leq \frac{C_f}{\mu_g \lambda} \tag{27}$$

Proof. It holds from definition of $\mathbf{y}_\lambda^*(\mathbf{w})$ that $\nabla_2 f(\mathbf{w}, \mathbf{y}_\lambda^*(\mathbf{w})) + \lambda \nabla_2 g(\mathbf{w}, \mathbf{y}_\lambda^*(\mathbf{w})) = \mathbf{0}$. Then we have

$$\|\mathbf{y}^*(\mathbf{w}) - \mathbf{y}_\lambda^*(\mathbf{w})\| \leq \frac{1}{\mu_g} \|\nabla_2 g(\mathbf{w}, \mathbf{y}_\lambda^*(\mathbf{w}))\| = \frac{1}{\mu_g \lambda} \|\nabla_2 f(\mathbf{w}, \mathbf{y}_\lambda^*(\mathbf{w}))\| \leq \frac{C_f}{\mu_g \lambda}$$

where the first inequality holds from strongly convexity of $g(\mathbf{w}, \cdot)$. □

Lemma 3. *Under assumption 1, for $\lambda \geq \frac{2L_f}{\mu_g}$, $\mathbf{y}_\lambda^*(\cdot)$ is $C_{y*} = \frac{2(L_f + \lambda L_g)}{\lambda \mu_g} = O(\kappa)$-lipschitz.*

Proof. It holds from the definition of $\mathbf{y}_\lambda^*(\mathbf{w})$ that $\nabla_2 \mathcal{L}_\lambda^*(\mathbf{w}, \mathbf{y}_\lambda^*(\mathbf{w})) = \mathbf{0}$. Differentiate LHS w.r.t. $\mathbf{w}$, we have

$$\nabla_{21}^2 \mathcal{L}_\lambda^*(\mathbf{w}, y^*(\mathbf{w})) + \nabla y^*(\mathbf{w}) \nabla_{22}^2 \mathcal{L}_\lambda^*(\mathbf{w}, y^*(\mathbf{w})) = \mathbf{0}.$$

Rearranging the term and multiplying both side by $\nabla_{22}^2 \mathcal{L}_\lambda^*(\mathbf{w}, y^*(\mathbf{w}))^{-1}$ yields

$$\nabla y_\lambda^*(\mathbf{w}) = -\left(\frac{1}{\lambda} \nabla_{21}^2 f(\mathbf{w}, y_\lambda^*(\mathbf{w})) + \nabla_{21}^2 g(\mathbf{w}, y_\lambda^*(\mathbf{w}))\right) \left(\frac{1}{\lambda} \nabla_{22}^2 f(\mathbf{w}, y_\lambda^*(\mathbf{w})) + \nabla_{22}^2 g(\mathbf{w}, y_\lambda^*(\mathbf{w}))\right)^{-1} \tag{28}$$

It finishes the proof by noting that $\nabla_{22}^2 f(\mathbf{w}, y_\lambda^*(\mathbf{w})) \preceq L_f \mathbf{I}, \nabla_{21}^2 f(\mathbf{w}, y_\lambda^*(\mathbf{w})) \preceq L_f \mathbf{I}$ and $\nabla_{22}^2 g(\mathbf{w}, y_\lambda^*(\mathbf{w})) \succeq \mu_g \mathbf{I}, \nabla_{21}^2 g(\mathbf{w}, y_\lambda^*(\mathbf{w})) \preceq L_g \mathbf{I}$ □

Lemma 4 (Lemma B.5 of Chen et al. (2025))**.** *Under assumption 1, if $\lambda \geq \frac{2L_f}{\mu_g}$, we have that for any t:*

$$\left\|\mathbf{s}_t^* - \mathbf{s}_{t-1}^*\right\|^2 = \|\mathbf{s}^*(\mathbf{w}_t) - \mathbf{s}^*(\mathbf{w}_{t-1})\|^2 \leq \frac{C_s^2}{\lambda^2} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2 \tag{29}$$

where $C_s = (\frac{1}{\mu_g} + \frac{2L_g}{\mu_g^2})(L_f + \frac{C_f Q_g}{\mu_g}) = O(\kappa^3)$.

A.2 LEMMA FOR THEOREM 1

In this section we introduce several important technical lemma for theorem 1.

The following lemma is a standard descent lemma for smooth function $\mathcal{L}_\lambda^*$.

Lemma 5. *Under assumption 1, 2, for $\eta \leq \frac{1}{2L_*}$ it holds that*

$$\mathbf{E}_t \mathcal{L}_\lambda^*(\mathbf{w}_{t+1}) \leq \mathcal{L}_\lambda^*(\mathbf{w}_t) + \frac{\eta}{2} \mathbf{E}_t \|\mathbf{v}_{t+1} - \nabla \mathcal{L}_\lambda^*(\mathbf{w}_t)\|^2 - \frac{\eta}{4} \mathbf{E}_t \|\mathbf{v}_{t+1}\|^2 - \frac{\eta}{2} \|\nabla \mathcal{L}_\lambda^*(\mathbf{w}_t)\|^2. \tag{30}$$

proof of lemma 5. . The proof is quite standard. We present for completeness.

$$\begin{aligned}\mathbf{E}_t\mathcal{L}_\lambda^*(\mathbf{w}_{t+1}) &\leq \mathcal{L}_\lambda^*(\mathbf{w}_t) + \mathbf{E}_t\langle\nabla\mathcal{L}_\lambda^*(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \frac{L_*}{2}\mathbf{E}_t\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|^2 \\ &\leq \mathcal{L}_\lambda^*(\mathbf{w}_t) - \eta\mathbf{E}_t\langle\nabla\mathcal{L}_\lambda^*(\mathbf{w}_t), \mathbf{v}_{t+1}\rangle + \frac{\eta^2 L_*}{2}\mathbf{E}_t\left\|\mathbf{v}_{t+1}\right\|^2 \\ &\leq \mathcal{L}_\lambda^*(\mathbf{w}_t) + \frac{\eta}{2}\left(\mathbf{E}_t\left\|\mathbf{v}_{t+1} - \nabla\mathcal{L}_\lambda^*(\mathbf{w}_t)\right\|^2 - \mathbf{E}_t\left\|\mathbf{v}_{t+1}\right\|^2 - \left\|\nabla\mathcal{L}_\lambda^*(\mathbf{w}_t)\right\|^2\right) \\ &\quad + \frac{\eta^2 L_*}{2}\mathbf{E}_t\left\|\mathbf{v}_{t+1}\right\|^2 \\ &\leq \mathcal{L}_\lambda^*(\mathbf{w}_t) + \frac{\eta}{2}\mathbf{E}_t\left\|\mathbf{v}_{t+1} - \nabla\mathcal{L}_\lambda^*(\mathbf{w}_t)\right\|^2 - \frac{\eta}{4}\mathbf{E}_t\left\|\mathbf{v}_{t+1}\right\|^2 - \frac{\eta}{2}\left\|\nabla\mathcal{L}_\lambda^*(\mathbf{w}_t)\right\|^2 \end{aligned} \tag{31}$$

where the first inequality holds from the L_*-smoothness of $\mathcal{L}_\lambda^*(\cdot)$ and the last inequality holds from $\eta \leq \frac{1}{2L_*}$. □

We then provide a recursive bound for the error of the momentum estimator.

Lemma 6 (Lemma A.2 of Kwon et al. (2024))**.** *Under assumption 1, 2, it holds that:*

$$var(G_{t+1}^{\hat{w}}) \coloneqq \mathbf{E}\left\|\hat{G}_{t+1}^w - G_{t+1}^w\right\|^2 \leq \frac{1}{B}\left(2\sigma_f^2 + 8\lambda^2\sigma_g^2\right). \tag{32}$$

If we further assume that assumption 3 holds and let $r_\lambda = \frac{C_f}{\mu_g\lambda}$, we have that

$$var(G_{t+1}^{\hat{w}}) \coloneqq \mathbf{E}\left\|\hat{G}_{t+1}^w - G_{t+1}^w\right\|^2 \leq \frac{1}{B}\left(2\sigma_f^2 + 8\frac{\tilde{L}_g^2 C_f^2}{\mu_g^2}\right). \tag{33}$$

Lemma 7. *Under assumption 1, 2, for $\beta \leq 1$ it holds that*

$$\mathbf{E}_t V_{t+1} \leq (1-\beta)V_t + \frac{2L_*^2\eta^2}{\beta}\left\|\mathbf{v}_{t+1}\right\|^2 + 4\beta\left\|G_{t+1}^w - \nabla\mathcal{L}_\lambda^*(\mathbf{w}_{t+1})\right\|^2 + \frac{\beta^2\lambda^2\sigma^2}{B} \tag{34}$$

where $G_{t+1}^w = \mathbf{E}_t\hat{G}_{t+1}^w$ and $\lambda^2\sigma^2 = Bvar(\hat{G}_{t+1}^w) = 2\sigma_f^2 + 8\lambda^2\sigma_g^2$.

The following lemma decomposes the error of the (expected) hypergradient estimator.

Lemma 8 (lemma A.1(ii) of Kwon et al. (2024))**.** *Under assumption 1, let $r_\lambda = \frac{C_f}{\mu_g\lambda}$, it holds that*

$$\left\|G_t^w - \nabla\mathcal{L}_\lambda^*(\mathbf{w}_t)\right\|^2 \leq 3L_{y*}^2\left\|\mathbf{y}_{t+1} - \mathbf{y}_\lambda^*(\mathbf{w}_t)\right\|^2 + 3\lambda^2 L_g^2\left\|\mathbf{s}_{t+1} - \mathbf{s}^*(\mathbf{w}_t)\right\|^2 + \frac{3C_f^2 L_{y*}^2}{\mu_g^2\lambda^2} \tag{35}$$

where $L_{y*} = L_f + \frac{C_f Q_g}{\mu_g} = O(\bar{\ell}\kappa)$.

Remark: The original lemma in Kwon et al. (2024) requires (MSS) assumption 3. However, note that LHS doesn't involve any randomness/stochastic gradient(when $\mathbf{w}_t$ is fixed), we point out that this assumption is not needed. In fact, $\tilde{L}_g$ doesn't appear in the RHS.

proof of lemma 7. The proof is quite standard and we present for completeness.

$$\begin{aligned}
\mathbf{E}_t V_{t+1} &= \mathbf{E}_t \left\| \beta \hat{G}^w_{t+1} + (1-\beta)\mathbf{v}_{t+1} - \nabla \mathcal{L}^*_\lambda(\mathbf{w}_{t+1}) \right\|^2 \\
&= \mathbf{E}_t \| \beta(\hat{G}^w_{t+1} - G^w_{t+1}) + \beta(G^w_{t+1} - \nabla \mathcal{L}^*_\lambda(\mathbf{w}_{t+1})) \\
&\quad + (1-\beta)(\mathbf{v}_{t+1} - \nabla \mathcal{L}^*_\lambda(\mathbf{w}_t) + \nabla \mathcal{L}^*_\lambda(\mathbf{w}_t) - \nabla \mathcal{L}^*_\lambda(\mathbf{w}_{t+1})) \|^2 \\
&\leq \frac{\beta^2 \lambda^2 \sigma^2}{B} + (1-\beta)^2(1+\beta) \left\| \mathbf{v}_{t+1} - \nabla \mathcal{L}^*_\lambda(\mathbf{w}_t) \right\|^2 \\
&\quad + (1+\frac{1}{\beta}) \left(2\beta^2 \left\| G^w_{t+1} - \nabla \mathcal{L}^*_\lambda(\mathbf{w}_{t+1}) \right\|^2 + 2(1-\beta)^2 \left\| \nabla \mathcal{L}^*_\lambda(\mathbf{w}_t) - \nabla \mathcal{L}^*_\lambda(\mathbf{w}_{t+1}) \right\|^2 \right) \\
&\leq \frac{\beta^2 \lambda^2 \sigma^2}{B} + (1-\beta) \left\| \mathbf{v}_{t+1} - \nabla \mathcal{L}^*_\lambda(\mathbf{w}_t) \right\|^2 \\
&\quad + 4\beta \left\| G^w_{t+1} - \nabla \mathcal{L}^*_\lambda(\mathbf{w}_{t+1}) \right\|^2 + \frac{2}{\beta} \left\| \nabla \mathcal{L}^*_\lambda(\mathbf{w}_t) - \nabla \mathcal{L}^*_\lambda(\mathbf{w}_{t+1}) \right\|^2 \\
&\leq \frac{\beta^2 \lambda^2 \sigma^2}{B} + (1-\beta) \left\| \mathbf{v}_{t+1} - \nabla \mathcal{L}^*_\lambda(\mathbf{w}_t) \right\|^2 + 4\beta \left\| G^w_{t+1} - \nabla \mathcal{L}^*_\lambda(\mathbf{w}_{t+1}) \right\|^2 + \frac{2L_*^2 \eta^2}{\beta} \left\| \mathbf{v}_{t+1} \right\|^2
\end{aligned} \tag{36}$$

where the first inequality holds from 32 and young's inequality, the second inequality holds from $\beta \leq 1$ □

Lemma 9 (Descent lemma for $\mathbf{y}$)**.** *Under assumption 1, 2, for $\gamma \leq \frac{\mu_y}{L_y^2}$, we have*

$$\mathbf{E}_t Y_t \leq (1 - \frac{\gamma \mu_y}{2}) Y_{t-1} + \frac{2C_{y*}^2 \eta^2}{\gamma \mu_y} \left\| \mathbf{v}_t \right\|^2 + \frac{\gamma^2 \sigma_y^2}{B} \tag{37}$$

where $\sigma_y^2 = \frac{2\sigma_f^2}{\lambda^2} + 2\sigma_g^2 = O(1)$ and $C_{y*} = \frac{2(L_f + \lambda L_g)}{\lambda \mu_g} = O(\kappa)$.

proof of lemma 9.

$$\begin{aligned}
\mathbf{E}_t Y_t &= \mathbf{E}_t \left\| \mathbf{y}_t - \gamma \hat{G}^y_t - \mathbf{y}^*_{\lambda,t} \right\|^2 \\
&= \left\| \mathbf{y}_t - \mathbf{y}^*_{\lambda,t} \right\|^2 - 2\gamma \mathbf{E}_t \langle \hat{G}^y_t, \mathbf{y}_t - \mathbf{y}^*_{\lambda,t} \rangle + \gamma^2 \mathbf{E}_t \left\| \hat{G}^y_t \right\|^2 \\
&\leq \left\| \mathbf{y}_t - \mathbf{y}^*_{\lambda,t} \right\|^2 - 2\gamma\mu_y \left\| \mathbf{y}_t - \mathbf{y}^*_{\lambda,t} \right\|^2 + \gamma^2 \left(L_y^2 \left\| \mathbf{y}_t - \mathbf{y}^*_{\lambda,t} \right\|^2 + \frac{\sigma_y^2}{B} \right) \\
&\leq (1 - \gamma\mu_y) \left\| \mathbf{y}_t - \mathbf{y}^*_{\lambda,t} \right\|^2 + \frac{\gamma^2 \sigma_y^2}{B} \\
&\leq (1 - \gamma\mu_y)(1 + \frac{\gamma\mu_y}{2}) \left\| \mathbf{y}_t - \mathbf{y}^*_{\lambda,t-1} \right\|^2 + (1 - \gamma\mu_y)(1 + \frac{2}{\gamma\mu_y}) \left\| \mathbf{y}^*_{\lambda,t-1} - \mathbf{y}^*_{\lambda,t} \right\|^2 + \frac{\gamma^2 \sigma_y^2}{B} \\
&\leq (1 - \frac{\gamma\mu_y}{2}) \left\| \mathbf{y}_t - \mathbf{y}^*_{\lambda,t-1} \right\|^2 + \frac{2C_{y*}^2 \eta^2}{\gamma\mu_y} \left\| \mathbf{v}_t \right\|^2 + \frac{\gamma^2 \sigma_y^2}{B}
\end{aligned} \tag{38}$$

where $\mu_y = \mu_g - \frac{L_f}{\lambda}$ and $L_y = \frac{L_f}{\lambda} + L_g$ are the strongly convexity coefficient and the smoothness coefficient of $\mathcal{L}_y^\lambda(\mathbf{w}_t, \cdot) = \frac{1}{\lambda} f(\mathbf{w}_t, \cdot) + g(\mathbf{w}_t, \cdot)$. The last equality holds from $G_t^{y*} = \nabla_2 \mathcal{L}_y^\lambda(\mathbf{w}_t, \mathbf{y}^*_{\lambda,t}) = \mathbf{0}$ and lemma 3. The first inequality holds from the strongly convexity and smoothness of $\mathcal{L}_y^\lambda(\mathbf{w}_t, \cdot)$ and the second inequality holds from $\gamma \leq \frac{\mu_y}{L_y^2}$. □

Lemma 10. *Under assumption 1, we have that the following two inequality hold:*

$$\langle G_t^s, \mathbf{s}_t - \mathbf{s}_t^* \rangle \geq \frac{\mu_g}{2} \left\| \mathbf{s}_t - \mathbf{s}_t^* \right\|^2 - \frac{1}{\mu_g} \left(\frac{1}{\lambda^2} L_f^2 + Q_g^2 r_\lambda^2 \right) \left\| \mathbf{y}_t - \mathbf{y}_\lambda^*(\mathbf{w}_t) \right\|^2, \tag{39}$$

$$\left\| G_t^s \right\|^2 \leq 2L_g^2 \left\| \mathbf{s}_t - \mathbf{s}_t^* \right\|^2 + 2 \left(\frac{2}{\lambda^2} L_f^2 + 2Q_g^2 r_\lambda^2 \right) \left\| \mathbf{y}_t - \mathbf{y}_\lambda^*(\mathbf{w}_t) \right\|^2. \tag{40}$$

proof of lemma 10. We first show that the mapping $H(\mathbf{w},\mathbf{y},\mathbf{s}) \coloneqq \frac{1}{\lambda}\nabla_2 f(\mathbf{w},\mathbf{y}) + \nabla_2 g(\mathbf{w},\mathbf{y}) - \nabla_2 g(\mathbf{w},\mathbf{y}-\mathbf{s})$ is μ_g-strongly monotone and L_g lipschitz in $\mathbf{s}$, i.e.

$$\langle H(\mathbf{w},\mathbf{y},\mathbf{s}_1) - H(\mathbf{w},\mathbf{y},\mathbf{s}_2), \mathbf{s}_1-\mathbf{s}_2\rangle \geq \mu_g \|\mathbf{s}_1-\mathbf{s}_2\|^2, \tag{41}$$

$$\mu_g^2 \|\mathbf{s}_1-\mathbf{s}_2\|^2 \leq \|H(\mathbf{w},\mathbf{y},\mathbf{s}_1) - H(\mathbf{w},\mathbf{y},\mathbf{s}_2)\|^2 \leq L_g^2 \|\mathbf{s}_1-\mathbf{s}_2\|^2. \tag{42}$$

The proof is straightforward, we directly plug the definition of H in the LHS of 41:

$$\begin{aligned}\text{LHS} &= \langle \nabla_2 g(\mathbf{w},\mathbf{y}-\mathbf{s}_2) - \nabla_2 g(\mathbf{w},\mathbf{y}-\mathbf{s}_1), \mathbf{s}_1-\mathbf{s}_2\rangle \\ &= \langle \nabla_2 g(\mathbf{w},\mathbf{y}-\mathbf{s}_2) - \nabla_2 g(\mathbf{w},\mathbf{y}-\mathbf{s}_1), (\mathbf{y}-\mathbf{s}_2)-(\mathbf{y}-\mathbf{s}_1)\rangle \\ &\geq \mu_g \|\mathbf{s}_1-\mathbf{s}_2\|^2 \end{aligned} \tag{43}$$

where the inequality holds from the strongly convexity of g. Then the first inequality of 42 holds obviously by combining the fact that

$$\|H(\mathbf{w},\mathbf{y},\mathbf{s}_1) - H(\mathbf{w},\mathbf{y},\mathbf{s}_2)\| \|\mathbf{s}_1-\mathbf{s}_2\| \geq \langle H(\mathbf{w},\mathbf{y},\mathbf{s}_1) - H(\mathbf{w},\mathbf{y},\mathbf{s}_2), \mathbf{s}_1-\mathbf{s}_2\rangle$$

with 41 while the second inequality holds from smoothness of g.
We note that $G_t^s = H(\mathbf{w}_t,\mathbf{y}_t,\mathbf{s}_t)$ and that

$$\begin{aligned} H(\mathbf{w}_t,\mathbf{y}_\lambda^*(\mathbf{w}_t),\mathbf{s}_t^*) &= H(\mathbf{w}_t,\mathbf{y}_\lambda^*(\mathbf{w}_t),\mathbf{y}_\lambda^*(\mathbf{w}_t)-\mathbf{y}^*(\mathbf{w}_t)) \\ &= \frac{1}{\lambda}\nabla_2 f(\mathbf{w}_t,\mathbf{y}_\lambda^*(\mathbf{w}_t)) + \nabla_2 g(\mathbf{w}_t,\mathbf{y}_\lambda^*(\mathbf{w}_t)) - \nabla_2 g(\mathbf{w}_t,\mathbf{y}^*(\mathbf{w}_t)) \\ &= \mathbf{0}-\mathbf{0}=\mathbf{0}, \end{aligned}$$

then the cross term becomes:

$$\begin{aligned} \langle G_t^s, \mathbf{s}_t-\mathbf{s}_t^*\rangle &= \langle H(\mathbf{w}_t,\mathbf{y}_t,\mathbf{s}_t) - H(\mathbf{w}_t,\mathbf{y}_t,\mathbf{s}_t^*), \mathbf{s}_t-\mathbf{s}_t^*\rangle \\ &\quad + \langle H(\mathbf{w}_t,\mathbf{y}_t,\mathbf{s}_t^*) - H(\mathbf{w}_t,\mathbf{y}_\lambda^*(\mathbf{w}_t),\mathbf{s}_t^*), \mathbf{s}_t-\mathbf{s}_t^*\rangle \\ &\geq \mu_g \|\mathbf{s}_t-\mathbf{s}_t^*\|^2 + \langle H(\mathbf{w}_t,\mathbf{y}_t,\mathbf{s}_t^*) - H(\mathbf{w}_t,\mathbf{y}_\lambda^*(\mathbf{w}_t),\mathbf{s}_t^*), \mathbf{s}_t-\mathbf{s}_t^*\rangle \\ &\geq \mu_g \|\mathbf{s}_t-\mathbf{s}_t^*\|^2 - \frac{1}{2\mu_g}\|H(\mathbf{w}_t,\mathbf{y}_t,\mathbf{s}_t^*) - H(\mathbf{w}_t,\mathbf{y}_\lambda^*(\mathbf{w}_t),\mathbf{s}_t^*)\|^2 - \frac{\mu_g}{2}\|\mathbf{s}_t-\mathbf{s}_t^*\|^2 \\ &\geq \frac{\mu_g}{2}\|\mathbf{s}_t-\mathbf{s}_t^*\|^2 - \frac{1}{2\mu_g}\|H(\mathbf{w}_t,\mathbf{y}_t,\mathbf{s}_t^*) - H(\mathbf{w}_t,\mathbf{y}_\lambda^*(\mathbf{w}_t),\mathbf{s}_t^*)\|^2. \end{aligned} \tag{44}$$

For the second term in RHS, we note that $H(\mathbf{w},\mathbf{y},\mathbf{s}) = \frac{1}{\lambda}\nabla_2 f(\mathbf{w},\mathbf{y}) + \int_0^1 \nabla_{22}^2 g(\mathbf{w},\mathbf{y}-\tau\mathbf{s})\mathbf{s}d\tau$, then

$$\begin{aligned} \|H(\mathbf{w}_t,\mathbf{y}_t,\mathbf{s}_t^*) - H(\mathbf{w}_t,\mathbf{y}_\lambda^*(\mathbf{w}_t),\mathbf{s}_t^*)\|^2 &\leq \frac{2}{\lambda^2}L_f^2\|\mathbf{y}_t-\mathbf{y}_\lambda^*(\mathbf{w}_t)\|^2 \\ &\quad +2\|\mathbf{s}^*\|^2\left\|\int_0^1 \left(\nabla_{22}^2 g(\mathbf{w},\mathbf{y}_t-\tau\mathbf{s}) - \nabla_{22}^2 g(\mathbf{w},\mathbf{y}_\lambda^*(\mathbf{w}_t)-\tau\mathbf{s})\right)d\tau\right\|^2 \\ &\leq \frac{2}{\lambda^2}L_f^2\|\mathbf{y}_t-\mathbf{y}_\lambda^*(\mathbf{w}_t)\|^2 + 2Q_g^2\|\mathbf{s}\|^2\|\mathbf{y}_t-\mathbf{y}_\lambda^*(\mathbf{w}_t)\|^2 \\ &\leq \left(\frac{2}{\lambda^2}L_f^2 + 2Q_g^2 r_\lambda^2\right)\|\mathbf{y}_t-\mathbf{y}_\lambda^*(\mathbf{w}_t)\|^2. \end{aligned} \tag{45}$$

Plugging 45 into 44 leads to 39. The 40 can be proved in a similar way:

$$\begin{aligned} \|G_t^s\|^2 &\leq 2\|H(\mathbf{w}_t,\mathbf{y}_t,\mathbf{s}_t) - H(\mathbf{w}_t,\mathbf{y}_t,\mathbf{s}_t^*)\|^2 + 2\|H(\mathbf{w}_t,\mathbf{y}_t,\mathbf{s}_t^*)\|^2 \\ &\leq 2L_g^2\|\mathbf{s}_t-\mathbf{s}_t^*\|^2 + 2\left(\frac{2}{\lambda^2}L_f^2 + 2Q_g^2 r_\lambda^2\right)\|\mathbf{y}_t-\mathbf{y}_\lambda^*(\mathbf{w}_t)\|^2. \end{aligned}$$

The second inequality holds from 42 and 45. □

Proposition 1 (Proposition 3.4 from Kwon et al. (2024))**.** *Under assumption 1, for any t, with $r_\lambda = \frac{C_f}{\lambda\mu_g}$,*

$$\|\mathbf{s}_{t+1}-\mathbf{s}_t^*\| \leq \|\bar{\mathbf{s}}_t-\mathbf{s}_t^*\|. \tag{46}$$

Lemma 11 (Descent lemma for $\mathbf{s}$). *Under assumption 1, 2, let* $\gamma \leq \min\{\frac{\mu_g}{4L_g^2}, \frac{1}{2\mu_g}\}$, *then it holds that*

$$\begin{aligned}\mathbf{E}_t S_t &\leq \mathbf{E}_t \left\|\bar{\mathbf{s}}_t - \mathbf{s}_t^*\right\|^2 \\ &\leq \left(1 - \frac{\gamma\mu_g}{4}\right) S_{t-1} + \frac{4}{\gamma\mu_g}\left\|\mathbf{s}_t^* - \mathbf{s}_{t-1}^*\right\|^2 + \frac{8\gamma}{\mu_g}\left(\frac{L_f^2}{\lambda^2} + Q_g^2 r_\lambda^2\right)\left(Y_{t-1} + \eta^2 C_{y*}^2 \left\|\mathbf{v}_t\right\|^2\right) \\ &\quad + \frac{2\gamma^2}{B}(\sigma_y^2 + \sigma_g^2). \end{aligned} \tag{47}$$

proof of lemma 11. The first inequality holds from proposition 1. We now prove the second inequality.

$$\begin{aligned}\mathbf{E}_t \left\|\bar{\mathbf{s}}_t - \mathbf{s}_t^*\right\|^2 &= \mathbf{E}_t \left\|\mathbf{s}_t - \gamma\hat{G}_t^s - \mathbf{s}_t^*\right\|^2 \\ &= \mathbf{E}_t \left\|\mathbf{s}_t - \mathbf{s}_t^*\right\|^2 + \gamma^2(\mathrm{var}(\hat{G}_t^s) + \mathbf{E}_t \left\|G_t^s\right\|^2) - 2\gamma\langle G_t^s, \mathbf{s}_t - \mathbf{s}_t^*\rangle \\ &\leq \mathbf{E}_t \left\|\mathbf{s}_t - \mathbf{s}_t^*\right\|^2 + \gamma^2\left(\frac{2\sigma_y^2 + 2\sigma_g^2}{B} + 2L_g^2\left\|\mathbf{s}_t - \mathbf{s}_t^*\right\|^2\right) \\ &\quad + 2\gamma^2\left(\frac{2}{\lambda^2}L_f^2 + 2Q_g^2 r_\lambda^2\right)\left\|\mathbf{y}_t - \mathbf{y}_\lambda^*(\mathbf{w}_t)\right\|^2 \\ &\quad - \gamma\mu_g\left\|\mathbf{s}_t - \mathbf{s}_t^*\right\|^2 + \frac{2\gamma}{\mu_g}\left(\frac{1}{\lambda^2}L_f^2 + Q_g^2 r_\lambda^2\right)\left\|\mathbf{y}_t - \mathbf{y}_\lambda^*(\mathbf{w}_t)\right\|^2 \\ &\leq (1 - \frac{\gamma\mu_g}{2})\mathbf{E}_t\left\|\mathbf{s}_t - \mathbf{s}_t^*\right\|^2 + \frac{2\gamma^2}{B}(\sigma_y^2 + \sigma_g^2) \\ &\quad + \frac{4\gamma(1 + 2\gamma\mu_g)}{\mu_g}\left(\frac{1}{\lambda^2}L_f^2 + Q_g^2 r_\lambda^2\right)\left(\left\|\mathbf{y}_t - \mathbf{y}_\lambda^*(\mathbf{w}_{t-1})\right\|^2 + \left\|\mathbf{y}_\lambda^*(\mathbf{w}_{t-1}) - \mathbf{y}_\lambda^*(\mathbf{w}_t)\right\|^2\right) \\ &\leq (1 - \frac{\gamma\mu_g}{4})\mathbf{E}_t\left\|\mathbf{s}_t - \mathbf{s}_{t-1}^*\right\|^2 + \frac{4}{\gamma\mu_g}\left\|\mathbf{s}_t^* - \mathbf{s}_{t-1}^*\right\|^2 + \frac{2\gamma^2}{B}(\sigma_y^2 + \sigma_g^2) \\ &\quad + \frac{8\gamma}{\mu_g}\left(\frac{1}{\lambda^2}L_f^2 + Q_g^2 r_\lambda^2\right)\left(Y_{t-1} + \eta^2 C_{y*}^2\left\|\mathbf{v}_t\right\|^2\right) \end{aligned} \tag{48}$$

where $\hat{G}_t^s = \hat{G}_t^y - \hat{G}_t^z$, the last inequality holds from Young's inequality. And the blue part is the newly introduced term according to the corrected analysis of cross term presented in lemma 10. □

A.3 Lemma for Theorem 2

Below we provide better bound with the extra assumption 3 for recursion of V_t and S_t.
Note that the proof given the extra mean-square smoothness assumption is quite similar to the corresponding lemma in the previous section, we only highlight the difference instead of provide the complete proof.

Lemma 12 (Recursive bound for momentum with MSS). *Under assumption 1, 2 and 3, for* $\beta \leq 1$ *it holds that*

$$\mathbf{E}_t V_{t+1} \leq (1 - \beta)V_t + \frac{2L_*^2\eta^2}{\beta}\left\|\mathbf{v}_{t+1}\right\|^2 + 4\beta\left\|G_{t+1}^w - \nabla\mathcal{L}_\lambda^*(\mathbf{w}_{t+1})\right\|^2 + \frac{\beta^2\tilde{\sigma}^2}{B} \tag{49}$$

where $G_{t+1}^w = \mathbf{E}_t\hat{G}_{t+1}^w$ *and* $\tilde{\sigma}^2 = B\mathit{var}(\hat{G}_{t+1}^w) = 2\sigma_f^2 + 8\frac{\tilde{L}_g^2 C_f^2}{\mu_g^2} = O(\bar{\ell}^2\kappa^2)$.

Proof. The proof is almost the same as that of lemma 7 except using 33 instead of 32 to bound the variance term of $\hat{G}_t^w$. □

Lemma 13 (Descent lemma for $\mathbf{s}$ with MSS)**.** *Suppose that assumption 1, 2 and 3 hold. Let* $\gamma \leq \min\{\frac{\mu_g}{4L_g^2}, \frac{1}{2\mu_g}\}$*, it holds that*

$$\begin{aligned}\mathbf{E}_t S_t &\leq \mathbf{E}_t \left\|\bar{\mathbf{s}}_t - \mathbf{s}_t^*\right\|^2 \\ &\leq \left(1 - \frac{\gamma\mu_g}{4}\right) S_{t-1} + \frac{4}{\gamma\mu_g}\left\|\mathbf{s}_t^* - \mathbf{s}_{t-1}^*\right\|^2 + \frac{8\gamma}{\mu_g}\left(\frac{L_f^2}{\lambda^2} + Q_g^2 r_\lambda^2\right)\left(Y_{t-1} + \eta^2 C_{y*}^2 \left\|\mathbf{v}_t\right\|^2\right) \\ &\quad + \frac{2\gamma^2}{B\lambda^2}(\sigma_f^2 + \frac{4C_f^2\tilde{L}_g^2}{\mu_g^2}). \end{aligned} \tag{50}$$

Proof. The main difference to the proof of lemma 11 is that we leverage a better bound for the variance term of $\hat{G}_t^s$ given the MSS assumption:

$$\begin{aligned}\text{var}(\hat{G}_t^s) &= \mathbf{E}_t \left\|\hat{G}_t^s - G_t^s\right\|^2 \\ &\leq \frac{1}{B}\left(\frac{2}{\lambda^2}\mathbf{E}_t \left\|\nabla_2 f(\mathbf{w}_t, \mathbf{y}_t, \xi_{fy}^t) - \nabla_2 f(\mathbf{w}_t, \mathbf{y}_t)\right\|^2\right. \\ &\quad + 4\mathbf{E}_t \left\|\nabla_2 g(\mathbf{w}_t, \mathbf{y}_t, \xi_{gy}^t) - \nabla_2 g(\mathbf{w}_t, \mathbf{z}_t, \xi_{gy}^t)\right\|^2 \\ &\quad \left. + 4\left\|\nabla_2 g(\mathbf{w}_t, \mathbf{y}_t) - \nabla_2 g(\mathbf{w}_t, \mathbf{z}_t)\right\|^2\right) \\ &\leq \frac{1}{B}\left(\frac{2\sigma_f^2}{\lambda^2} + 8\tilde{L}_g^2\left\|\mathbf{y}_t - \mathbf{z}_t\right\|^2\right) \\ &\leq \frac{1}{B}\left(\frac{2\sigma_f^2}{\lambda^2} + 8\frac{C_f^2\tilde{L}_g^2}{\mu_g^2\lambda^2}\right) \end{aligned} \tag{51}$$

where ξ_{fy}^t, ξ_{gy}^t are randomly sample from $\mathcal{B}_{fy}^t, \mathcal{B}_{gy}^t$. For the following derivation, we just need to follow inequality 48 but replace with the above bound for the variance term. □

B Proof of Theorem 1

Proof of theorem 1. By plugging 35 into 34 and plugging 29 into 47 we have:

$$\begin{aligned}\mathbf{E}_t V_{t+1} &\leq (1-\beta)V_t + \frac{2L_*^2\eta^2}{\beta}\left\|\mathbf{v}_{t+1}\right\|^2 + 12\beta L_{y*}^2\left\|\mathbf{y}_{t+2} - \mathbf{y}_\lambda^*(\mathbf{w}_{t+1})\right\|^2 \\ &\quad + 12\beta\lambda^2 L_g^2\left\|\mathbf{s}_{t+2} - \mathbf{s}^*(\mathbf{w}_{t+1})\right\|^2 + \frac{12\beta C_f^2 L_{y*}^2}{\mu_g^2\lambda^2} + \frac{\beta^2\lambda^2\sigma^2}{B} \\ &= (1-\beta)V_t + \frac{2L_*^2\eta^2}{\beta}\left\|\mathbf{v}_{t+1}\right\|^2 + 12\beta L_{y*}^2 Y_{t+1} + 12\beta\lambda^2 L_g^2 S_{t+1} + \frac{12\beta C_f^2 L_{y*}^2}{\mu_g^2\lambda^2} + \frac{\beta^2\lambda^2\sigma^2}{B}, \end{aligned} \tag{52}$$

$$\begin{aligned}\mathbf{E}_t S_t &\leq \mathbf{E}_t \left\|\bar{\mathbf{s}}_t - \mathbf{s}_t^*\right\|^2 \\ &\leq \left(1 - \frac{\gamma\mu_g}{4}\right) S_{t-1} + \frac{4C_s^2\eta^2}{\lambda^2\gamma\mu_g}\left\|\mathbf{v}_t\right\|^2 + \frac{8\gamma}{\mu_g}\left(\frac{L_f^2}{\lambda^2} + Q_g^2 r_\lambda^2\right)\left(Y_{t-1} + \eta^2 C_{y*}^2\left\|\mathbf{v}_t\right\|^2\right) \\ &\quad + \frac{2\gamma^2}{B}(\sigma_y^2 + \sigma_g^2). \end{aligned} \tag{53}$$

Combining 30, 52, 37 and 53, then summing over $t = 0, 1, \cdots, T-1$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\left\|\nabla\mathcal{L}_\lambda^*(\mathbf{w}_t)\right\|^2 \le{}& \frac{2(\mathcal{L}_\lambda^*(\mathbf{w}_0)-\inf_{\mathbf{w}}\mathcal{L}_\lambda^*(\mathbf{w}))}{\eta T}+\frac{\mathbf{E}V_0}{\beta T}\\ &+\frac{24\left((L_{y*}^2+32\kappa^2(L_f^2+\frac{Q_g^2C_f^2}{\mu_g^2}))\mathbf{E}Y_1+2\lambda^2L_g^2\mathbf{E}S_1\right)}{\gamma\mu_g T}\\ &-\left(\frac{1}{2}-\frac{2L_*^2\eta^2}{\beta^2}-\frac{48\eta^2\left(C_{y*}^2(L_{y*}^2+32\kappa^2(L_f^2+\frac{Q_g^2C_f^2}{\mu_g^2}))+4L_g^2C_s^2\right)}{\gamma^2\mu_g^2}\right.\\ &\left.-384\eta^2\kappa^2C_{y*}^2(L_f^2+\frac{Q_g^2C_f^2}{\mu_g^2})\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\left\|\mathbf{v}_{t+1}\right\|^2\\ &+\frac{\beta\lambda^2\sigma^2}{B}+\frac{24\gamma\sigma_y^2}{\mu_g B}\left(L_{y*}^2+32\kappa^2(L_f^2+\frac{Q_g^2C_f^2}{\mu_g^2})\right)+\frac{12C_f^2L_{y*}^2}{\mu_g^2\lambda^2}\\ &+\frac{96\gamma\lambda^2L_g^2}{\mu_g B}(\sigma_y^2+\sigma_g^2).\end{aligned}\tag{54}$$

Denote $\Delta = \mathcal{L}_\lambda^*(\mathbf{w}_0) - \inf_{\mathbf{w}}\mathcal{L}_\lambda^*(\mathbf{w})$. With lemma 1 we have that

$$\Delta \le F(\mathbf{w}_0) - \inf_{\mathbf{w}} F(\mathbf{w}) + O(\frac{\bar{\ell}\kappa^2}{\lambda}) \le O(1)$$

Note that $\mathbf{v}_1 = \beta\hat{G}_0^w = O(\beta\lambda)$, we have:

$$\mathbf{E}V_0 \le 2\mathbf{E}\left\|\mathbf{v}_1\right\|^2 + 2\left\|\nabla\mathcal{L}_\lambda^*(\mathbf{w}_0)\right\|^2 \le O(\beta^2\lambda^2) + 4L_*\Delta \le O(\beta^2\lambda^2+\kappa^3).$$

With eqn. 37 we have that

$$\mathbf{E}Y_1 \le (1-\frac{\gamma\mu_y}{2})Y_0 + \frac{2C_{y*}^2\eta^2}{\gamma\mu_y}\left\|\mathbf{v}_1\right\|^2 + \frac{\gamma^2\sigma_y^2}{B}.$$

According the setting of β, γ and λ that we show later, the second and the last term will be dominated by the first term. Therefore, we have $\mathbf{E}Y_1 = O(Y_0)$. Besides,

$$Y_0 = \left\|\mathbf{y}_1 - \mathbf{y}_\lambda^*(\mathbf{w}_0)\right\|^2 = \left\|\mathbf{y}_0 - \gamma\hat{G}_t^y - \mathbf{y}_\lambda^*(\mathbf{w}_0)\right\|^2 \le 2\left\|\mathbf{y}_0 - \mathbf{y}_\lambda^*(\mathbf{w}_0)\right\|^2 + 2\gamma^2\left\|\hat{G}_0^y\right\|^2 \le O(\frac{1}{\kappa}),$$

where the last inequality holds from the warm up stage for $\mathbf{y}$. Combining the 2 inequalities above yields that $\mathbf{E}Y_1 \le O(\frac{1}{\kappa})$. In addition, $\mathbf{E}S_1 \le 2\mathbf{E}\left\|\mathbf{s}_2\right\|^2 + 2\mathbf{E}\left\|\mathbf{s}^*(\mathbf{w}_1)\right\|^2 \le O(r_\lambda^2) = O(\frac{\kappa^2}{\lambda^2})$.

To make RHS of eqn. 54$\le \epsilon^2$ with constant level batchsize $B = O(1)$, we require the following condition holds:

$$\beta\lambda^2\sigma^2 \le O(\epsilon^2), \gamma\bar{\ell}\kappa^5 \le O(\epsilon^2), \frac{\bar{\ell}^2\kappa^4}{\lambda^2} \le O(\epsilon^2), \bar{\ell}\kappa\gamma\lambda^2 \le O(\epsilon^2),\tag{55}$$

$$\max\left\{\frac{L_*^2\eta^2}{\beta^2}, \frac{\eta^2O(\kappa^8)}{\gamma^2}, \eta^2\bar{\ell}^2\kappa^6\right\} \le O(1),\tag{56}$$

$$\frac{\Delta}{\eta T} \le O(\epsilon^2), \frac{V_0}{\beta T} \le O(\epsilon^2), \frac{O(\bar{\ell}\kappa^5Y_1 + \lambda^2\bar{\ell}\kappa S_1)}{\gamma T} \le O(\epsilon^2).\tag{57}$$

Therefore, by setting $\lambda = \Theta(\frac{\kappa^3}{\epsilon}), \beta = \Theta(\frac{\epsilon^4}{\kappa^6}), \gamma = \Theta(\frac{\epsilon^4}{\kappa^7}), \eta = \Theta(\frac{\epsilon^4}{\kappa^{11}})$, it's sufficient for algorithm 1 to return an $O(\epsilon)$-stationary point of $\mathcal{L}_\lambda^*(\mathbf{w})$ within $T = O(\frac{\kappa^{11}}{\epsilon^6})$ iterations. Then according to lemma 1, it is also an $O(\epsilon)$-stationary point of $F(\mathbf{w})$. □

C Proof of Theorem 2

The proof of theorem 2 is quite similar to that of theorem 1 except that we need to leverage better recursion lemma of $\mathbf{v}$ and $\mathbf{s}$.

Proof of theorem 2. By plugging 35 into 49 and plugging 29 into 50, we have

$$\begin{aligned}\mathbf{E}_t V_{t+1} &\leq (1-\beta)V_t + \frac{2L_*^2\eta^2}{\beta}\left\|\mathbf{v}_{t+1}\right\|^2 + 12\beta L_{y*}^2\left\|\mathbf{y}_{t+2} - \mathbf{y}_\lambda^*(\mathbf{w}_{t+1})\right\|^2 \\ &\quad + 12\beta\lambda^2 L_g^2\left\|\mathbf{s}_{t+2} - \mathbf{s}^*(\mathbf{w}_{t+1})\right\|^2 + \frac{12\beta C_f^2 L_{y*}^2}{\mu_g^2\lambda^2} + \frac{\beta^2\tilde{\sigma}^2}{B} \\ &= (1-\beta)V_t + \frac{2L_*^2\eta^2}{\beta}\left\|\mathbf{v}_{t+1}\right\|^2 + 12\beta L_{y*}^2 Y_{t+1} + 12\beta\lambda^2 L_g^2 S_{t+1} + \frac{12\beta C_f^2 L_{y*}^2}{\mu_g^2\lambda^2} + \frac{\beta^2\tilde{\sigma}^2}{B},\end{aligned} \tag{58}$$

$$\begin{aligned}\mathbf{E}_t S_t &\leq \mathbf{E}_t\left\|\bar{\mathbf{s}}_t - \mathbf{s}_t^*\right\|^2 \\ &\leq \left(1 - \frac{\gamma\mu_g}{4}\right)S_{t-1} + \frac{4C_s^2\eta^2}{\lambda^2\gamma\mu_g}\left\|\mathbf{v}_t\right\|^2 + \frac{8\gamma}{\mu_g}\left(\frac{L_f^2}{\lambda^2} + Q_g^2 r_\lambda^2\right)\left(Y_{t-1} + \eta^2 C_{y*}^2\left\|\mathbf{v}_t\right\|^2\right) \\ &\quad + \frac{2\gamma^2}{B\lambda^2}(\sigma_f^2 + \frac{4C_f^2\tilde{L}_g^2}{\mu_g^2}).\end{aligned} \tag{59}$$

Combine 30, 58, 37 and 59, then sum over $t = 0, 1, \cdots, T-1$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\left\|\nabla\mathcal{L}_\lambda^*(\mathbf{w}_t)\right\|^2 &\leq \frac{2(\mathcal{L}_\lambda^*(\mathbf{w}_0) - \inf_{\mathbf{w}}\mathcal{L}_\lambda^*(\mathbf{w}))}{\eta T} + \frac{\mathbf{E}V_0}{\beta T} \\ &\quad + \frac{24\left((L_{y*}^2 + 32\kappa^2(L_f^2 + \frac{Q_g^2C_f^2}{\mu_g^2}))\mathbf{E}Y_1 + 2\lambda^2 L_g^2 S_1\right)}{\gamma\mu_g T} \\ &\quad - \left(\frac{1}{2} - \frac{2L_*^2\eta^2}{\beta^2} - \frac{48\eta^2\left(C_{y*}^2(L_{y*}^2 + 32\kappa^2(L_f^2 + \frac{Q_g^2C_f^2}{\mu_g^2})) + 4L_g^2C_s^2\right)}{\gamma^2\mu_g^2}\right. \\ &\quad \left. -384\eta^2\kappa^2 C_{y*}^2(L_f^2 + \frac{Q_g^2C_f^2}{\mu_g^2})\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\left\|\mathbf{v}_{t+1}\right\|^2 \\ &\quad + \frac{\beta\tilde{\sigma}^2}{B} + \frac{24\gamma\sigma_y^2}{\mu_g B}\left(L_{y*}^2 + 32\kappa^2(L_f^2 + \frac{Q_g^2C_f^2}{\mu_g^2})\right) + \frac{12C_f^2L_{y*}^2}{\mu_g^2\lambda^2} \\ &\quad + \frac{96\gamma L_g^2}{\mu_g B}(\sigma_f^2 + \frac{4C_f^2\tilde{L}_g^2}{\mu_g^2}).\end{aligned} \tag{60}$$

As shown in proof of theorem 1, we have that:

$$\mathcal{L}_\lambda^*(\mathbf{w}_0) - \inf_{\mathbf{w}}\mathcal{L}_\lambda^*(\mathbf{w}) \leq O(1), \mathbf{E}V_0 \leq O(\kappa^3), \mathbf{E}Y_1 \leq O(\frac{1}{\kappa}), \mathbf{E}S_1 \leq \frac{\kappa^2}{\lambda^2}$$

Therefore, to make RHS of eqn. 60$\leq \epsilon^2$ with constant level batchsize $B = O(1)$, and note that $\tilde{\sigma}^2 = O(\bar{\ell}^2\kappa^2)$, we require the following condition holds:

$$\beta\bar{\ell}^2\kappa^2 \leq O(\epsilon^2), \gamma\bar{\ell}\kappa^5 \leq O(\epsilon^2), \frac{\bar{\ell}^2\kappa^4}{\lambda^2} \leq O(\epsilon^2), \bar{\ell}^3\kappa^3\gamma \leq O(\epsilon^2), \tag{61}$$

$$\max\left\{\frac{L_*^2\eta^2}{\beta^2}, \frac{\eta^2 O(\kappa^8)}{\gamma^2}, \eta^2\bar{\ell}^2\kappa^6\right\} \leq O(1), \tag{62}$$

$$\frac{\Delta}{\eta T} \leq O(\epsilon^2), \frac{V_0}{\beta T} \leq O(\epsilon^2), \frac{O(\bar{\ell}\kappa^5 Y_1 + \lambda^2\bar{\ell}\kappa S_1)}{\gamma T} \leq O(\epsilon^2). \tag{63}$$

Therefore, by setting $\lambda = \Theta(\frac{\kappa^3}{\epsilon}), \beta = \Theta(\frac{\epsilon^2}{\kappa^2}), \gamma = \Theta(\frac{\epsilon^2}{\kappa^5}), \eta = \Theta(\frac{\epsilon^2}{\kappa^9})$, it's sufficient for algorithm 1 to return an $O(\epsilon)$-stationary point of $\mathcal{L}_\lambda^*(\mathbf{w})$ within $T = O(\frac{\kappa^9}{\epsilon^4})$ iterations. Then according to lemma 1, it is also an $O(\epsilon)$-stationary point of $F(\mathbf{w})$. □

D COUNTER EXAMPLE OF INEQUALITY 24

We provide a counterexample showing that even under Assumptions 1– 3, the inequality

$$\langle G_t^s, s_t - s_t^* \rangle \geq c\|s_t - s_t^*\|^2$$

need not hold for any positive constant c. Consider a scalar lower-level variable $y \in \mathbb{R}$ and define

$$f(w, y) = \cos y - 1, \qquad g(w, y) = \frac{1}{2}y^2,$$

where both functions are independent of w. These functions are infinitely differentiable, f is globally 1-Lipschitz and 1-smooth, and g is 1-smooth and 1-strongly convex in y, with a constant Hessian. For any $\lambda \geq 2$, the penalized lower-level objective satisfies

$$\frac{\partial^2}{\partial y^2}\left(\lambda^{-1} f(w, y) + g(w, y)\right) = 1 - \lambda^{-1}\cos y \geq \frac{1}{2}.$$

Since its first derivative vanishes at $y = 0$, we have

$$y_\lambda^*(w) = y^*(w) = 0, \qquad s^*(w) = 0, \qquad F(w) = 0.$$

Thus, Assumption 1 holds. Moreover, taking exact gradients as the stochastic oracles satisfies Assumption 2 with zero variance and Assumption 3 with $\widetilde{L}_g = 1$.

Now choose any $y_t \in (0, 1]$ and set

$$z_t = y_t - \frac{\sin y_t}{\lambda}.$$

Then

$$s_t = y_t - z_t = \frac{\sin y_t}{\lambda} > 0, \qquad |s_t| \leq \frac{1}{\lambda} = \frac{C_f}{\mu_g \lambda} = r_\lambda,$$

so these iterates satisfy the projected-coupling constraint. Nevertheless, the corresponding gap-update direction is

$$\begin{aligned} G_t^s &= \lambda^{-1}\nabla_2 f(w_t, y_t) + \nabla_2 g(w_t, y_t) - \nabla_2 g(w_t, z_t) \\ &= -\frac{\sin y_t}{\lambda} + y_t - z_t = 0. \end{aligned}$$

Consequently, for every $c > 0$,

$$\langle G_t^s, s_t - s_t^* \rangle = 0 < c\,\frac{\sin^2 y_t}{\lambda^2} = c\|s_t - s_t^*\|^2,$$

which contradicts the proposed inequality.

E MORE EXPERIMENT DETAIL

We use linear schedule for λ of F2SA/F3SA which start from λ_0 and tune the final λ; We further tune β for our method SICO in $\{0.01, 0.1, 0.3, 0.5, 1\}$. For stoc-SGHA we further tune τ in $\{10^{-4}, 10^{-3}\}$; p_w in $\{0.005, 0.01, 0.02\}$; p_λ in $\{0.00001, 0.0001\}$ and β in $\{0.16, 0.24, 0.32\}$; we fix $C_\lambda = 10$.