Enhanced Bilevel Optimization via Bregman Distance

Feihu Huang[1,2], **Junyi Li**[1], **Shangqian Gao**[1], **Heng Huang**[1]
[1]Electrical & Computer Engineering, University of Pittsburgh, Pittsburgh, PA, United States
[2]College of Computer Science & Technology, Nanjing University of Aeronautics & Astronautics, Nanjing, China
huangfeihu2018@gmail.com, junyili.ai@gmail.com, shg84@pitt.edu, heng.huang@pitt.edu

Abstract

Bilevel optimization has been recently used in many machine learning problems such as hyperparameter optimization, policy optimization, and meta learning. Although many bilevel optimization methods have been proposed, they still suffer from the high computational complexities and do not consider the more general bilevel problems with nonsmooth regularization. In the paper, thus, we propose a class of enhanced bilevel optimization methods with using Bregman distance to solve bilevel optimization problems, where the outer subproblem is nonconvex and possibly nonsmooth, and the inner subproblem is strongly convex. Specifically, we propose a bilevel optimization method based on Bregman distance (BiO-BreD) to solve deterministic bilevel problems, which achieves a lower computational complexity than the best known results. Meanwhile, we also propose a stochastic bilevel optimization method (SBiO-BreD) to solve stochastic bilevel problems based on stochastic approximated gradients and Bregman distance. Moreover, we further propose an accelerated version of SBiO-BreD method (ASBiO-BreD) using the variance-reduced technique, which can achieve a lower computational complexity than the best known computational complexities with respect to condition number κ and target accuracy ϵ for finding an ϵ-stationary point. We conduct data hyper-cleaning task and hyper-representation learning task to demonstrate that our new algorithms outperform related bilevel optimization approaches.

1 Introduction

Bilevel optimization can effectively solve the problems with a hierarchical structure, thus it recently has been widely used in many machine learning tasks such as hyper-parameter optimization [37, 20, 9, 38], meta learning [9, 31, 22], neural network architecture search [30], reinforcement learning [15], and image processing [31]. In the paper, we consider solving the following nonconvex-strongly-convex bilevel optimization problem:

$$\begin{aligned} &\min_{x\in\mathcal{X}\subseteq\mathbb{R}^{d_1}} f(x, y^*(x)) + h(x), && \text{(Outer)} \\ &\text{s.t. } y^*(x) \in \arg\min_{y\in\mathbb{R}^{d_2}} g(x, y), && \text{(Inner)} \end{aligned} \tag{1}$$

where function $F(x) = f(x, y^*(x)) : \mathcal{X} \to \mathbb{R}$ is smooth and possibly nonconvex, and function $h(x)$ is convex and possibly nonsmooth, and function $g(x, y) : \mathcal{X} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is μ-strongly convex in $y \in \mathbb{R}^{d_2}$. The constraint set $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ is compact and convex. Problem (1) covers a rich class of nonconvex objective functions with nonsmooth regularization, and is more general than the existing nonconvex bilevel optimization formulation in [11, 22] that does not consider any nonsmooth regularization. Here the function $h(x)$ can be the nonsmooth regularization term such as $h(x) = \lambda\|x\|_1$.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 1: Comparisons of the representative bilevel optimization algorithms for finding an ϵ-stationary point of the **deterministic** nonconvex-strongly-convex Problem (1) with $h(x)$ or without $h(x)$, *i.e.*, $\|\nabla F(x)\|^2 \leq \epsilon$ or its equivalent variants. $Gc(f,\epsilon)$ and $Gc(g,\epsilon)$ denote the number of gradient evaluations *w.r.t.* $f(x,y)$ and $g(x,y)$; $JV(g,\epsilon)$ denotes the number of Jacobian-vector products; $HV(g,\epsilon)$ is the number of Hessian-vector products; $\kappa = L/\mu$ is the conditional number. $\surd$ means that the algorithms solve both the **smooth** and **nonsmooth** bilevel optimizations.

Algorithm	**Reference**	$Gc(f,\epsilon)$	$Gc(g,\epsilon)$	$JV(g,\epsilon)$	$HV(g,\epsilon)$	**Nonsmooth**
AID-BiO	[11]	$O(\kappa^4\epsilon^{-1})$	$O(\kappa^5\epsilon^{-5/4})$	$O(\kappa^4\epsilon^{-1})$	$O(\kappa^{4.5}\epsilon^{-1})$	
AID-BiO	[22]	$O(\kappa^3\epsilon^{-1})$	$O(\kappa^4\epsilon^{-1})$	$O(\kappa^3\epsilon^{-1})$	$O(\kappa^{3.5}\epsilon^{-1})$	
ITD-BiO	[22]	$O(\kappa^3\epsilon^{-1})$	$\tilde{O}(\kappa^4\epsilon^{-1})$	$\tilde{O}(\kappa^4\epsilon^{-1})$	$\tilde{O}(\kappa^4\epsilon^{-1})$	
BiO-BreD	Ours	$O(\kappa^2\epsilon^{-1})$	$\tilde{O}(\kappa^3\epsilon^{-1})$	$\tilde{O}(\kappa^3\epsilon^{-1})$	$\tilde{O}(\kappa^3\epsilon^{-1})$	$\surd$

Table 2: Comparisons of the representative bilevel optimization algorithms for finding an ϵ-stationary point of the **stochastic** nonconvex-strongly-convex problem (2) with $h(x)$ or without $h(x)$, *i.e.*, $\mathbb{E}\|\nabla F(x)\|^2 \leq \epsilon$ or its equivalent variants. Since some algorithms do not provide the explicit dependence on κ, we use $p(\kappa)$.

Algorithm	**Reference**	$Gc(f,\epsilon)$	$Gc(g,\epsilon)$	$JV(g,\epsilon)$	$HV(g,\epsilon)$	**Nonsmooth**
TTSA	[15]	$O(p(\kappa)\epsilon^{-2.5})$	$O(p(\kappa)\epsilon^{-2.5})$	$O(p(\kappa)\epsilon^{-2.5})$	$O(p(\kappa)\epsilon^{-2.5})$	
STABLE	[5]	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	
SMB	[13]	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	
VRBO	[41]	$O(p(\kappa)\epsilon^{-1.5})$	$O(p(\kappa)\epsilon^{-1.5})$	$O(p(\kappa)\epsilon^{-1.5})$	$O(p(\kappa)\epsilon^{-1.5})$	
SUSTAIN	[23]	$O(p(\kappa)\epsilon^{-1.5})$	$O(p(\kappa)\epsilon^{-1.5})$	$O(p(\kappa)\epsilon^{-1.5})$	$O(p(\kappa)\epsilon^{-1.5})$	
VR-saBiAdam	[18]	$O(p(\kappa)\epsilon^{-1.5})$	$O(p(\kappa)\epsilon^{-1.5})$	$O(p(\kappa)\epsilon^{-1.5})$	$O(p(\kappa)\epsilon^{-1.5})$	
BSA	[11]	$O(\kappa^6\epsilon^{-2})$	$O(\kappa^9\epsilon^{-3})$	$O(\kappa^6\epsilon^{-2})$	$\tilde{O}(\kappa^6\epsilon^{-2})$	
stocBiO	[22]	$O(\kappa^5\epsilon^{-2})$	$O(\kappa^9\epsilon^{-2})$	$O(\kappa^5\epsilon^{-2})$	$\tilde{O}(\kappa^6\epsilon^{-2})$	
SBiO-BreD	Ours	$O(\kappa^5\epsilon^{-2})$	$O(\kappa^5\epsilon^{-2})$	$O(\kappa^5\epsilon^{-2})$	$\tilde{O}(\kappa^6\epsilon^{-2})$	$\surd$
ASBiO-BreD	Ours	$O(\kappa^5\epsilon^{-1.5})$	$O(\kappa^5\epsilon^{-1.5})$	$O(\kappa^5\epsilon^{-1.5})$	$\tilde{O}(\kappa^6\epsilon^{-1.5})$	$\surd$

Many recent machine learning research problems utilize the stochastic loss functions. Thus, we also consider the following stochastic bilevel optimization problem:

$$\min_{x\in\mathcal{X}\subseteq\mathbb{R}^{d_1}} \mathbb{E}_{\xi\sim\mathcal{D}}\big[f(x,y^*(x);\xi)\big] + h(x), \qquad \text{(Outer)} \tag{2}$$
$$\text{s.t. } y^*(x) \in \arg\min_{y\in\mathbb{R}^{d_2}} \mathbb{E}_{\zeta\sim\mathcal{D}'}\big[g(x,y;\zeta)\big], \qquad \text{(Inner)}$$

where function $F(x) = \mathbb{E}_{\xi}\big[F(x;\xi)\big] = \mathbb{E}_{\xi}\big[f(x,y^*(x);\xi)\big]$ is smooth and possibly nonconvex, and function $h(x)$ is convex and possibly nonsmooth, and function $g(x,y) = \mathbb{E}_{\zeta}\big[g(x,y;\zeta)\big] : \mathcal{X}\times\mathbb{R}^{d_2} \to \mathbb{R}$ is μ-strongly convex in $y \in \mathbb{R}^{d_2}$. ξ and ζ are random variables following unknown distributions $\mathcal{D}$ and $\mathcal{D}'$, respectively. Both Problem (1) and Problem (2) have been used in many machine learning tasks with a hierarchical structure, such as hyper-parameter meta-learning [9, 22] and neural network architecture search [30].

Many bilevel optimization methods recently have been developed to solve these problems. For example, [11, 22] introduced a class of effective methods to solve the above deterministic Problem (1) and stochastic Problem (2) with $h(x) = 0$. However, these methods suffer from high computational complexity issue. More recently, multiple accelerated methods were designed for stochastic Problem (2) with $h(x) = 0$. Specifically, [5, 23, 14, 41] proposed accelerated bilevel optimization algorithms via using the variance reduced techniques of SARAH/SPIDER/SNVRG [36, 8, 40, 43] and STORM [6]. However, these accelerated methods obtain a lower computational complexity without considering the condition number, which also accounts for an important part of the computational complexity (please see Tables 1 and 2). Meanwhile, these accelerated methods only focus on the special case of the stochastic bilevel optimization Problem (2) with $h(x) = 0$.

To fill in the gaps, in the paper, we propose a class of efficient bilevel optimization methods with lower computational complexity to solve the bilevel optimization Problems (1) and (2), where the outer subproblem is nonconvex and possibly nonsmooth, and the inner subproblem is strongly convex. Specifically, we use the mirror decent iteration to update the variable x based on the Bregman distance. Our main contributions are summarized as follows:

(i) We propose a class of enhanced bilevel optimization methods based on Bregman distance to solve the nonconvex-strongly-convex bilevel optimization problems. Moreover, we provide a comprehensive convergence analysis framework for our proposed methods.

(ii) An efficient bilevel optimization method based on Bregman distances (BiO-BreD) is proposed to solve the deterministic bilevel Problem (1). We prove that our BiO-BreD achieves a lower sample complexity than the best known results (please see Table 1).

(iii) We introduce an efficient bilevel optimization method based on adaptive Bregman distances (SBiO-BreD) to solve the stochastic bilevel Problem (2). Moreover, we design an accelerated version of SBiO-BreD algorithm (ASBiO-BreD) via using the variance reduced technique, which achieves a lower sample complexity than the best known results (please see Table 2).

Note that our methods can solve the constrained bilevel optimization with nonsmooth regularization but not rely on any form of constraint set and nonsmooth regularization. In the other words, our methods can solve the unconstrained bilevel optimization without nonsmooth regularization studied in [11, 22]. Naturally, our convergence analysis can be applied to both the constrained bilevel optimization with nonsmooth regularization and the unconstrained bilevel optimization without nonsmooth regularization.

2 Related Works

In this section, we will revisit the existing bilevel optimization algorithms and Bregman distance based methods.

2.1 Bilevel Optimization Methods

Bilevel optimization recently has attracted increasing interest in many machine learning applications such as model-agnostic meta-learning, neural network architecture search, and policy optimization. Thus, recently many algorithms [9, 11, 15, 34, 35, 22, 28] have been proposed to solve the bilevel optimization problems. Specifically, [11] proposed a class of approximation methods for bilevel optimization and studied convergence properties of the proposed methods under convexity assumption. [34, 35] developed the gradient-based descent aggregation methods for convex bilevel optimization. [37] presented a nonlinear primal–dual algorithm for nonsmooth convex bilevel optimization in parameter learning problems.

In parallel, [15] introduced a two-timescale stochastic algorithm framework for nonconvex stochastic bilevel optimization in reinforcement learning. Multiple accelerated bilevel approximation methods were developed later. Specifically, [22] proposed faster bilevel optimization methods based on the approximated implicit differentiation (AID) and iterative differentiation (ITD), respectively. [5, 23, 14, 41] presented several accelerated bilevel methods for the stochastic bilevel problems using variance-reduced techniques. More recently, [18] proposed a class of efficient adaptive methods for nonconvex-strongly-convex bilevel optimization problems. At the same time, the lower bound of bilevel optimization methods has been studied in [21] for these nonconvex-strongly-convex bilevel optimization problems. In addition, [34, 27, 32, 33] designed a class of value-function-based and gradient-based bilevel methods for nonconvex bilevel optimization problems and studied asymptotic convergence properties of these methods. [38] analyzed a class of special nonconvex nonsmooth bilevel optimization methods for selecting the best hyperparameter value for the nonsmooth ℓ_p regularization with $0 < p \leq 1$.

2.2 Bregman Distance-Based Methods

Bregman distance-based method (*a.k.a*, mirror descent method) [4, 1] is a powerful optimization tool because it uses the Bregman distance to fit the geometry of optimization problems. Bregman distance was first proposed in [2], and later extended in [3]. [4] introduced the first proximal minimization algorithm with Bregman function. [1] studied the mirror descent for convex optimization. [7] presented an effective variant of mirror descent, *i.e.* composite objective mirror descent, for regularized convex optimization. Subsequently, [42] studied the convergence properties of mirror descent algorithm for solving nonsmooth nonconvex problems. [26] integrated the variance reduced technique to the mirror descent algorithm for stochastic convex optimization. The variance-reduced adaptive stochastic mirror descent algorithm [29] was proposed to solve the nonsmooth nonconvex finite-sum optimization. More recently, [16] studied Bregman gradient methods for policy optimization.

3 Preliminaries

3.1 Notations

Let I_d denote a d-dimensional identity matrix. $\mathcal{U}\{1, 2, \cdots, K\}$ denotes a uniform distribution over a discrete set $\{1, 2, \cdots, K\}$. $\|\cdot\|$ denotes the ℓ_2-norm for vectors and spectral norm for matrices,

respectively. For two vectors x and y, $\langle x, y\rangle$ denotes their inner product. $\nabla_x f(x,y)$ and $\nabla_y f(x,y)$ are the partial derivatives *w.r.t.* variables x and y. Given the mini-batch samples $\mathcal{B} = \{\xi^i\}_{i=1}^b$, we define $\nabla f(x;\mathcal{B}) = \frac{1}{b}\sum_{i=1}^b \nabla f(x;\xi^i)$. For two sequences $\{a_n, b_n\}_{i=1}^n$, $a_n = O(b_n)$ denotes that $a_n \leq Cb_n$ for some constant $C > 0$. The notation $\tilde{O}(\cdot)$ hides logarithmic terms. Given a convex closed set $\mathcal{X}$, we define a projection operation $\mathcal{P}_{\mathcal{X}}(x_0) = \arg\min_{x\in\mathcal{X}} \|x - x_0\|^2$. $\partial h(x)$ is the subgradient set of function $h(x)$.

3.2 Some Mild Assumptions

Assumption 1. *Function $F(x) = f(x, y^*(x))$ is possibly nonconvex w.r.t. x, and function $g(x,y)$ is μ-strongly convex w.r.t. y. For stochastic case, the same assumptions hold for $f(x, y^*(x); \xi)$ and $g(x, y; \zeta)$, respectively.*

Assumption 2. *Functions $f(x,y)$ and $g(x,y)$ satisfy*

1) *$\|\nabla_y f(x,y)\| \leq C_{fy}$ and $\|\nabla^2_{xy} g(x,y)\| \leq C_{gxy}$ for any $x \in \mathcal{X}$ and $y \in \mathbb{R}^{d_2}$;*
2) *The partial derivatives $\nabla_x f(x,y)$, $\nabla_y f(x,y)$, $\nabla_x g(x,y)$ and $\nabla_y g(x,y)$ are L-Lipschitz, e.g., for $x, x_1, x_2 \in \mathcal{X}$ and $y, y_1, y_2 \in \mathbb{R}^{d_2}$,*

$$\|\nabla_x f(x_1,y) - \nabla_x f(x_2,y)\| \leq L\|x_1 - x_2\|,\ \|\nabla_x f(x,y_1) - \nabla_x f(x,y_2)\| \leq L\|y_1 - y_2\|.$$

For stochastic case, the same assumptions hold for $f(x,y;\xi)$ and $g(x,y;\zeta)$ for any ξ and ζ.

Assumption 3. *The partial derivatives $\nabla^2_{xy} g(x,y)$ and $\nabla^2_{yy} g(x,y)$ are L_{gxy}-Lipschitz and L_{gyy}-Lipschitz, e.g., for all $x, x_1, x_2 \in \mathcal{X}$ and $y, y_1, y_2 \in \mathbb{R}^{d_2}$*

$$\|\nabla^2_{xy} g(x_1,y) - \nabla^2_{xy} g(x_2,y)\| \leq L_{gxy}\|x_1 - x_2\|,\ \|\nabla^2_{xy} g(x,y_1) - \nabla^2_{xy} g(x,y_2)\| \leq L_{gxy}\|y_1 - y_2\|.$$

For stochastic case, the same assumptions hold for $\nabla^2_{xy} g(x,y;\zeta)$ and $\nabla^2_y g(x,y;\zeta)$ for any ζ.

Assumption 4. *Function $h(x)$ for any $x \in \mathcal{X}$ is convex but possibly nonsmooth.*

Assumption 5. *Function $\Phi(x) = F(x) + h(x)$ is bounded below, i.e., $\Phi^* = \inf_{x\in\mathcal{X}} \Phi(x) > -\infty$.*

Assumptions 1-3 are commonly used in bilevel optimization methods [11, 22, 23]. According to Assumption 1, $\|f(x,y_1) - f(x,y_2)\| = \|\nabla_y f(x,y_\tau)(y_1 - y_2)\| \leq \|\nabla_y f(x,y_\tau)\|\|y_1 - y_2\| \leq C_{fy}\|y_1 - y_2\|$, where $y_\tau = \tau y_1 + (1-\tau)y_2$ and $\tau \in [0,1]$. Thus $\|\nabla_y f(x,y)\| \leq C_{fy}$ is similar to the assumption that the function f is M-Lipschitz in [22]. From the proofs in [22], we can find that they still use the norm bounded partial derivative $\|\nabla_y f(x,y)\| \leq M$. Similarly, according to Assumption 1, we have $\|\nabla_y g(x_1,y) - \nabla_y g(x_2,y)\| \leq L\|x_1 - x_2\|$. Since $\|\nabla_y g(x_1,y) - \nabla_y g(x_2,y)\| = \|\nabla^2_{xy} g(x_{\tau'},y)(x_1 - x_2)\| \leq \|\nabla^2_{xy} g(x_{\tau'},y)\|\|x_1 - x_2\| \leq C_{gxy}\|x_1 - x_2\|$, where $x_{\tau'} = \tau' x_1 + (1-\tau')x_2$ and $\tau' \in [0,1]$, we can let $C_{gxy} = L$ as in [22]. From the proofs in [22], we can find that they still use the norm bounded partial derivative $\|\nabla^2_{xy} g(x,y)\| \leq L$ for all x, y. Throughout the paper, we let $C_{gxy} = L$. Assumption 4 is generally used for regularization such as $h(x) = \|x\|_1$. Assumption 5 ensures the feasibility of Problems (1) and (2).

When we use the first-order methods to solve the above bilevel optimization Problems (1) and (2), we can easily obtain the partial (stochastic) derivative $\nabla_y g(x,y)$ or $\nabla_y g(x,y;\zeta)$ to update variable y. However, it is hard to get the (stochastic) gradient $\nabla F(x) = \frac{\partial f(x,y^*(x))}{\partial x}$ or $\nabla F(x;\xi) = \frac{\partial f(x,y^*(x);\xi)}{\partial x}$, when there is no closed form solution for the inner problem of Problems (1) and (2). Thus, a key point of solving the Problems (1) and (2) is to estimate the gradient $\nabla F(x)$. The following lemma provides one gradient estimator of $\nabla F(x)$.

Lemma 1. *(Lemma 2.1 in [11]) Under the above Assumptions (1, 2, 3), we have, for any $x \in \mathcal{X}$*

$$\begin{aligned}\nabla F(x) &= \nabla_x f(x,y^*(x)) + \nabla y^*(x)^T \nabla_y f(x,y^*(x)) \\ &= \nabla_x f(x,y^*(x)) - \nabla^2_{xy} g(x,y^*(x))[\nabla^2_{yy} g(x,y^*(x))]^{-1}\nabla_y f(x,y^*(x)).\end{aligned} \tag{3}$$

Lemma 1 provides a natural estimator of $\nabla F(x)$, defined as, for all $x \in \mathcal{X}, y \in \mathbb{R}^{d_2}$

$$\bar{\nabla} f(x,y) = \nabla_x f(x,y) - \nabla^2_{xy} g(x,y)\big(\nabla^2_{yy} g(x,y)\big)^{-1}\nabla_y f(x,y). \tag{4}$$

Next, we show some properties of $\nabla F(x)$, $y^*(x)$ and $\bar{\nabla} f(x,y)$ in the following lemma:

Algorithm 1 Deterministic BiO-BreD Algorithm

Input: T, $K \geq 1$, learning rates $\gamma > 0$, $\lambda > 0$;
initialize: $x_0 \in \mathcal{X}$ and $y_{-1}^K = y_0 \in \mathbb{R}^{d_2}$;
for $t = 0, 1, \cdots, T-1$ **do**
  Let $y_t^0 = y_{t-1}^K$;
  for $k = 1, \cdots, K$ **do**
    Update $y_t^k = y_t^{k-1} - \lambda \nabla_y g(x_t, y_t^{k-1})$;
  end for
  Compute partial derivative $w_t = \frac{\partial f(x_t, y_t^K)}{\partial x}$ via backpropagation *w.r.t.* x_t;
  Given a ρ-strongly convex mirror function ψ_t;
  Update $x_{t+1} = \arg\min_{x \in \mathcal{X}} \big\{ \langle w_t, x \rangle + h(x) + \frac{1}{\gamma} D_{\psi_t}(x, x_t) \big\}$;
end for
Output: Uniformly and randomly choose from $\{x_t, y_t\}_{t=1}^T$.

Lemma 2. *(Lemma 2.2 in [11]) Under the Assumptions (1, 2, 3), for all $x, x_1, x_2 \in \mathcal{X}$ and $y \in \mathbb{R}^{d_2}$, we have $\|\bar{\nabla} f(x, y) - \nabla F(x)\| \leq L_y \|y^*(x) - y\|$*

$$\|y^*(x_1) - y^*(x_2)\| \leq \kappa \|x_1 - x_2\|, \ \|\nabla F(x_1) - \nabla F(x_2)\| \leq L_F \|x_1 - x_2\|,$$

where $L_y = L + \frac{L^2}{\mu} + \frac{C_{fy} L_{gxy}}{\mu} + \frac{L_{gyy} C_{fy} L}{\mu^2}$, $\kappa = \frac{L}{\mu}$, *and* $L_F = L + \frac{2L^2 + L_{gxy} C_{fy}^2}{\mu} + \frac{L_{gyy} C_{fy} L + L^3 + L_{gxy} C_{fy} L}{\mu^2} + \frac{L_{gyy} C_{fy} L^2}{\mu^3}$.

4 Bilevel Optimization via Bregman Distance Methods

In this section, we propose a class of enhanced bilevel optimization methods based on Bregman distance to solve the deterministic Problem (1) and the stochastic Problem (2), respectively.

4.1 Deterministic BiO-BreD Algorithm

In this subsection, we propose an efficient deterministic bilevel optimization method via Bregman distances (BiO-BreD) to solve the deterministic Problem (1). Algorithm 1 summarizes the algorithmic framework of our BiO-BreD method.

Given a ρ-strongly convex and continuously-differentiable function $\psi(x)$, *i.e.*, $\langle x_1 - x_2, \nabla \psi(x_1) - \nabla \psi(x_2) \rangle \geq \rho \|x_1 - x_2\|^2$, we define a Bregman distance [3, 4] for any $x_1, x_2 \in \mathcal{X}$:

$$D_\psi(x_1, x_2) = \psi(x_1) - \psi(x_2) - \langle \nabla \psi(x_2), x_1 - x_2 \rangle.$$

In Algorithm 1, we use the mirror descent iteration to update the variable x at $t+1$-th step:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \big\{ \langle w_t, x \rangle + h(x) + \frac{1}{\gamma} D_{\psi_t}(x, x_t) \big\}, \tag{5}$$

where $\gamma > 0$ is stepsize, and w_t is an estimator of $\nabla F(x_t)$. Here the mirror function ψ_t can be dynamic as the algorithm is running. Let $\psi_t(x) = \frac{1}{2}\|x\|^2$, we have $D_{\psi_t}(x, x_t) = \frac{1}{2}\|x - x_t\|^2$. When $\mathcal{X} = \mathbb{R}^{d_1}$, the above subproblem (5) is equivalent to the proximal gradient descent. When $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ and $h(x) = 0$, the above subproblem (5) is equivalent to the projection gradient descent. Let $\psi_t(x) = \frac{1}{2} x^T H_t x$, we have $D_{\psi_t}(x, x_t) = \frac{1}{2}(x - x_t)^T H_t (x - x_t)$. When H_t is an approximated Hessian matrix, the above subproblem (5) is equivalent to the proximal quasi-Newton decent. When H_t is an adaptive matrix as used in [19], the above subproblem (5) is equivalent to the proximal adaptive gradient decent.

In Algorithm 1, we use gradient estimator $w_t = \frac{\partial f(x_t, y_t^K)}{\partial x}$ to estimate $\nabla F(x_t)$, where the partial derivative $w_t = \frac{\partial f(x_t, y_t^K)}{\partial x}$ is obtained by the backpropagation *w.r.t.* x_t.

4.2 SBiO-BreD Algorithm

In this subsection, we introduce an efficient stochastic bilevel optimization method via Bregman distance (SBiO-BreD) to solve the stochastic bilevel optimization Problem (2). Algorithm 2 describes the algorithmic framework of our SBiO-BreD method.

Algorithm 2 Stochastic BiO-BreD (SBiO-BreD) Algorithm

Input: $T, K \geq 1$, stepsizes $\gamma > 0$, $\lambda > 0$, $\{\eta_t\}_{t=1}^T$;
initialize: $x_0 \in \mathcal{X}$ and $y_0 \in \mathbb{R}^{d_2}$;
for $t = 0, 1, \cdots, T-1$ **do**
  Draw randomly b independent samples $\mathcal{B}_t = \{\zeta_t^i\}_{i=1}^b$, and compute stochastic partial derivatives $v_t = \nabla_y g(x_t, y_t; \mathcal{B}_t)$;
  Update $y_{t+1} = y_t - \lambda \eta_t v_t$;
  Draw randomly $b(K+1)$ independent samples $\bar{\mathcal{B}}_t = \{\xi_{t,i}, \zeta_{t,i}^0 \cdots, \zeta_{t,i}^{K-1}\}_{i=1}^b$, and compute stochastic partial derivatives $w_t = \bar{\nabla} f(x_t, y_t; \bar{\mathcal{B}}_t)$;
  Given a ρ-strongly convex mirror function ψ_t;
  Update $x_{t+1} = \arg\min_{x \in \mathcal{X}} \big\{ \langle w_t, x \rangle + h(x) + \frac{1}{\gamma} D_{\psi_t}(x, x_t) \big\}$;
end for
Output: Uniformly and randomly choose from $\{x_t, y_t\}_{t=1}^T$.

Given $K \geq 1$ and draw $K+1$ independent samples $\bar{\xi} = \{\xi, \zeta^0, \cdots, \zeta^{K-1}\}$, as in [15, 23], we definite a stochastic gradient estimator:

$$\bar{\nabla} f(x, y, \bar{\xi}) = \nabla_x f(x, y; \xi) - \nabla_{xy}^2 g(x, y; \zeta^0) \Big[\frac{K}{L} \prod_{i=1}^{k} \big(I_{d_2} - \frac{1}{L} \nabla_{yy}^2 g(x, y; \zeta^i) \big) \Big] \nabla_y f(x, y; \xi), \quad (6)$$

where $k \sim \mathcal{U}\{0, 1, \cdots, K-1\}$ is a uniform random variable independent on $\bar{\xi}$. It is easy to verify that $\bar{\nabla} f(x, y, \bar{\xi})$ is a biased estimator of $\bar{\nabla} f(x, y)$, *i.e.* $\mathbb{E}_{\bar{\xi}}\big[\bar{\nabla} f(x, y; \bar{\xi})\big] \neq \bar{\nabla} f(x, y)$. For the gradient estimator (6), thus we define a bias $R(x, y) = \bar{\nabla} f(x, y) - \mathbb{E}_{\bar{\xi}}\big[\bar{\nabla} f(x, y; \bar{\xi})\big] : \mathcal{X} \times \mathbb{R}^{d_2} \to \mathbb{R}$.

Lemma 3. *(Lemma 2.1 in [23]) Under the about Assumptions (1, 2, 3), for any $K \geq 1$, the gradient estimator in* (6) *satisfies*

$$\|R(x, y)\| \leq \frac{L C_{fy}}{\mu} \big(1 - \frac{\mu}{L}\big)^K.$$

Lemma 5 shows that the bias $R(x, y)$ decays exponentially fast with number K, and with choosing $K = \frac{L}{\mu} \log(L C_{fy} T / \mu)$, we have $\|R(x, y)\| \leq \frac{1}{T}$. Let $\frac{L C_{fy}}{\mu} \big(1 - \frac{\mu}{L}\big)^K \leq \frac{1}{T}$, we have $K \log(1 - \frac{\mu}{L}) \leq \log(\frac{\mu}{L C_{fy} T})$. Due to $\mu < L$, we have $K \geq \log(\frac{C_{fy} L T}{\mu}) / \log(\frac{L}{L - \mu})$. Further due to $\frac{\mu}{L} \leq \log(\frac{L}{L-\mu})$, let $K = \frac{L}{\mu} \log(L C_{fy} T / \mu)$, we have $\|R(x, y)\| \leq \frac{1}{T}$. Note that here we use $C_{gxy} = L$.

To simplify notations, let $\bar{\xi}_t^i = \{\xi_{t,i}, \zeta_{t,i}^0 \cdots, \zeta_{t,i}^{K-1}\}$. In Algorithm 2, we use mini-batch stochastic gradient estimator $w_t = \bar{\nabla} f(x_t, y_t; \bar{\mathcal{B}}_t) = \frac{1}{b} \sum_{i=1}^b \bar{\nabla} f(x_t, y_t; \bar{\xi}_t^i)$, where $\bar{\nabla} f(x_t, y_t; \bar{\xi}_t^i)$

$$= \nabla_x f(x_t, y_t; \xi_{t,i}) - \nabla_{xy}^2 g(x_t, y_t; \zeta_{t,i}^0) \Big[\frac{K}{L} \prod_{j=1}^{k} \big(I_{d_2} - \frac{1}{L} \nabla_{yy}^2 g(x_t, y_t; \zeta_{t,i}^j) \big) \Big] \nabla_y f(x_t, y_t; \xi_{t,i}),$$

with $k \sim \mathcal{U}\{0, 1, \cdots, K-1\}$. Let $R(x_t, y_t) = w_t - \bar{\nabla} f(x_t, y_t) = \bar{\nabla} f(x_t, y_t; \bar{\mathcal{B}}_t) - \bar{\nabla} f(x_t, y_t)$, we have $\mathbb{E}[\bar{\nabla} f(x_t, y_t; \bar{\mathcal{B}}_t)] = R(x_t, y_t) + \bar{\nabla} f(x_t, y_t)$. According to the above Lemma 5, it is easy to verify that $\|R(x_t, y_t)\| \leq \frac{L C_{fy}}{\mu} \big(1 - \frac{\mu}{L}\big)^K$.

4.3 ASBiO-BreD Algorithm

In this subsection, we propose an accelerated version of SBiO-BreD method (ASBiO-BreD) to solve the stochastic bilevel optimization Problem (2) via using variance reduced technique of SARAH/SPIDER/SNVRG [36, 8, 40, 43]. Algorithm 3 shows the algorithmic framework of the ASBiO-BreD method.

In Algorithm 3, we use the variance reduced technique of SPIDER to accelerate SBiO-BreD algorithm. When $\mod(t, q) = 0$, we draw a relative large batch samples $\mathcal{B}_t = \{\zeta_t^i\}_{i=1}^b$ and $\bar{\mathcal{B}}_t = \{\bar{\xi}_t^i\}_{i=1}^b$ to estimate our stochastic partial derivatives v_t and w_t, respectively. When $\mod(t, q) \neq 0$, we draw a mini-batch samples $\mathcal{I}_t = \{\xi_t^i\}_{i=1}^{b_1}$ and $\bar{\mathcal{I}}_t = \{\bar{\xi}_t^i\}_{i=1}^{b_1}$ $(b > b_1)$ to estimate v_t and w_t, respectively. Let $R(x_t, y_t) = \bar{\nabla} f(x_t, y_t; \bar{\mathcal{I}}_t) - \bar{\nabla} f(x_t, y_t)$ when $\mod(t, q) \neq 0$, we have $\mathbb{E}[\bar{\nabla} f(x_t, y_t; \bar{\mathcal{I}}_t)] = R(x_t, y_t) + \bar{\nabla} f(x_t, y_t)$ and $\|R(x_t, y_t)\| \leq \frac{L C_{fy}}{\mu} \big(1 - \frac{\mu}{L}\big)^K$.

Algorithm 3 Accelerated Stochastic BiO-BreD (ASBiO-BreD) Algorithm

Input: $T, K \geq 1$, q, stepsizes $\gamma > 0$, $\lambda > 0$, $\{\eta_t\}_{t=1}^T$, mini-batch sizes b and b_1;
initialize: $x_0 \in \mathcal{X}$ and $y_0 \in \mathbb{R}^{d_2}$;
for $t = 0, 1, \cdots, T-1$ **do**
  if $\mod(t, q) = 0$ **then**
    Draw randomly b independent samples $\mathcal{B}_t = \{\zeta_t^i\}_{i=1}^b$, and compute stochastic partial derivative $v_t = \nabla_y g(x_t, y_t; \mathcal{B}_t)$;
    Draw randomly $b(K+1)$ independent samples $\bar{\mathcal{B}}_t = \{\xi_{t,i}, \zeta_{t,i}^0 \cdots, \zeta_{t,i}^{K-1}\}_{i=1}^b$, and compute stochastic partial derivative $w_t = \bar{\nabla} f(x_t, y_t; \bar{\mathcal{B}}_t)$;
  else
    Generate randomly b_1 independent samples $\mathcal{I}_t = \{\zeta_t^i\}_{i=1}^{b_1}$, and compute stochastic partial derivative $v_t = \nabla_y g(x_t, y_t; \mathcal{I}_t) - \nabla_y g(x_{t-1}, y_{t-1}; \mathcal{I}_t) + v_{t-1}$;
    Generate randomly $b_1(K+1)$ independent samples $\bar{\mathcal{I}}_t = \{\xi_{t,i}, \zeta_{t,i}^0 \cdots, \zeta_{t,i}^{K-1}\}_{i=1}^{b_1}$, and compute stochastic partial derivative $w_t = \bar{\nabla} f(x_t, y_t; \bar{\mathcal{I}}_t) - \bar{\nabla} f(x_{t-1}, y_{t-1}; \bar{\mathcal{I}}_t) + w_{t-1}$;
  end if
  Update $y_{t+1} = y_t - \lambda \eta_t v_t$;
  Given a ρ-strongly convex mirror function ψ_t;
  Update $x_{t+1} = \arg\min_{x \in \mathcal{X}} \big\{ \langle w_t, x \rangle + h(x) + \frac{1}{\gamma} D_{\psi_t}(x, x_t) \big\}$;
end for
Output: Uniformly and randomly choose from $\{x_t, y_t\}_{t=1}^T$.

5 Convergence Analysis

In this section, we study the convergence properties of our new algorithms (*i.e.*, BiO-BreD, SBiO-BreD, and ASBiO-BreD) under mild conditions. All proofs are provided in the Appendix B.

We begin with introducing a useful convergence metric $\|\mathcal{G}_t\|^2$ or $\mathbb{E}\|\mathcal{G}_t\|^2$ to measure convergence properties of our algorithms. Given the generated parameter vector x_t at the t-th iteration in our algorithms, as in [10, 29], we define the generalized gradient at the t-th iteration as:

$$\mathcal{G}_t = \frac{1}{\gamma}(x_t - x_{t+1}^+), \quad x_{t+1}^+ = \arg\min_{x \in \mathcal{X}} \big\{ \langle \nabla F(x_t), x \rangle + h(x) + \frac{1}{\gamma} D_{\psi_t}(x, x_t) \big\},$$

where $F(x) = f(x, y^*(x))$ or $F(x) = \mathbb{E}_\xi[f(x, y^*(x); \xi)]$. When $\psi_t(x) = \frac{1}{2}\|x\|^2$, $\mathcal{X} = \mathbb{R}^{d_1}$ and $h(x) = c$ is a constant, we have $\|\mathcal{G}_t\|^2 = \|\nabla F(x_t)\|^2$, which is a common convergence metric used in [11, 22]. When $\psi(x) = \frac{1}{2}\|x\|^2$, $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ and $h(x) = c$ is a constant, our convergence metric is $\|\mathcal{G}_t\|^2 = \|\frac{1}{\gamma}(x_t - \mathcal{P}_{\mathcal{X}}(x_t - \gamma \nabla F(x_t))\|^2$, which was also used in [15].

Next, we provide some useful lemmas and some mild assumptions.

Lemma 4. *(Lemma 3.1 in [23]) Under the above Assumptions (1, 2, 3), stochastic gradient estimator $\bar{\nabla} f(x, y; \bar{\xi})$ is L_K-Lipschitz continuous, e.g., for $x_1, x_2 \in \mathcal{X}$ and $y \in \mathbb{R}^{d_2}$,*

$$\mathbb{E}_{\bar{\xi}}\|\bar{\nabla} f(x_1, y; \bar{\xi}) - \bar{\nabla} f(x_2, y; \bar{\xi})\|^2 \leq L_K^2 \|x_1 - x_2\|^2,$$

where $L_K^2 = 2L^2 + 6L^4 \frac{K}{2\mu L - \mu^2} + 6C_{fy}^2 L_{gxy}^2 \frac{K}{2\mu L - \mu^2} + 6L^4 \frac{K^3 L_{gyy}^2}{(L-\mu)^2(2\mu L - \mu^2)}$.

Lemma 5. *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithms 2 and 3. Under the above assumptions, given $0 < \eta_t \leq 1$ for all $t \geq 1$ and $0 < \lambda \leq \frac{1}{6L}$, we have*

$$\begin{aligned}\|y_{t+1} - y^*(x_{t+1})\|^2 \leq (1 - \frac{\eta_t \mu \lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t \lambda^2}{4}\|v_t\|^2 \\ + \frac{25\eta_t \lambda}{6\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2 + \frac{25\kappa^2}{6\eta_t \mu \lambda}\|x_{t+1} - x_t\|^2.\end{aligned}$$

The above lemma 5 basically follows the Lemma 28 of [17] used for minimax optimization.

Assumption 6. *The stochastic partial derivative $\nabla_y g(x, y; \zeta)$ satisfies $\mathbb{E}[\nabla_y g(x, y; \zeta)] = \nabla_y g(x, y)$ and $\mathbb{E}\|\nabla_y g(x, y; \zeta) - \nabla_y g(x, y)\|^2 \leq \sigma^2$. The estimated stochastic partial derivative $\bar{\nabla} f(x, y; \bar{\xi})$ defined in* (6) *satisfies $\mathbb{E}_{\bar{\xi}}\big[\bar{\nabla} f(x, y; \bar{\xi})\big] = \bar{\nabla} f(x, y) + R(x, y)$ and $\mathbb{E}_{\bar{\xi}}\|\bar{\nabla} f(x, y; \bar{\xi}) - \bar{\nabla} f(x, y) - R(x, y)\|^2 \leq \sigma^2$.*

Assumption 7. *The mirror functions $\{\psi_t(x)\}_{t=0}^T$ are ρ-strongly convex, where $\rho > 0$.*

Assumption 6 is commonly used in stochastic bilevel optimization methods [15, 23]. Assumption 7 shows that the constant ρ can be seen as a lower bound of the strong convexity of all the mirror functions $\psi_t(x)$ for all $t \geq 0$, which is widely used in mirror descent algorithms [29] and adaptive gradient algorithms [19].

5.1 Convergence Analysis of BiO-BreD Algorithm

In this subsection, we provide the convergence properties of our BiO-BreD algorithm.

Theorem 1. *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 1. Let $0 < \gamma \leq \frac{3\rho}{4L_F}$, $0 < \lambda < \frac{1}{L}$, $K = \log(T)/\log(\frac{1}{1-\lambda\mu}) + 1$ and $\|y_t^0 - y^*(x_t)\|^2 \leq \Delta$ for all $t \geq 0$, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{G}_t\|^2 \leq \frac{16\big(\Phi(x_0) - \Phi^*\big)}{3T\gamma\rho} + \frac{22\Delta L_1^2}{\rho^2 T} + \frac{22\Delta L_2^2}{\rho^2 T} + \frac{22L_3^2}{\rho^2 T^2}, \tag{7}$$

where $\kappa = \frac{L}{\mu}$, $L_1 = \frac{L(L+\mu)}{\mu}$, $L_2 = \frac{2C_{fy}(\mu L_{gxy} + L L_{gyy})}{\mu^2}$ and $L_3 = \frac{LC_{fy}}{\mu}$.

Remark 1. *Without loss of generality, let $L \geq \frac{1}{\mu}$, $\lambda = \frac{1}{2L}$, $\gamma = \frac{3\rho}{4L_F}$ and $\rho = O(L)$. It is easy to verify that our BiO-BreD algorithm has a convergence rate of $O\big(\frac{\kappa^2}{T}\big)$. Let $\frac{\kappa^2}{T} = \epsilon$, we have $T = \kappa^2\epsilon^{-1}$. Due to $K = \log(T)/\log(\frac{1}{1-\lambda\mu}) + 1$, we choose $K = O(\kappa\log(\frac{1}{\epsilon}))$ for finding ϵ-stationary point of the problem* (1)*, we need the gradient complexity: $Gc(f,\epsilon) = 2T = O(\kappa^2\epsilon^{-1})$ and $Gc(g,\epsilon) = KT = \tilde{O}(\kappa^3\epsilon^{-1})$, and the Jacobian-vector and Hessian-vector product complexities: $JV(g,\epsilon) = KT = \tilde{O}(\kappa^3\epsilon^{-1})$ and $HV(g,\epsilon) = KT = \tilde{O}(\kappa^3\epsilon^{-1})$.*

5.2 Convergence Analysis of SBiO-BreD Algorithm

In this subsection, we provide the convergence properties of our SBiO-BreD algorithm.

Theorem 2. *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 2. Let $\Delta = \|y_0 - y^*(x_0)\|^2$, $K = \frac{L}{\mu}\log(\frac{LC_{fy}T}{\mu})$, $0 < \eta = \eta_t \leq 1$, $0 < \gamma \leq \min\big(\frac{3\rho}{4L_F}, \frac{9\eta\rho\mu\lambda}{800\kappa^2}, \frac{\eta\mu\rho\lambda}{47L_y^2}\big)$ and $0 < \lambda \leq \frac{1}{6L}$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}_t\|^2 \leq \frac{32(\Phi(x_0) - \Phi^*)}{3T\gamma\rho} + \frac{32\Delta}{3T\gamma\rho} + \frac{752\sigma^2}{3\rho^2 b} + \frac{400\eta\lambda\sigma^2}{9\gamma\rho\mu b} + \frac{752}{3\rho^2T^2}. \tag{8}$$

Remark 2. *Without loss of generality, let $L \geq \frac{1}{\mu}$, $\lambda = \frac{1}{6L}$, $\gamma = \min\big(\frac{3\rho}{4L_F}, \frac{9\eta\rho\mu\lambda}{800\kappa^2}, \frac{\eta\mu\rho\lambda}{47L_y^2}\big)$ and $\rho = O(L)$, we have $\gamma\rho = O(\frac{1}{\kappa^3})$ It is easily verified that our SBiO-BreD algorithm has a convergence rate of $O\big(\frac{\kappa^3}{T} + \frac{\kappa^2}{b}\big)$. Let $\frac{\kappa^3}{T} = \frac{\epsilon}{2}$ and $\frac{\kappa^2}{b} = \frac{\epsilon}{2}$, we have $T = 2\kappa^3\epsilon^{-1}$ and $b = 2\kappa^2\epsilon^{-1}$. Due to $K = \frac{L}{\mu}\log(\frac{LC_{fy}T}{\mu})$, we have $K = O(\kappa\log(\frac{\kappa^4}{\epsilon})) = \tilde{O}(\kappa)$. For finding ϵ-stationary point of the problem* (2)*, we need the gradient complexity: $Gc(f,\epsilon) = 2bT = \kappa^5\epsilon^{-2}$ and $Gc(g,\epsilon) = bT = O(\kappa^5\epsilon^{-2})$, and the Jacobian-vector and Hessian-vector product complexities: $JV(g,\epsilon) = bT = O(\kappa^5\epsilon^{-2})$ and $HV(g,\epsilon) = KbT = \tilde{O}(\kappa^6\epsilon^{-2})$.*

5.3 Convergence Analysis of ASBiO-BreD Algorithm

In this subsection, we provide the convergence properties of our ASBiO-BreD algorithm.

Theorem 3. *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 3. Let $\Delta = \|y_0 - y^*(x_0)\|^2$, $b_1 = q$, $K = \frac{L}{\mu}\log(\frac{LC_{fy}T}{\mu})$, $0 < \eta = \eta_t \leq 1$, $0 < \gamma \leq \min\big(\frac{3\rho}{38L_K^2\eta}, \frac{3\rho}{4L_F}, \frac{2\rho\eta\mu\lambda}{19L_y^2}, \frac{\rho\eta}{8}, \frac{9\rho\eta\mu\lambda}{400\kappa^2}\big)$ and $0 < \lambda \leq \min\big(\frac{1}{6L}, \frac{9\mu}{100\eta^2L^2}\big)$, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathcal{G}_t\|^2 \leq \frac{32(\Phi(x_0) - \Phi^*)}{3T\gamma\rho} + \frac{32\Delta}{3T\gamma\rho} + \frac{152}{3T^2\rho^2} + \frac{4}{\eta\rho\gamma}\big(\frac{1}{L^2} + \frac{1}{L_K^2}\big)\frac{\sigma^2}{b}. \tag{9}$$

Remark 3. *Without loss of generality, let $L \geq \frac{1}{\mu}$, $\lambda = \min\big(\frac{1}{6L}, \frac{9\mu}{100\eta^2L^2}\big)$, $\gamma = \min\big(\frac{3\rho}{38L_K^2\eta}, \frac{3\rho}{4L_F}, \frac{2\rho\eta\mu\lambda}{19L_y^2}, \frac{\rho\eta}{8}, \frac{9\rho\eta\mu\lambda}{400\kappa^2}\big)$ and $\rho = O(L)$, we have $\gamma\rho = O(\frac{1}{\kappa^4})$. It is easily verified that our ASBiO-BreD algorithm has a convergence rate of $O\big(\frac{\kappa^4}{T} + \frac{\kappa^2}{b}\big)$. Let $\frac{\kappa^4}{T} = \frac{\epsilon}{2}$ and $\frac{\kappa^2}{b} = \frac{\epsilon}{2}$, we have*

$T = 2\kappa^4\epsilon^{-1}$ *and* $b = 2\kappa^2\epsilon^{-1}$. *Due to* $K = \frac{L}{\mu}\log(\frac{LC_{fy}T}{\mu})$, *we have* $K = O(\kappa\log(\frac{\kappa^4}{\epsilon})) = \tilde{O}(\kappa)$. *Let* $b_1 = q = \kappa\epsilon^{-0.5}$. *For finding* ϵ*-stationary point of the problem* (2), *we need the gradient complexity:* $Gc(f,\epsilon) = 2(\frac{bT}{q} + 2b_1T) = O(\kappa^5\epsilon^{-1.5})$ *and* $Gc(g,\epsilon) = \frac{bT}{q} + 2b_1T = O(\kappa^5\epsilon^{-1.5})$, *and the Jacobian-vector and Hessian-vector product complexities:* $JV(g,\epsilon) = \frac{bT}{q} + 2b_1T = O(\kappa^5\epsilon^{-1.5})$ *and* $HV(g,\epsilon) = K(\frac{bT}{q} + 2b_1T) = \tilde{O}(\kappa^6\epsilon^{-1.5})$.

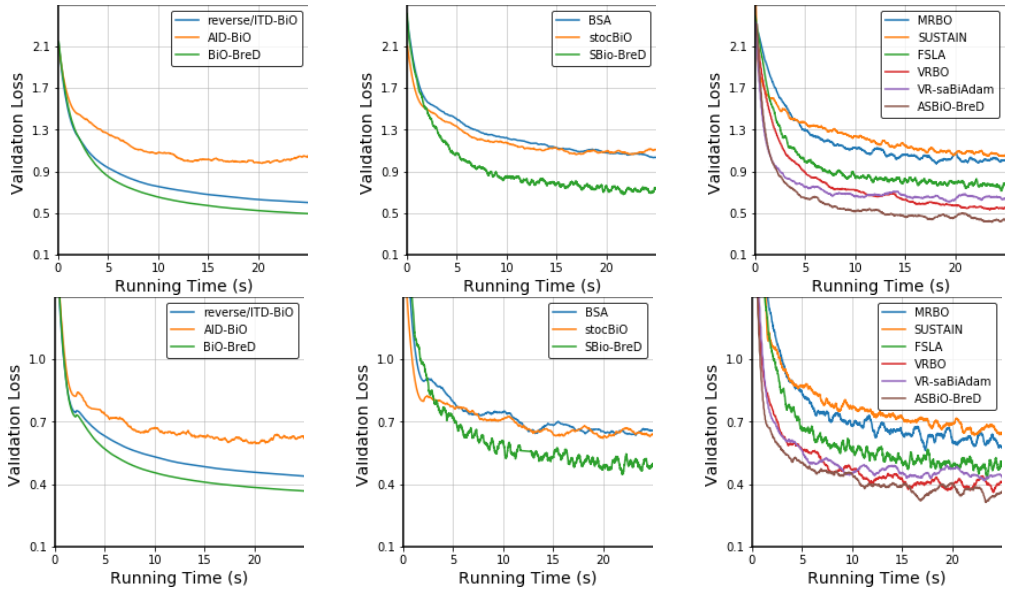

Figure 1: Validation Loss vs. Running Time for different methods. We compare our BiO-BreD with deterministic baselines (the first column), SBiO-BreD with stochastic baselines (the second column); ASBiO-BreD with momentum-based or SPIDER/SARAH based baselines (the last column). We test two values of ϱ: large noise setting $\varrho = 0.8$ (top row) and small noise setting $\varrho = 0.4$ (bottom row).

6 Numerical Experiments

In this section, we perform two tasks to demonstrate the efficiency of our algorithms: 1) data hyper-cleaning task [39] over the MNIST dataset [25]; 2) hyper-representation learning task [9] over the Omniglot dataset [24]. In the experiment, we compare our algorithms (*i.e.*, BiO-BreD, SBiO-BreD, and ASBiO-BreD) with the following bilevel optimization algorithms: reverse [9]/AID-BiO [11, 22], AID-CG [12], AID-FP [12], stocBiO [22]), MRBO [21], VRBO [21], FSLA [28], SUSTAIN [23], and VR-saBiAdam [18]. All experiments are averaged over 5 runs and we use a server with AMD EPYC 7763 64-Core CPU and 1 NVIDIA RTX A5000.

We use Bregman function $\psi_t(x) = \frac{1}{2}x^T H_t x$ to generate the Bregman distance in our algorithms, where H_t is the adaptive matrix as used in [19], *i.e.* the exponential moving average of the square of the gradient and we use coefficient 0.99 in all experiments.

6.1 Data Hyper-cleaning

In this subsection, we perform data hyper-cleaning over the MNIST dataset [25]. The formulation of this problem is as follows:

$$\min_{\lambda}\ l_{val}\big(\lambda, w^*(\lambda)\big) := \frac{1}{|D_{\mathcal{V}}|}\sum_{(x_i,y_i)\in D_{\mathcal{V}}} l\big(x_i^T w^*(\lambda), y_i\big)$$

$$\text{s.t.}\ \ w^*(\lambda) = \arg\min_{w}\ l_{tr}(\lambda, w) := \frac{1}{|D_{\mathcal{T}}|}\sum_{(x_i,y_i)\in D_{\mathcal{T}}} \sigma(\lambda_i) l(x_i^T w, y_i) + C\|w\|^2,$$

where $l(\cdot)$ denotes the cross entropy loss, $D_{\mathcal{T}}$ and $D_{\mathcal{V}}$ are training and validation datasets, respectively. Here $\lambda = \{\lambda_i\}_{i\in\mathcal{D}_{\mathcal{T}}}$ are hyper-parameters and $C \geq 0$ is a tuning parameter, $\sigma(\cdot)$ denotes the sigmoid function. In experiment, we set $C = 0.001$. The dataset includes a training set and a validation set

Table 3: Validation accuracy $vs.$ Running Time (5-way-1-shot) for different methods (with L_1 regularization)

Time	AID_BiO	ITD_BiO	MRBO	FSLA	VRBO	VR-saBiAdam	ASBiO-BreD
20s	0.6509	0.6411	0.6103	0.6539	0.5951	0.6812	0.6653
40s	0.7365	0.7210	0.6971	0.7399	0.6805	0.7141	0.7403
60s	0.7762	0.7721	0.7519	0.7661	0.7429	0.7523	**0.7830**

Table 4: Validation accuracy $vs.$ Running Time (5-way-5-shot) for different methods (with L_1 regularization)

Time	AID_BiO	ITD_BiO	MRBO	FSLA	VRBO	VR-saBiAdam	ASBiO-BreD
20s	0.8316	0.8131	0.8174	0.7993	0.7730	0.7753	0.8529
40s	0.8779	0.8621	0.8634	0.8485	0.8305	0.8188	0.8967
60s	0.9032	0.8968	0.8819	0.8824	0.8745	0.8640	**0.9313**

where each contains 5000 images. A portion of the training data are corrupted by randomly changing their labels, and we denote the portion of corrupted images as ϱ.

The detailed experimental setup is described in the Appendix A.1. For hyper-parameters, we perform grid search for our algorithms and other baselines to choose the best setting. The experimental results are summarized in Figure 1. As shown by the figure, BiO-BreD outperforms the reverse algorithm; SBiO-BreD outperforms AID-FP/stocBiO and AID-CG methods, and ASBiO-BreD outperforms the other SPIDER based algorithm MRBO and several momentum-based variance reduction methods: MRBO, SUSTAIN, FSLA, and VR-saBiAdam.

6.2 Hyper-representation Learning

In this subsection, we perform the hyper-representation learning task over the Omniglot dataset [24]. The formulation of this problem is as follows:

$$\min_{\lambda} \ l_{val}\big(\lambda, w^*(\lambda)\big) := \mathbb{E}_{\xi}\Big[\frac{1}{|D_{\mathcal{V},\xi}|}\sum_{(x_i,y_i)\in D_{\mathcal{V},\xi}} l\Big(w_{\xi}^*(\lambda)^T\phi(x_i;\lambda), y_i\Big); \xi\Big] + \alpha\|\lambda\|_1$$

$$\text{s.t. } w_{\xi}^*(\lambda) = \arg\min_{w} \ l_{tr}(\lambda, w; \xi) := \frac{1}{|D_{\mathcal{T},\xi}|}\sum_{(x_i,y_i)\in D_{\mathcal{T},\xi}} l\Big(w^T\phi(x_i;\lambda), y_i\Big) + C\|w\|^2,$$

where $l(\cdot)$ denotes the cross entropy loss, $D_{\mathcal{T},\xi}$ and $D_{\mathcal{V},\xi}$ are training and validation datasets for randomly sampled meta task ξ. Here $\phi(\cdot,\cdot)$ is a four-layers convolutional neural network with max-pooling and 32 filters per layer [9], which denotes a representation mapping. λ denotes the parameter vector of the representation mapping $\phi(\cdot,\cdot)$, and $C \geq 0$ is a tuning parameter to guarantee the inner problem to be strongly convex. The term $\alpha\|\lambda\|_1$ imposes the sparsity of hyper-representations. In the experiment, we set $\alpha = 0.001$ and $C = 0.01$.

The detailed experimental setup is described in the Appendix A.2. The results of validation accuracy (test accuracy) are summarized in Table 3 and 4. From these results, our ASBiO-BreD algorithm outperforms other baselines in the non-smooth case. We also consider the smooth case, where the upper level problem is not added the L_1 regularization. The results without L_1 regularization are given in the Appendix A.2.

7 Conclusions

In the paper, we proposed a class of enhanced bilevel optimization methods based on the Bregman distance to solve the nonconvex-strongly-convex bilevel optimization problems possibly with nonsmooth regularization. Moreover, we provided a comprehensive theoretical analysis framework to analyze our methods. The theoretical results show that our methods outperform the best known computational complexities with respect to the condition number κ and the target accuracy ϵ for finding an ϵ-stationary point.

Acknowledgments and Disclosure of Funding

This work was partially supported by NSF IIS 1838627, 1837956, 1956002, 2211492, CNS 2213701, CCF 2217003, DBI 2225775. FH was a postdoctoral researcher at the University of Pittsburgh. FH and HH are the corresponding authors.

References

[1] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[2] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[3] Y. Censor and A. Lent. An iterative row-action method for interval convex programming. *Journal of Optimization theory and Applications*, 34(3):321–353, 1981.

[4] Y. Censor and S. A. Zenios. Proximal minimization algorithm withd-functions. *Journal of Optimization Theory and Applications*, 73(3):451–464, 1992.

[5] T. Chen, Y. Sun, and W. Yin. A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*, 2021.

[6] A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex sgd. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 15236–15245, 2019.

[7] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, pages 14–26. Citeseer, 2010.

[8] C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018.

[9] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018.

[10] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

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

[12] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.

[13] Z. Guo, Y. Xu, W. Yin, R. Jin, and T. Yang. On stochastic moving-average estimators for non-convex optimization. *arXiv preprint arXiv:2104.14840*, 2021.

[14] Z. Guo and T. Yang. Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

[15] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

[16] F. Huang, S. Gao, and H. Huang. Bregman gradient policy optimization. In *International Conference on Learning Representations*, 2022.

[17] F. Huang, S. Gao, J. Pei, and H. Huang. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal of Machine Learning Research*, 23(36):1–70, 2022.

[18] F. Huang and H. Huang. Biadam: Fast adaptive bilevel optimization methods. *arXiv preprint arXiv:2106.11396*, 2021.

[19] F. Huang, J. Li, and H. Huang. Super-adam: faster and universal framework of adaptive gradients. *Advances in Neural Information Processing Systems*, 34:9074–9085, 2021.

[20] S. Jenni and P. Favaro. Deep bilevel learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 618–633, 2018.

[21] K. Ji and Y. Liang. Lower bounds and accelerated algorithms for bilevel optimization. *arXiv preprint arXiv:2102.03926*, 2021.

[22] K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892. PMLR, 2021.

[23] P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in Neural Information Processing Systems*, 34:30271–30283, 2021.

[24] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[26] L. Lei and M. I. Jordan. On the adaptivity of stochastic gradient-based optimization. *SIAM Journal on Optimization*, 30(2):1473–1500, 2020.

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

[28] J. Li, B. Gu, and H. Huang. A fully single loop algorithm for bilevel optimization without hessian inverse. *arXiv preprint arXiv:2112.04660*, 2021.

[29] W. Li, Z. Wang, Y. Zhang, and G. Cheng. Variance reduction on adaptive stochastic mirror descent. *arXiv preprint arXiv:2012.13760*, 2020.

[30] H. Liu, K. Simonyan, and Y. Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations*, 2018.

[31] R. Liu, J. Gao, J. Zhang, D. Meng, and Z. Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *arXiv preprint arXiv:2101.11517*, 2021.

[32] R. Liu, X. Liu, X. Yuan, S. Zeng, and J. Zhang. A value-function-based interior-point method for non-convex bi-level optimization. *arXiv preprint arXiv:2106.07991*, 2021.

[33] R. Liu, Y. Liu, S. Zeng, and J. Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34, 2021.

[34] R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *International Conference on Machine Learning*, pages 6305–6315. PMLR, 2020.

[35] R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang. A general descent aggregation framework for gradient-based bi-level optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[36] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.

[37] P. Ochs, R. Ranftl, T. Brox, and T. Pock. Bilevel optimization with nonsmooth lower level problems. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 654–665. Springer, 2015.

[38] T. Okuno, A. Takeda, A. Kawana, and M. Watanabe. On lp-hyperparameter learning via bilevel nonsmooth optimization. *Journal of Machine Learning Research*, 22(245):1–47, 2021.

[39] A. Shaban, C.-A. Cheng, N. Hatch, and B. Boots. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1723–1732. PMLR, 2019.

[40] Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, pages 2403–2413, 2019.

[41] J. Yang, K. Ji, and Y. Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34:13670–13682, 2021.

[42] S. Zhang and N. He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.

[43] D. Zhou, P. Xu, and Q. Gu. Stochastic nested variance reduction for nonconvex optimization. *Journal of machine learning research*, 2020.

A Experimental Details

In this section, we introduce more details of our experiments. we compare our algorithms (*i.e.*, BiO-BreD, SBiO-BreD, and ASBiO-BreD) with the following bilevel optimization algorithms: reverse [9]/AID-BiO [11, 22], AID-CG [12], AID-FP [12], stocBiO [22]), MRBO [21], VRBO [21], FSLA [28], SUSTAIN [23], and VR-saBiAdam [18]. We do not include results for STABLE [5]/SVRB [14], because they require matrix inversion which does not make sufficient progress compared to other baselines within a given time range. SMB/SEMA [13] method resembles SUSTAIN, thus we do not include it in the comparison.

A.1 Data Hyper-cleaning

In this subsection, we perform data hyper-cleaning over the MNIST dataset [25]. The formulation of this problem is as follows:

$$\min_{\lambda} \ l_{val}\big(\lambda, w^*(\lambda)\big) := \frac{1}{|D_{\mathcal{V}}|} \sum_{(x_i,y_i)\in D_{\mathcal{V}}} l\big(x_i^T w^*(\lambda), y_i\big)$$
$$\text{s.t. } w^*(\lambda) = \arg\min_{w} \ l_{tr}(\lambda, w) := \frac{1}{|D_{\mathcal{T}}|} \sum_{(x_i,y_i)\in D_{\mathcal{T}}} \sigma(\lambda_i) l(x_i^T w, y_i) + C\|w\|^2,$$

where $l(\cdot)$ denotes the cross entropy loss, $D_{\mathcal{T}}$ and $D_{\mathcal{V}}$ are training and validation dataset, respectively. Here $\lambda = \{\lambda_i\}_{i\in\mathcal{D}_{\mathcal{T}}}$ are hyper-parameters and $C \geq 0$ is a tuning parameter, $\sigma(\cdot)$ denotes the sigmoid function. In experiment, we set $C = 0.001$.

For training/validation batch-size, we use batch-size of 32, while for VRBO and our ASBiO-BreD, we choose larger batch-size 5000 (parameter b in Algorithm 3) and sampling interval (parameter q in Algorithm 3) is set as 3. For stocBiO/AID-FP, AID-CG and reverse, we use the warm-start trick as our BiO-BreD algorithm, *i.e.* the inner variable starts from the state of last iteration (Line 4 of Algorithm 1). We fine tune the number of inner-loop iterations and set it to be 50 for these algorithms. For MRBO, VRBO, SUSTAIN and our SBiO-BreD/ASBiO-BreD, we set $K = 3$ to evaluate the hyper-gradient. For FSLA, $K = 1$ as the hyper-gradient is evaluated recursively. As for learning rates, we set 1000 as the outer learning rate for all algorithms except our algorithms which use 0.1 as we change the learning rate adaptively. As for the inner learning rates, we set the stepsize as 0.05 for reverse, BiO-BreD, AID-CG, stocBiO/AID-FP, MRBO/SUSTAIN, FSLA and our SBiO-BreD; we set the stepsize as 0.2 for VRBO, VR-saBiAdam and our ASBiO-BreD; we set the stepsize as 1 for SUSTAIN.

A.2 Hyper-representation Learning

In this subsection, we perform the hyper-representation learning task over the Omniglot dataset [24]. The formulation of this problem (without L_1 regularization) is as follows:

$$\min_{\lambda} \ l_{val}\big(\lambda, w^*(\lambda)\big) := \mathbb{E}_{\xi}\Big[\frac{1}{|D_{\mathcal{V},\xi}|} \sum_{(x_i,y_i)\in D_{\mathcal{V},\xi}} l\Big(w_{\xi}^*(\lambda)^T \phi(x_i;\lambda), y_i\Big); \xi\Big]$$
$$\text{s.t. } w_{\xi}^*(\lambda) = \arg\min_{w} \ l_{tr}(\lambda, w; \xi) := \frac{1}{|D_{\mathcal{T},\xi}|} \sum_{(x_i,y_i)\in D_{\mathcal{T},\xi}} l\Big(w^T \phi(x_i;\lambda), y_i\Big) + C\|w\|^2,$$

where $l(\cdot)$ denotes the cross entropy loss, $D_{\mathcal{T},\xi}$ and $D_{\mathcal{V},\xi}$ are training and validation dataset for randomly sampled meta task ξ. Here $\phi(\cdot,\cdot)$ is a four-layers convolutional neural network with maxpooling and 32 filters per layer [9], which denotes a representation mapping. λ denotes the parameter vector of the representation mapping $\phi(\cdot,\cdot)$, and $C \geq 0$ is a tuning parameter to guarantee the inner problem to be strongly convex. In the experiment, we set $C = 0.01$.

In every hyper-iteration, we choose 4 meta tasks, while for VRBO and our ASBiO-BreD, we choose larger batch-size 16 (parameter b in Algorithm 3) and sampling interval (parameter q in Algorithm 3) is set as 3. For stocBiO/AID-FP, AID-CG and reverse, we use the warm-start trick as our BiO-BreD algorithm, *i.e.* the inner variable starts from the state of last iteration (Line 4 of Algorithm 1). We fine tune the number of inner-loop iterations and set it to be 16 for these algorithms. For MRBO, VRBO, SUSTAIN and our SBiO-BreD/ASBiO-BreD, we set $K = 5$ to evaluate the hyper-gradient. For FSLA, $K = 1$ as the hyper-gradient is evaluated recursively. As for learning rates, we set 1000 as the outer learning rate for all algorithms except our algorithms which use 0.001 as we change the learning rate adaptively. As for the inner learning rates, we set the stepsize as 0.4 for all algorithms.

The experimental results are summarized in Figure 2. As shown by the figure, BiO-BreD outperforms the reverse algorithm; SBiO-BreD outperforms AID-FP/stocBiO and AID-CG methods, while SBiO-BreD outperforms another SPIDER based algorithm MRBO and several momentum-based variance reduction methods: MRBO, SUSTAIN, FSLA and VR-saBiAdam.

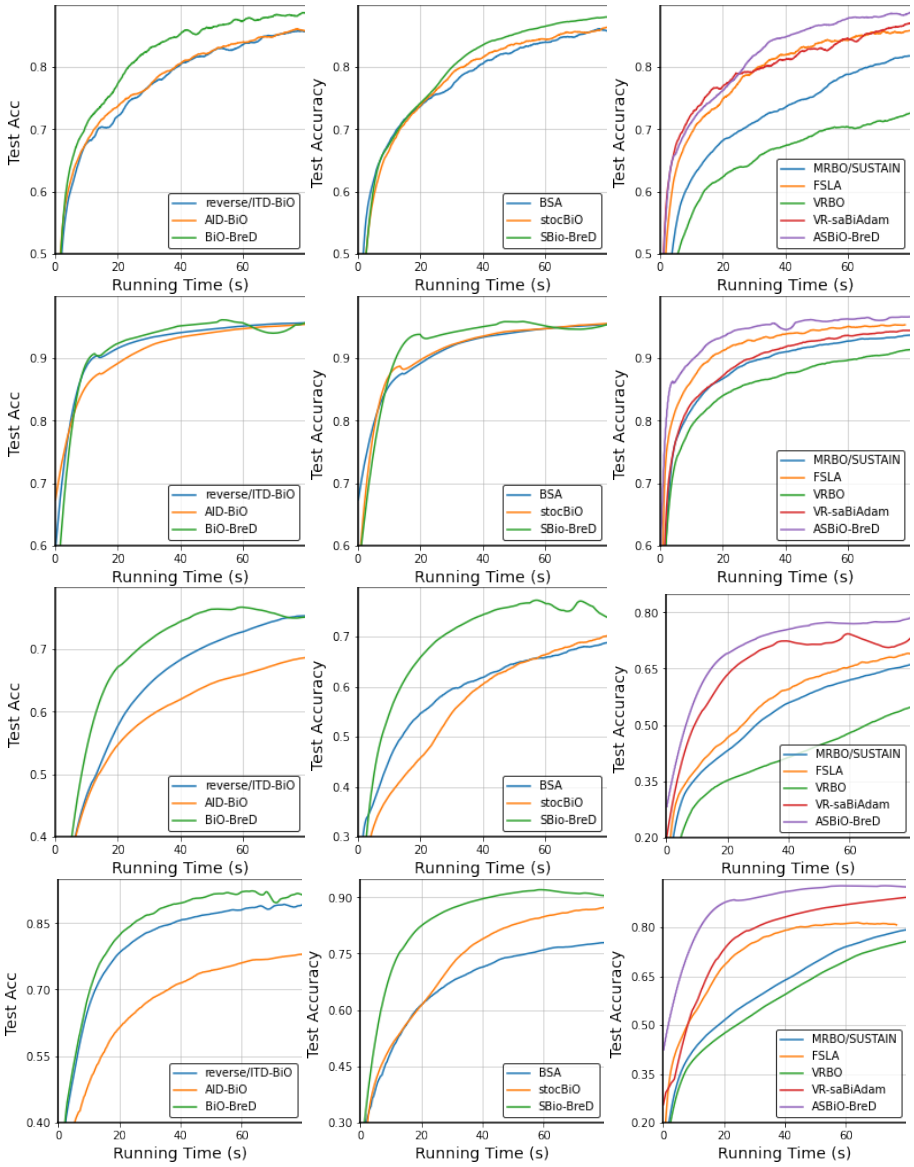

Figure 2: Validation Accuracy (Test Accuracy) $vs.$ Running Time for different methods for the Omniglot Dataset. We compare our BiO-BreD with deterministic baselines (the first column), SBiO-BreD with stochastic baselines (the second column); ASBiO-BreD with momentum-based or SPIDER/SARAH based baselines (the last column). The first row shows results for 5-way-1-shot case; the second row shows results for 5-way-5-shot case; the third row shows results for 20-way-1-shot case; the last row shows results for 20-way-5-shot case.

B Detailed Convergence Analysis

In this section, we provide the detailed convergence analysis of our algorithms. We first gives some useful lemmas.

Lemma 6. *(Proposition 2. [22]) The gradient* $\frac{\partial f(x_t,y_t^K)}{\partial x}$ *is the following analytical form:*

$$\frac{\partial f(x_t,y_t^K)}{\partial x} = \nabla_x f(x_t,y_t^K) - \lambda \sum_{k=0}^{K-1} \nabla^2_{xy} g(x_t,y_t^k) \prod_{j=k+1}^{K-1} \big(I_{d_2} - \lambda \nabla^2_{yy} g(x_t,y_t^j)\big) \nabla_y f(x_t,y_t^K).$$

The above lemma 6 shows an analytical form of w_t in Algorithm 1.

Lemma 7. *(Lemma 6. [22]) Under the above Assumptions, given the sequence $\{x_t, y_t\}_{t=1}^T$ generated from Algorithm 1, and $0 < \lambda < \frac{1}{L}$, we have*

$$\|\frac{\partial f(x_t, y_t^K)}{\partial x} - \nabla F(x_t)\| \leq \big(L_1(1-\lambda\mu)^{\frac{K}{2}} + L_2(1-\lambda\mu)^{\frac{K-1}{2}}\big)\|y_t^0 - y^*(x_t)\| + L_3(1-\lambda\mu)^K, \quad (10)$$

where $L_1 = \frac{L(L+\mu)}{\mu}$, $L_2 = \frac{2C_{fy}(\mu L_{gxy} + L L_{gyy})}{\mu^2}$ and $L_3 = \frac{LC_{fy}}{\mu}$.

The above lemma 7 shows the variance of gradient estimator $w_t = \frac{\partial f(x_t, y_t^K)}{\partial x}$ decays exponentially fast with iteration number K.

Lemma 8. *Given i.i.d. random variables $\{\zeta_i\}_{i=1}^n$ with zero mean, we have $\mathbb{E}\|\frac{1}{n}\sum_{i=1}^n \zeta_i\|^2 = \frac{1}{n}\mathbb{E}\|\zeta_i\|^2$ for any $i \in [n]$.*

Lemma 9. *(Lemma 1 in [10]) Let $x_{t+1} = \arg\min_{x\in\mathcal{X}} \big\{\langle w_t, x\rangle + h(x) + \frac{1}{\gamma}D_{\psi_t}(x, x_t)\big\}$ and $\tilde{\mathcal{G}}_t = \frac{1}{\gamma}(x_t - x_{t+1})$, we have, for all $t \geq 1$*

$$\langle w_t, \tilde{\mathcal{G}}_t\rangle \geq \rho\|\tilde{\mathcal{G}}_t\|^2 + \frac{1}{\gamma}\big(h(x_{t+1}) - h(x_t)\big), \quad (11)$$

where $\rho > 0$ depends on ρ-strongly convex function $\psi_t(x)$.

Lemma 10. *(Lemma 2 in [10]) Let $\{x_t\}_{t=1}^T$ be generated from Algorithms 1, 2 and 3, and define $x_{t+1}^+ = \arg\min_{x\in\mathcal{X}} \big\{\langle \nabla F(x_t), x\rangle + h(x) + \frac{1}{\gamma}D_{\psi_t}(x, x_t)\big\}$, and let $\mathcal{G}_t = \frac{1}{\gamma}(x_t - x_{t+1}^+)$, $\tilde{\mathcal{G}}_t = \frac{1}{\gamma}(x_t - x_{t+1})$, we have*

$$\|\mathcal{G}_t - \tilde{\mathcal{G}}_t\| \leq \frac{1}{\rho}\|\nabla F(x_t) - w_t\|, \quad (12)$$

where $F(x_t) = f(x_t, y^*(x_t))$ and $\rho > 0$ depends on ρ-strongly convex function $\psi_t(x)$.

Lemma 11. *(Restatement of Lemma 5) Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithms 2 and 3. Under the above assumptions, given $0 < \eta_t \leq 1$ for all $t \geq 1$ and $0 < \lambda \leq \frac{1}{6L}$, we have*

$$\begin{aligned}\|y_{t+1} - y^*(x_{t+1})\|^2 \leq (1 - \frac{\eta_t\mu\lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t\lambda^2}{4}\|v_t\|^2 \\ + \frac{25\eta_t\lambda}{6\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2 + \frac{25\kappa^2}{6\eta_t\mu\lambda}\|x_{t+1} - x_t\|^2, \end{aligned} \quad (13)$$

where $\kappa = L/\mu$.

Proof. We first use the step $y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t)$ and $\tilde{y}_{t+1} = y_t - \lambda v_t$ instead of the step 5 in Algorithm 2 and step 11 in Algorithm 3, i.e., $y_{t+1} = y_t - \lambda\eta_t v_t$. This proof mainly follows the proof of Lemma 28 in [17].

According to Assumption 1, i.e., the function $g(x, y)$ is μ-strongly convex w.r.t y, we have

$$\begin{aligned} g(x_t, y) &\geq g(x_t, y_t) + \langle \nabla_y g(x_t, y_t), y - y_t\rangle + \frac{\mu}{2}\|y - y_t\|^2 \\ &= g(x_t, y_t) + \langle v_t, y - \tilde{y}_{t+1}\rangle + \langle \nabla_y g(x_t, y_t) - v_t, y - \tilde{y}_{t+1}\rangle \\ &\quad + \langle \nabla_y g(x_t, y_t), \tilde{y}_{t+1} - y_t\rangle + \frac{\mu}{2}\|y - y_t\|^2. \end{aligned} \quad (14)$$

According to Assumption 2, i.e., the function $g(x, y)$ is L-smooth, we have

$$g(x_t, \tilde{y}_{t+1}) \leq g(x_t, y_t) + \langle \nabla_y g(x_t, y_t), \tilde{y}_{t+1} - y_t\rangle + \frac{L}{2}\|\tilde{y}_{t+1} - y_t\|^2. \quad (15)$$

By combining the about inequalities (14) with (15), we have

$$\begin{aligned} g(x_t, y) \geq g(x_t, \tilde{y}_{t+1}) + \langle v_t, y - \tilde{y}_{t+1}\rangle + \langle \nabla_y g(x_t, y_t) - v_t, y - \tilde{y}_{t+1}\rangle \\ + \frac{\mu}{2}\|y - y_t\|^2 - \frac{L}{2}\|\tilde{y}_{t+1} - y_t\|^2. \end{aligned} \quad (16)$$

According to $\tilde{y}_{t+1} = y_t - \lambda v_t$, we have

$$\begin{aligned}\langle v_t, y - \tilde{y}_{t+1}\rangle &= \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, \tilde{y}_{t+1} - y\rangle \\ &= \frac{1}{\lambda}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y_t - y\rangle. \end{aligned} \quad (17)$$

By plugging the inequalities (17) into (16), we have

$$g(x_t, y) \geq g(x_t, \tilde{y}_{t+1}) + \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y_t - y\rangle + \frac{1}{\lambda}\|\tilde{y}_{t+1} - y_t\|^2 + \langle \nabla_y g(x_t, y_t) - v_t, y - \tilde{y}_{t+1}\rangle + \frac{\mu}{2}\|y - y_t\|^2 - \frac{L}{2}\|\tilde{y}_{t+1} - y_t\|^2. \tag{18}$$

Let $y = y^*(x_t)$, then we have

$$g(x_t, y^*(x_t)) \geq g(x_t, \tilde{y}_{t+1}) + \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y_t - y^*(x_t)\rangle + (\frac{1}{\lambda} - \frac{L}{2})\|\tilde{y}_{t+1} - y_t\|^2 + \langle \nabla_y g(x_t, y_t) - v_t, y^*(x_t) - \tilde{y}_{t+1}\rangle + \frac{\mu}{2}\|y^*(x_t) - y_t\|^2. \tag{19}$$

Due to the strongly-convexity of $g(\cdot, y)$ and $y^*(x_t) = \arg\min_{y\in\mathcal{Y}} g(x_t, y)$, we have $g(x_t, y^*(x_t)) \leq g(x_t, \tilde{y}_{t+1})$. Thus, we obtain

$$0 \geq \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y_t - y^*(x_t)\rangle + \langle \nabla_y g(x_t, y_t) - v_t, y^*(x_t) - \tilde{y}_{t+1}\rangle + (\frac{1}{\lambda} - \frac{L}{2})\|\tilde{y}_{t+1} - y_t\|^2 + \frac{\mu}{2}\|y^*(x_t) - y_t\|^2. \tag{20}$$

By $y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t)$, we have

$$\begin{aligned}\|y_{t+1} - y^*(x_t)\|^2 &= \|y_t + \eta_t(\tilde{y}_{t+1} - y_t) - y^*(x_t)\|^2 \\ &= \|y_t - y^*(x_t)\|^2 + 2\eta_t\langle \tilde{y}_{t+1} - y_t, y_t - y^*(x_t)\rangle + \eta_t^2\|\tilde{y}_{t+1} - y_t\|^2.\end{aligned} \tag{21}$$

Then we obtain

$$\langle \tilde{y}_{t+1} - y_t, y_t - y^*(x_t)\rangle = \frac{1}{2\eta_t}\|y_{t+1} - y^*(x_t)\|^2 - \frac{1}{2\eta_t}\|y_t - y^*(x_t)\|^2 - \frac{\eta_t}{2}\|\tilde{y}_{t+1} - y_t\|^2. \tag{22}$$

Consider the upper bound of the term $\langle \nabla_y g(x_t, y_t) - v_t, y^*(x_t) - \tilde{y}_{t+1}\rangle$, we have

$$\begin{aligned}&\langle \nabla_y g(x_t, y_t) - v_t, y^*(x_t) - \tilde{y}_{t+1}\rangle \\ &= \langle \nabla_y g(x_t, y_t) - v_t, y^*(x_t) - y_t\rangle + \langle \nabla_y g(x_t, y_t) - v_t, y_t - \tilde{y}_{t+1}\rangle \\ &\geq -\frac{1}{\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2 - \frac{\mu}{4}\|y^*(x_t) - y_t\|^2 - \frac{1}{\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2 - \frac{\mu}{4}\|y_t - \tilde{y}_{t+1}\|^2 \\ &= -\frac{2}{\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2 - \frac{\mu}{4}\|y^*(x_t) - y_t\|^2 - \frac{\mu}{4}\|y_t - \tilde{y}_{t+1}\|^2.\end{aligned} \tag{23}$$

By plugging the inequalities (22) and (23) into (20), we obtain

$$\begin{aligned}&\frac{1}{2\eta_t\lambda}\|y_{t+1} - y^*(x_t)\|^2 \\ &\leq (\frac{1}{2\eta_t\lambda} - \frac{\mu}{4})\|y_t - y^*(x_t)\|^2 + (\frac{\eta_t}{2\lambda} + \frac{\mu}{4} + \frac{L}{2} - \frac{1}{\lambda})\|\tilde{y}_{t+1} - y_t\|^2 + \frac{2}{\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2 \\ &\leq (\frac{1}{2\eta_t\lambda} - \frac{\mu}{4})\|y_t - y^*(x_t)\|^2 + (\frac{3L}{4} - \frac{1}{2\lambda})\|\tilde{y}_{t+1} - y_t\|^2 + \frac{2}{\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2 \\ &= (\frac{1}{2\eta_t\lambda} - \frac{\mu}{4})\|y_t - y^*(x_t)\|^2 - \big(\frac{3}{8\lambda} + \frac{1}{8\lambda} - \frac{3L}{4}\big)\|\tilde{y}_{t+1} - y_t\|^2 + \frac{2}{\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2 \\ &\leq (\frac{1}{2\eta_t\lambda} - \frac{\mu}{4})\|y_t - y^*(x_t)\|^2 - \frac{3}{8\lambda}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{2}{\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2,\end{aligned} \tag{24}$$

where the second inequality holds by $L \geq \mu$ and $0 < \eta_t \leq 1$, and the last inequality is due to $0 < \lambda \leq \frac{1}{6L}$. It implies that

$$\|y_{t+1} - y^*(x_t)\|^2 \leq (1 - \frac{\eta_t\mu\lambda}{2})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{4\eta_t\lambda}{\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2. \tag{25}$$

Next, we decompose the term $\|y_{t+1} - y^*(x_{t+1})\|^2$ as follows:

$$\begin{aligned}\|y_{t+1} - y^*(x_{t+1})\|^2 &= \|y_{t+1} - y^*(x_t) + y^*(x_t) - y^*(x_{t+1})\|^2 \\ &= \|y_{t+1} - y^*(x_t)\|^2 + 2\langle y_{t+1} - y^*(x_t), y^*(x_t) - y^*(x_{t+1})\rangle + \|y^*(x_t) - y^*(x_{t+1})\|^2 \\ &\leq (1 + \frac{\eta_t\mu\lambda}{4})\|y_{t+1} - y^*(x_t)\|^2 + (1 + \frac{4}{\eta_t\mu\lambda})\|y^*(x_t) - y^*(x_{t+1})\|^2 \\ &\leq (1 + \frac{\eta_t\mu\lambda}{4})\|y_{t+1} - y^*(x_t)\|^2 + (1 + \frac{4}{\eta_t\mu\lambda})\kappa^2\|x_t - x_{t+1}\|^2,\end{aligned} \tag{26}$$

where the first inequality holds by Cauchy-Schwarz inequality and Young's inequality, and the second inequality is due to Lemma 2, and the last equality holds by $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$.

By combining the above inequalities (25) and (26), we have

$$\begin{aligned}\|y_{t+1} - y^*(x_{t+1})\|^2 \leq (1 + \frac{\eta_t\mu\lambda}{4})(1 - \frac{\eta_t\mu\lambda}{2})\|y_t - y^*(x_t)\|^2 - (1 + \frac{\eta_t\mu\lambda}{4})\frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 \\ + (1 + \frac{\eta_t\mu\lambda}{4})\frac{4\eta_t\lambda}{\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2 + (1 + \frac{4}{\eta_t\mu\lambda})\kappa^2\|x_t - x_{t+1}\|^2.\end{aligned}$$

Since $0 < \eta_t \leq 1$, $0 < \lambda \leq \frac{1}{6L}$ and $L \geq \mu$, we have $\lambda \leq \frac{1}{6L} \leq \frac{1}{6\mu}$ and $\eta_t \leq 1 \leq \frac{1}{6\mu\lambda}$. Then by using ≥ 1, we have

$$\begin{aligned}(1 + \frac{\eta_t\mu\lambda}{4})(1 - \frac{\eta_t\mu\lambda}{2}) &= 1 - \frac{\eta_t\mu\lambda}{2} + \frac{\eta_t\mu\lambda}{4} - \frac{\eta_t^2\mu^2\lambda^2}{8} \leq 1 - \frac{\eta_t\mu\lambda}{4}, \\ -(1 + \frac{\eta_t\mu\lambda}{4})\frac{3\eta_t}{4} &\leq -\frac{3\eta_t}{4}, \\ (1 + \frac{\eta_t\mu\lambda}{4})\frac{4\eta_t\lambda}{\mu} &\leq (1 + \frac{1}{24})\frac{4\eta_t\lambda}{\mu} = \frac{25\eta_t\lambda}{6\mu}, \\ (1 + \frac{4}{\eta_t\mu\lambda})\kappa^2 &\leq \kappa^2 + \frac{4\kappa^2}{\eta_t\mu\lambda} \leq \frac{\kappa^2}{6\eta_t\mu\lambda} + \frac{4\kappa^2}{\eta_t\mu\lambda} = \frac{25\kappa^2}{6\eta_t\mu\lambda}.\end{aligned}$$

Thus we have

$$\begin{aligned}\|y_{t+1} - y^*(x_{t+1})\|^2 \leq (1 - \frac{\eta_t\mu\lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t\lambda^2}{4}\|v_t\|^2 \\ + \frac{25\eta_t\lambda}{6\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2 + \frac{25\kappa^2}{6\eta_t\mu\lambda}\|x_{t+1} - x_t\|^2. \end{aligned} \tag{27}$$

□

B.1 Convergence Analysis of the BiO-BreD Algorithm

In this subsection, we provide the convergence analysis of our BiO-BreD algorithm.

Theorem 4. *(Restatement of Theorem 1) Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 1. Let $0 < \gamma \leq \frac{3\rho}{4L_F}$, $0 < \lambda < \frac{1}{L}$, $K = \log(T)/\log(\frac{1}{1-\lambda\mu}) + 1$ and $\|y_t^0 - y^*(x_t)\|^2 \leq \Delta$ for all $t \geq 0$, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{G}_t\|^2 \leq \frac{16\big(\Phi(x_0) - \Phi^*\big)}{3T\gamma\rho} + \frac{22\Delta L_1^2}{\rho^2 T} + \frac{22\Delta L_2^2}{\rho^2 T} + \frac{22L_3^2}{\rho^2 T^2}, \tag{28}$$

where $L_1 = \frac{L(L+\mu)}{\mu}$, $L_2 = \frac{2C_{fy}(\mu L_{gxy} + LL_{gyy})}{\mu^2}$ and $L_3 = \frac{LC_{fy}}{\mu}$.

Proof. According to the above Lemma 2, the function $F(x)$ has L_F-Lipschitz continuous gradient. Let $\tilde{\mathcal{G}}_t = \frac{1}{\gamma}(x_t - x_{t+1})$, we have

$$\begin{aligned}F(x_{t+1}) &\leq F(x_t) + \langle\nabla F(x_t), x_{t+1} - x_t\rangle + \frac{L_F}{2}\|x_{t+1} - x_t\|^2 \\ &= F(x_t) - \gamma\langle\nabla F(x_t), \tilde{\mathcal{G}}_t\rangle + \frac{\gamma^2 L_F}{2}\|\tilde{\mathcal{G}}_t\|^2 \\ &= F(x_t) - \gamma\langle\frac{\partial f(x_t, y_t^K)}{\partial x}, \tilde{\mathcal{G}}_t\rangle + \gamma\langle\frac{\partial f(x_t, y_t^K)}{\partial x} - \nabla F(x_t), \tilde{\mathcal{G}}_t\rangle + \frac{\gamma^2 L_F}{2}\|\tilde{\mathcal{G}}_t\|^2 \\ &\leq F(x_t) - \gamma\rho\|\tilde{\mathcal{G}}_t\|^2 - h(x_{t+1}) + h(x_t) + \gamma\langle\frac{\partial f(x_t, y_t^K)}{\partial x} - \nabla F(x_t), \tilde{\mathcal{G}}_t\rangle + \frac{\gamma^2 L_F}{2}\|\tilde{\mathcal{G}}_t\|^2 \\ &\leq F(x_t) + (\frac{\gamma^2 L_F}{2} - \frac{3\gamma\rho}{4})\|\tilde{\mathcal{G}}_t\|^2 - h(x_{t+1}) + h(x_t) + \frac{\gamma}{\rho}\|\frac{\partial f(x_t, y_t^K)}{\partial x} - \nabla F(x_t)\|^2, \end{aligned} \tag{29}$$

where the second last inequality holds by the above Lemma 9, and the last inequality holds by the following inequality

$$\begin{aligned}\langle\frac{\partial f(x_t, y_t^K)}{\partial x} - \nabla F(x_t), \tilde{\mathcal{G}}_t\rangle &\leq \|\frac{\partial f(x_t, y_t^K)}{\partial x} - \nabla F(x_t)\|\|\tilde{\mathcal{G}}_t\| \\ &\leq \frac{1}{\rho}\|\frac{\partial f(x_t, y_t^K)}{\partial x} - \nabla F(x_t)\|^2 + \frac{\rho}{4}\|\tilde{\mathcal{G}}_t\|^2. \end{aligned} \tag{30}$$

According to the above Lemma 7, we have

$$\begin{aligned}
&\|\frac{\partial f(x_t, y_t^K)}{\partial x} - \nabla F(x_t)\|^2 \\
&\leq 3\big(L_1^2(1-\lambda\mu)^K + L_2^2(1-\lambda\mu)^{K-1}\big)\|y_t^0 - y^*(x_t)\|^2 + 3L_3^2(1-\lambda\mu)^{2K} \\
&\leq 3\Delta L_1^2(1-\lambda\mu)^K + 3\Delta L_2^2(1-\lambda\mu)^{K-1} + 3L_3^2(1-\lambda\mu)^{2K},
\end{aligned} \tag{31}$$

where the last inequality holds by $\|y_t^0 - y^*(x_t)\|^2 \leq \Delta$ for all $t \geq 0$.

Let $\Phi(x) = F(x) + h(x)$, plugging (31) into (29), we have

$$\begin{aligned}
\Phi(x_{t+1}) &\leq \Phi(x_t) + (\frac{\gamma^2 L_F}{2} - \frac{3\gamma\rho}{4})\|\tilde{\mathcal{G}}_t\|^2 + \frac{\gamma}{\rho}\|\frac{\partial f(x_t, y_t^K)}{\partial x} - \nabla F(x_t)\|^2 \\
&\leq \Phi(x_t) - \frac{3\gamma\rho}{8}\|\tilde{\mathcal{G}}_t\|^2 + \frac{3\gamma\Delta L_1^2}{\rho}(1-\lambda\mu)^K + \frac{3\gamma\Delta L_2^2}{\rho}(1-\lambda\mu)^{K-1} + \frac{3\gamma L_3^2}{\rho}(1-\lambda\mu)^{2K},
\end{aligned} \tag{32}$$

where the last inequality is due to $0 < \gamma \leq \frac{3\rho}{4L_F}$ and the above inequality (31). According to Lemma 10, the difference between $\tilde{\mathcal{G}}_t$ and $\mathcal{G}_t$ are bounded, we have

$$\begin{aligned}
\|\mathcal{G}_t\|^2 &\leq 2\|\tilde{\mathcal{G}}_t\|^2 + 2\|\tilde{\mathcal{G}}_t - \mathcal{G}_t\|^2 \\
&\leq 2\|\tilde{\mathcal{G}}_t\|^2 + \frac{2}{\rho^2}\|w_t - \nabla F(x_t)\|^2 \\
&= 2\|\tilde{\mathcal{G}}_t\|^2 + \frac{2}{\rho^2}\|\frac{\partial f(x_t, y_t^K)}{\partial x} - \nabla F(x_t)\|^2 \\
&\leq 2\|\tilde{\mathcal{G}}_t\|^2 + \frac{6\Delta L_1^2}{\rho^2}(1-\lambda\mu)^K + \frac{6\Delta L_2^2}{\rho^2}(1-\lambda\mu)^{K-1} + \frac{6L_3^2}{\rho^2}(1-\lambda\mu)^{2K}.
\end{aligned} \tag{33}$$

Thus we have

$$-\|\tilde{\mathcal{G}}_t\|^2 \leq -\frac{1}{2}\|\mathcal{G}_t\|^2 + \frac{3\Delta L_1^2}{\rho^2}(1-\lambda\mu)^K + \frac{3\Delta L_2^2}{\rho^2}(1-\lambda\mu)^{K-1} + \frac{3L_3^2}{\rho^2}(1-\lambda\mu)^{2K}. \tag{34}$$

By plugging (34) into (29), we have

$$\begin{aligned}
\Phi(x_{t+1}) &\leq \Phi(x_t) - \frac{3\gamma\rho}{16}\|\mathcal{G}_t\|^2 + \frac{3\gamma\rho}{8}\big(\frac{3\Delta L_1^2}{\rho^2}(1-\lambda\mu)^K + \frac{3\Delta L_2^2}{\rho^2}(1-\lambda\mu)^{K-1} + \frac{3L_3^2}{\rho^2}(1-\lambda\mu)^{2K}\big) \\
&\quad + \frac{3\gamma\Delta L_1^2}{\rho}(1-\lambda\mu)^K + \frac{3\gamma\Delta L_2^2}{\rho}(1-\lambda\mu)^{K-1} + \frac{3\gamma L_3^2}{\rho}(1-\lambda\mu)^{2K} \\
&= \Phi(x_t) - \frac{3\gamma\rho}{16}\|\mathcal{G}_t\|^2 + \frac{33\gamma\Delta L_1^2}{8\rho}(1-\lambda\mu)^K + \frac{33\gamma\Delta L_2^2}{8\rho}(1-\lambda\mu)^{K-1} + \frac{33\gamma L_3^2}{8\rho}(1-\lambda\mu)^{2K}.
\end{aligned} \tag{35}$$

Thus, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\|\mathcal{G}_t\|^2 &\leq \frac{16\big(\Phi(x_0) - \Phi(x_T)\big)}{3T\gamma\rho} + \frac{22\Delta L_1^2}{\rho^2}(1-\lambda\mu)^K + \frac{22\Delta L_2^2}{\rho^2}(1-\lambda\mu)^{K-1} + \frac{22L_3^2}{\rho^2}(1-\lambda\mu)^{2K} \\
&\leq \frac{16\big(\Phi(x_0) - \Phi^*\big)}{3T\gamma\rho} + \frac{22\Delta L_1^2}{\rho^2}(1-\lambda\mu)^K + \frac{22\Delta L_2^2}{\rho^2}(1-\lambda\mu)^{K-1} + \frac{22L_3^2}{\rho^2}(1-\lambda\mu)^{2K} \\
&\leq \frac{16\big(\Phi(x_0) - \Phi^*\big)}{3T\gamma\rho} + \frac{22\Delta L_1^2}{\rho^2 T} + \frac{22\Delta L_2^2}{\rho^2 T} + \frac{22L_3^2}{\rho^2 T^2},
\end{aligned} \tag{36}$$

where the last inequality holds by $K = \log(T)/\log(\frac{1}{1-\lambda\mu}) + 1$ and $\lambda < \frac{1}{L}$.

□

B.2 Convergence Analysis of the SBiO-BreD Algorithm

In this subsection, we provide the convergence analysis of our SBiO-BreD algorithm. Let $R(x_t, y_t) = \bar{\nabla} f(x_t, y_t) - \bar{\nabla} f(x_t, y_t; \mathcal{B}_t)$ for all $t \geq 0$.

Theorem 5. *(Restatement of Theorem 2) Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 2. Let $K = \frac{L}{\mu}\log(\frac{LC_{fy}T}{\mu})$, $0 < \eta = \eta_t \leq 1$, $0 < \gamma \leq \min\big(\frac{3\rho}{4L_F}, \frac{9\eta\rho\mu\lambda}{800\kappa^2}, \frac{\eta\mu\rho\lambda}{47L_y^2}\big)$ and $0 < \lambda \leq \frac{1}{6L}$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}_t\|^2 \leq \frac{32(\Phi(x_0) - \Phi^*)}{3T\gamma\rho} + \frac{32\Delta}{3T\gamma\rho} + \frac{752\sigma^2}{3\rho^2 b} + \frac{400\eta\lambda\sigma^2}{9\gamma\rho\mu b} + \frac{752}{3\rho^2 T^2}, \tag{37}$$

where $\Delta = \|y_0 - y^*(x_0)\|^2$.

Proof. According to the above Lemma 2, the function $F(x)$ has L_F-Lipschitz continuous gradient. Let $\tilde{\mathcal{G}}_t = \frac{1}{\gamma}(x_t - x_{t+1})$, we have

$$\begin{aligned}
F(x_{t+1}) &\leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{L_F}{2}\|x_{t+1} - x_t\|^2 \\
&= F(x_t) - \gamma \langle \nabla F(x_t), \tilde{\mathcal{G}}_t \rangle + \frac{\gamma^2 L_F}{2}\|\tilde{\mathcal{G}}_t\|^2 \\
&= F(x_t) - \gamma \langle w_t, \tilde{\mathcal{G}}_t \rangle + \gamma \langle w_t - \nabla F(x_t), \tilde{\mathcal{G}}_t \rangle + \frac{\gamma^2 L_F}{2}\|\tilde{\mathcal{G}}_t\|^2 \\
&\leq F(x_t) - \gamma\rho\|\tilde{\mathcal{G}}_t\|^2 - h(x_{t+1}) + h(x_t) + \gamma \langle w_t - \nabla F(x_t), \tilde{\mathcal{G}}_t \rangle + \frac{\gamma^2 L_F}{2}\|\tilde{\mathcal{G}}_t\|^2 \\
&\leq F(x_t) + (\frac{\gamma^2 L_F}{2} - \frac{3\gamma\rho}{4})\|\tilde{\mathcal{G}}_t\|^2 - h(x_{t+1}) + h(x_t) + \frac{\gamma}{\rho}\|w_t - \nabla F(x_t)\|^2, \qquad (38)
\end{aligned}$$

where the second last inequality holds by the above Lemma 9, and the last inequality holds by the following inequality

$$\begin{aligned}
\langle w_t - \nabla F(x_t), \tilde{\mathcal{G}}_t \rangle &\leq \|w_t - \nabla F(x_t)\|\|\tilde{\mathcal{G}}_t\| \\
&\leq \frac{1}{\rho}\|w_t - \nabla F(x_t)\|^2 + \frac{\rho}{4}\|\tilde{\mathcal{G}}_t\|^2. \qquad (39)
\end{aligned}$$

According to the above Lemma 2, we have

$$\begin{aligned}
\|w_t - \nabla F(x_t)\|^2 &= \|w_t - \bar{\nabla} f(x_t, y_t) + \bar{\nabla} f(x_t, y_t) - \nabla F(x_t)\|^2 \\
&\leq 2\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + 2\|\bar{\nabla} f(x_t, y_t) - \nabla F(x_t)\|^2 \\
&\leq 2\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + 2L_y^2\|y_t - y^*(x_t)\|^2. \qquad (40)
\end{aligned}$$

Let $\Phi(x) = F(x) + h(x)$, plugging (40) into (38), we have

$$\begin{aligned}
\Phi(x_{t+1}) &\leq \Phi(x_t) + (\frac{\gamma^2 L_F}{2} - \frac{3\gamma\rho}{4})\|\tilde{\mathcal{G}}_t\|^2 + \frac{2\gamma}{\rho}\|w_t - \nabla_x f(x_t, y_t)\|^2 + \frac{2L_y^2\gamma}{\rho}\|y_t - y^*(x_t)\|^2 \\
&\leq \Phi(x_t) - \frac{3\gamma\rho}{8}\|\tilde{\mathcal{G}}_t\|^2 + \frac{2\gamma}{\rho}\|w_t - \nabla_x f(x_t, y_t)\|^2 + \frac{2L_y^2\gamma}{\rho}\|y_t - y^*(x_t)\|^2, \qquad (41)
\end{aligned}$$

where the last inequality is due to $0 < \gamma \leq \frac{3\rho}{4L_F}$. According to Lemma 10, the difference between $\tilde{\mathcal{G}}_t$ and $\mathcal{G}_t$ are bounded, we have

$$\begin{aligned}
\|\mathcal{G}_t\|^2 &\leq 2\|\tilde{\mathcal{G}}_t\|^2 + 2\|\tilde{\mathcal{G}}_t - \mathcal{G}_t\|^2 \\
&\leq 2\|\tilde{\mathcal{G}}_t\|^2 + \frac{2}{\rho^2}\|w_t - \nabla F(x_t)\|^2 \\
&\leq 2\|\tilde{\mathcal{G}}_t\|^2 + \frac{4}{\rho^2}\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + \frac{4L_y^2}{\rho^2}\|y_t - y^*(x_t)\|^2. \qquad (42)
\end{aligned}$$

Thus we have

$$-\|\tilde{\mathcal{G}}_t\|^2 \leq -\frac{1}{2}\|\mathcal{G}_t\|^2 + \frac{2}{\rho^2}\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + \frac{2L_y^2}{\rho^2}\|y_t - y^*(x_t)\|^2. \qquad (43)$$

By plugging (43) into (41), we have

$$\begin{aligned}
\Phi(x_{t+1}) &\leq \Phi(x_t) - \frac{3\gamma\rho}{16}\|\mathcal{G}_t\|^2 + \frac{3\gamma\rho}{8}\big(\frac{2}{\rho^2}\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + \frac{2L_y^2}{\rho^2}\|y_t - y^*(x_t)\|^2\big) \\
&\quad + \frac{2\gamma}{\rho}\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + \frac{2L_y^2\gamma}{\rho}\|y_t - y^*(x_t)\|^2 \\
&= \Phi(x_t) - \frac{3\gamma\rho}{16}\|\mathcal{G}_t\|^2 + \frac{11\gamma}{4\rho}\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + \frac{11L_y^2\gamma}{4\rho}\|y_t - y^*(x_t)\|^2. \qquad (44)
\end{aligned}$$

Next, we define a useful Lyapunov function, for any $t \geq 1$

$$\Omega_t = \mathbb{E}\big[\Phi(x_t) + \|y_t - y^*(x_t)\|^2\big]. \qquad (45)$$

According to Lemma 11, we have

$$\|y_{t+1}-y^*(x_{t+1})\|^2-\|y_t-y^*(x_t)\|^2\le-\frac{\eta_t\mu\lambda}{4}\|y_t-y^*(x_t)\|^2-\frac{3\eta_t\lambda^2}{4}\|v_t\|^2 + \frac{25\eta_t\lambda}{6\mu}\|\nabla_y g(x_t,y_t)-v_t\|^2+\frac{25\kappa^2}{6\eta_t\mu\lambda}\|x_{t+1}-x_t\|^2. \quad (46)$$

Then we have

$$\begin{aligned}
\Omega_{t+1}-\Omega_t &= \mathbb{E}\big[\Phi(x_{t+1})-\Phi(x_t)+\|y_{t+1}-y^*(x_{t+1})\|^2-\|y_t-y^*(x_t)\|^2\big]\\
&\le -\frac{3\gamma\rho}{16}\mathbb{E}\|\mathcal{G}_t\|^2+\frac{11\gamma}{4\rho}\mathbb{E}\|w_t-\bar\nabla f(x_t,y_t)\|^2+\frac{11L_y^2\gamma}{4\rho}\mathbb{E}\|y_t-y^*(x_t)\|^2-\frac{\eta_t\mu\lambda}{4}\mathbb{E}\|y_t-y^*(x_t)\|^2\\
&\quad-\frac{3\eta_t\lambda^2}{4}\mathbb{E}\|v_t\|^2+\frac{25\eta_t\lambda}{6\mu}\mathbb{E}\|\nabla_y g(x_t,y_t)-v_t\|^2+\frac{25\kappa^2}{6\eta_t\mu\lambda}\mathbb{E}\|x_{t+1}-x_t\|^2\\
&= -\frac{3\gamma\rho}{16}\mathbb{E}\|\mathcal{G}_t\|^2+\frac{11\gamma}{4\rho}\mathbb{E}\|w_t-\bar\nabla f(x_t,y_t)\|^2+\frac{11L_y^2\gamma}{4\rho}\mathbb{E}\|y_t-y^*(x_t)\|^2-\frac{\eta_t\mu\lambda}{4}\mathbb{E}\|y_t-y^*(x_t)\|^2\\
&\quad-\frac{3\eta_t\lambda^2}{4}\mathbb{E}\|v_t\|^2+\frac{25\eta_t\lambda}{6\mu}\mathbb{E}\|\nabla_y g(x_t,y_t)-v_t\|^2+\frac{25\kappa^2\gamma^2}{6\eta_t\mu\lambda}\mathbb{E}\|\tilde{\mathcal{G}}_t\|^2\\
&\le -(\frac{3\gamma\rho}{16}-\frac{25\kappa^2\gamma^2}{3\eta_t\mu\lambda})\mathbb{E}\|\mathcal{G}_t\|^2+(\frac{11\gamma}{4\rho}+\frac{50\kappa^2\gamma^2}{3\eta_t\mu\lambda\rho^2})\mathbb{E}\|w_t-\bar\nabla f(x_t,y_t)\|^2\\
&\quad+(\frac{11L_y^2\gamma}{4\rho}+\frac{50\kappa^2\gamma^2L_y^2}{3\eta_t\mu\lambda\rho^2}-\frac{\eta_t\mu\lambda}{4})\mathbb{E}\|y_t-y^*(x_t)\|^2-\frac{3\eta_t\lambda^2}{4}\mathbb{E}\|v_t\|^2\\
&\quad+\frac{25\eta_t\lambda}{6\mu}\mathbb{E}\|\nabla_y g(x_t,y_t)-v_t\|^2, \qquad (47)
\end{aligned}$$

where the last inequality holds by the following inequality

$$\begin{aligned}
\|\tilde{\mathcal{G}}_t\|^2 &\le 2\|\mathcal{G}_t\|^2+2\|\tilde{\mathcal{G}}_t-\mathcal{G}_t\|^2\\
&\le 2\|\mathcal{G}_t\|^2+\frac{2}{\rho^2}\|w_t-\nabla F(x_t)\|^2\\
&\le 2\|\mathcal{G}_t\|^2+\frac{4}{\rho^2}\|w_t-\bar\nabla f(x_t,y_t)\|^2+\frac{4L_y^2}{\rho^2}\|y_t-y^*(x_t)\|^2. \qquad (48)
\end{aligned}$$

Let $\eta=\eta_t$ for all $t\ge 0$. By using $0<\gamma\le\frac{9\eta\rho\mu\lambda}{800\kappa^2}$, we have

$$\frac{3\gamma\rho}{32}\ge\frac{50\kappa^2\gamma^2}{6\eta\lambda\mu},\quad \frac{3L_y^2\gamma}{16\rho}\ge\frac{50\kappa^2\gamma^2L_y^2}{3\eta_t\mu\lambda\rho^2},\quad \frac{3\gamma}{16\rho}\ge\frac{50\kappa^2\gamma^2}{3\eta_t\mu\lambda\rho^2}. \quad (49)$$

Let $\frac{\eta\mu\lambda}{4}\ge\frac{47L_y^2\gamma}{4\rho}$, we have $0<\gamma\le\frac{\eta\mu\rho\lambda}{47L_y^2}$. Given $0<\gamma\le\min\big(\frac{9\eta\rho\mu\lambda}{800\kappa^2},\frac{\eta\mu\rho\lambda}{47L_y^2}\big)$, we have

$$\Omega_{t+1}-\Omega_t\le-\frac{3\gamma\rho}{32}\mathbb{E}\|\mathcal{G}_t\|^2+\frac{47\gamma}{4\rho}\mathbb{E}\|w_t-\bar\nabla f(x_t,y_t)\|^2+\frac{25\eta\lambda}{6\mu}\mathbb{E}\|v_t-\nabla_y g(x_t,y_t)\|^2. \quad (50)$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}\|\mathcal{G}_t\|^2\\
&\le\frac{32(\Omega_t-\Omega_{t+1})}{3\gamma\rho}+\frac{376}{3\rho^2}\mathbb{E}\|w_t-\bar\nabla f(x_t,y_t)\|^2+\frac{400\eta\lambda}{9\gamma\rho\mu}\mathbb{E}\|v_t-\nabla_y g(x_t,y_t)\|^2\\
&=\frac{32(\Omega_t-\Omega_{t+1})}{3\gamma\rho}+\frac{376}{3\rho^2}\mathbb{E}\|w_t-\bar\nabla f(x_t,y_t)-R(x_t,y_t)+R(x_t,y_t)\|^2+\frac{400\eta\lambda}{9\gamma\rho\mu}\mathbb{E}\|v_t-\nabla_y g(x_t,y_t)\|^2\\
&\le\frac{32(\Omega_t-\Omega_{t+1})}{3\gamma\rho}+\frac{752}{3\rho^2}\mathbb{E}\|w_t-\bar\nabla f(x_t,y_t)-R(x_t,y_t)\|^2+\frac{752}{3\rho^2}\|R(x_t,y_t)\|^2+\frac{400\eta\lambda}{9\gamma\rho\mu}\mathbb{E}\|v_t-\nabla_y g(x_t,y_t)\|^2\\
&\le\frac{32(\Omega_t-\Omega_{t+1})}{3\gamma\rho}+\frac{752\sigma^2}{3\rho^2b}+\frac{400\eta\lambda\sigma^2}{9\gamma\rho\mu b}+\frac{752}{3\rho^2}\mathbb{E}\|R(x_t,y_t)\|^2, \qquad (51)
\end{aligned}$$

where the last inequality holds by Assumption 6 and $w_t=\bar\nabla f(x_t,y_t;\mathcal{B}_t)=\frac{1}{b}\sum_{i\in\mathcal{B}_t}\bar\nabla f(x_t,y_t,\bar\xi_t^i)$, $v_t=\nabla_y g(x_t,y_t;\mathcal{B}_t)=\frac{1}{b}\sum_{i\in\mathcal{B}_t}\nabla_y g(x_t,y_t,\xi_t^i)$.

Taking average over $t = 0, 2, \cdots, T-1$ on both sides of the above inequality (51), we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathcal{G}_t\|^2 &\leq \frac{32(\Omega_0 - \Omega_T)}{3T\gamma\rho} + \frac{752\sigma^2}{3\rho^2 b} + \frac{400\eta\lambda\sigma^2}{9\gamma\rho\mu b} + \frac{752}{3\rho^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|R(x_t, y_t)\|^2 \\ &= \frac{32(\Phi(x_0) + \|y_0 - y^*(x_0)\|^2)}{3T\gamma\rho} - \frac{32\mathbb{E}(\Phi(x_T) + \|y_T - y^*(x_T)\|^2)}{3T\gamma\rho} \\ &\quad + \frac{752\sigma^2}{3\rho^2 b} + \frac{400\eta\lambda\sigma^2}{9\gamma\rho\mu b} + \frac{752}{3\rho^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|R(x_t, y_t)\|^2 \\ &\leq \frac{32(\Phi(x_0) - \Phi^*)}{3T\gamma\rho} + \frac{32\Delta}{3T\gamma\rho} + \frac{752\sigma^2}{3\rho^2 b} + \frac{400\eta\lambda\sigma^2}{9\gamma\rho\mu b} + \frac{752}{3\rho^2 T^2}, \end{aligned} \tag{52}$$

where the last inequality holds by Assumption 5 and $\mathbb{E}\|R(x_t, y_t)\| \leq \frac{1}{T}$ for all $t \geq 1$ by choosing $K = \frac{L}{\mu}\log(\frac{LC_{fy}T}{\mu})$.

□

B.3 Convergence Analysis of the ASBiO-BreD Algorithm

In this subsection, we provide the convergence analysis of our ASBiO-BreD algorithm. When $\mod(t, q) \neq 0$, let $R(x_t, y_t) = \bar{\nabla} f(x_t, y_t) - \bar{\nabla} f(x_t, y_t; \bar{\mathcal{I}}_t)$ for all $t \geq 0$, when $\mod(t, q) = 0$, let $R(x_t, y_t) = \bar{\nabla} f(x_t, y_t) - \bar{\nabla} f(x_t, y_t; \bar{\mathcal{B}}_t)$.

Lemma 12. *Suppose the stochastic gradients v_t and w_t be generated from Algorithm 3, we have*

$$\mathbb{E}\|\bar{\nabla} f(x_t, y_t) + R(x_t, y_t) - w_t\|^2 \leq \frac{2L_K^2}{b_1}\sum_{i=(n_t-1)q}^{t-1}\big(\mathbb{E}\|x_{i+1} - x_i\|^2 + \mathbb{E}\|y_{i+1} - y_i\|^2\big) + \frac{\sigma^2}{b}, \tag{53}$$

$$\mathbb{E}\|\nabla_y g(x_t, y_t) - v_t\|^2 \leq \frac{2L^2}{b_1}\sum_{i=(n_t-1)q}^{t-1}\big(\mathbb{E}\|x_{i+1} - x_i\|^2 + \mathbb{E}\|y_{i+1} - y_i\|^2\big) + \frac{\sigma^2}{b}. \tag{54}$$

Proof. We first prove the inequality (53). According to the definition of w_{t-1} in Algorithm 3, we have

$$w_t - w_{t-1} = \bar{\nabla} f(x_t, y_t; \bar{\mathcal{I}}_t) - \bar{\nabla} f(x_{t-1}, y_{t-1}; \bar{\mathcal{I}}_t). \tag{55}$$

Then we have

$$\begin{aligned}&\mathbb{E}\|\bar{\nabla} f(x_t, y_t) + R(x_t, y_t) - w_t\|^2 \\ &= \mathbb{E}\|\bar{\nabla} f(x_t, y_t) + R(x_t, y_t) - w_{t-1} - (w_t - w_{t-1})\|^2 \\ &= \mathbb{E}\|\bar{\nabla} f(x_t, y_t) + R(x_t, y_t) - w_{t-1} - \bar{\nabla} f(x_t, y_t; \bar{\mathcal{I}}_t) + \bar{\nabla} f(x_{t-1}, y_{t-1}; \bar{\mathcal{I}}_t)\|^2 \\ &= \mathbb{E}\|\bar{\nabla} f(x_{t-1}, y_{t-1}) + R(x_{t-1}, y_{t-1}) - w_{t-1} + \bar{\nabla} f(x_t, y_t) + R(x_t, y_t) - \bar{\nabla} f(x_{t-1}, y_{t-1}) - R(x_{t-1}, y_{t-1}) \\ &\quad - \bar{\nabla} f(x_t, y_t; \bar{\mathcal{I}}_t) + \bar{\nabla} f(x_{t-1}, y_{t-1}; \bar{\mathcal{I}}_t)\|^2 \\ &= \mathbb{E}\|\bar{\nabla} f(x_{t-1}, y_{t-1}) + R(x_{t-1}, y_{t-1}) - w_{t-1}\|^2 + \mathbb{E}\|\bar{\nabla} f(x_t, y_t) + R(x_t, y_t) - \bar{\nabla} f(x_{t-1}, y_{t-1}) - R(x_{t-1}, y_{t-1}) \\ &\quad - \bar{\nabla} f(x_t, y_t; \bar{\mathcal{I}}_t) + \bar{\nabla} f(x_{t-1}, y_{t-1}; \bar{\mathcal{I}}_t)\|^2 \\ &= \mathbb{E}\|\bar{\nabla} f(x_{t-1}, y_{t-1}) + R(x_{t-1}, y_{t-1}) - w_{t-1}\|^2 + \frac{1}{b_1}\mathbb{E}\|\bar{\nabla} f(x_t, y_t) + R(x_t, y_t) - \bar{\nabla} f(x_{t-1}, y_{t-1}) - R(x_{t-1}, y_{t-1}) \\ &\quad - \big(\bar{\nabla} f(x_t, y_t; \xi_t^1) - \bar{\nabla} f(x_{t-1}, y_{t-1}; \xi_t^1)\big)\|^2 \\ &\leq \mathbb{E}\|\bar{\nabla} f(x_{t-1}, y_{t-1}) + R(x_{t-1}, y_{t-1}) - w_{t-1}\|^2 + \frac{1}{b_1}\mathbb{E}\|\bar{\nabla} f(x_t, y_t; \xi_t^1) - \bar{\nabla} f(x_{t-1}, y_{t-1}; \xi_t^1)\|^2 \\ &\leq \mathbb{E}\|\bar{\nabla} f(x_{t-1}, y_{t-1}) + R(x_{t-1}, y_{t-1}) - w_{t-1}\|^2 + \frac{2L_K^2}{b_1}\big(\|x_t - x_{t-1}\|^2 + \|y_t - y_{t-1}\|^2\big), \end{aligned} \tag{56}$$

where the fourth equality follows by $\mathbb{E}_{\bar{\mathcal{I}}_t}\big[\bar{\nabla} f(x_t, y_t) + R(x_t, y_t) - \bar{\nabla} f(x_{t-1}, y_{t-1}) - R(x_{t-1}, y_{t-1}) - \big(\bar{\nabla} f(x_t, y_t; \bar{\mathcal{I}}_t) - \bar{\nabla} f(x_{t-1}, y_{t-1}; \bar{\mathcal{I}}_t)\big)\big] = 0$; the fifth equality holds by Lemma 8 and $\bar{\nabla} f(x_t, y_t; \bar{\mathcal{I}}_t) = \frac{1}{b_1}\sum_{i\in\bar{\mathcal{I}}_t}\bar{\nabla} f(x_t, y_t; \bar{\xi}_t^i)$, $\bar{\nabla} f(x_{t-1}, y_{t-1}; \bar{\mathcal{I}}_t) = \frac{1}{b_1}\sum_{i\in\bar{\mathcal{I}}_t}\bar{\nabla} f(x_{t-1}, y_{t-1}; \bar{\xi}_t^i)$; the second last inequality holds by the inequality $\mathbb{E}\|\zeta - \mathbb{E}[\zeta]\|^2 \leq \mathbb{E}\|\zeta\|^2$; the last inequality is due to Lemma 4.

Throughout the paper, let $n_t = [t/q]$ such that $(n_t - 1)q \leq t \leq n_t q - 1$. Telescoping (56) over t from $(n_t - 1)q + 1$ to t, we have

$$\begin{aligned}\mathbb{E}\|\bar{\nabla} f(x_t, y_t) + R(x_t, y_t) - w_t\|^2 &\leq \frac{2L_K^2}{b_1} \sum_{i=(n_t-1)q}^{t-1} \big(\mathbb{E}\|x_{i+1} - x_i\|^2 + \mathbb{E}\|y_{i+1} - y_i\|^2\big) \\ &\quad + \mathbb{E}\|\bar{\nabla} f(x_{(n_t-1)q}, y_{(n_t-1)q}) + R(x_{(n_t-1)q}, y_{(n_t-1)q}) - w_{(n_t-1)q}\|^2 \\ &\leq \frac{2L_K^2}{b_1} \sum_{i=(n_t-1)q}^{t-1} \big(\mathbb{E}\|x_{i+1} - x_i\|^2 + \mathbb{E}\|y_{i+1} - y_i\|^2\big) + \frac{\sigma^2}{b}, \end{aligned} \tag{57}$$

where the last inequality is due to Assumption 6 and $w_{(n_t-1)q} = \frac{1}{b}\sum_{i\in\mathcal{B}_{(n_t-1)q}} \bar{\nabla} f(x_{(n_t-1)q}, y_{(n_t-1)q}, \xi^i_{(n_t-1)q})$. Similarly, we can obtain

$$\mathbb{E}\|\nabla_y g(x_t, y_t) - v_t\|^2 \leq \frac{2L^2}{b_1} \sum_{i=(n_t-1)q}^{t-1} \big(\mathbb{E}\|x_{i+1} - x_i\|^2 + \mathbb{E}\|y_{i+1} - y_i\|^2\big) + \frac{\sigma^2}{b}. \tag{58}$$

□

Theorem 6. *(Restatement of Theorem 3) Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 3. Let $b_1 = q$, $K = \frac{L}{\mu}\log(\frac{LC_{fy}T}{\mu})$, $0 < \eta = \eta_t \leq 1$, $0 < \gamma \leq \min\big(\frac{3\rho}{38L_K^2\eta}, \frac{3\rho}{4L_F}, \frac{2\rho\eta\mu\lambda}{19L_y^2}, \frac{\rho\eta}{8}, \frac{9\rho\eta\mu\lambda}{400\kappa^2}\big)$ and $0 < \lambda \leq \min(\frac{1}{6L}, \frac{9\mu}{100\eta^2L^2})$, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathcal{G}_t\|^2 \leq \frac{32(\Phi(x_0) - \Phi^*)}{3T\gamma\rho} + \frac{32\Delta}{3T\gamma\rho} + \frac{152}{3T^2\rho^2} + \frac{4}{\eta\rho\gamma}\big(\frac{1}{L^2} + \frac{1}{L_K^2}\big)\frac{\sigma^2}{b}, \tag{59}$$

where $\Delta = \|y_0 - y^*(x_0)\|^2$.

Proof. This proof is similar to the proof of Theorem 5. According to the above Lemma 2, the function $F(x)$ has L_F-Lipschitz continuous gradient. Let $\tilde{\mathcal{G}}_t = \frac{1}{\gamma}(x_t - x_{t+1})$, we have

$$\begin{aligned} F(x_{t+1}) &\leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t\rangle + \frac{L_F}{2}\|x_{t+1} - x_t\|^2 \\ &= F(x_t) - \gamma\langle \nabla F(x_t), \tilde{\mathcal{G}}_t\rangle + \frac{\gamma^2 L_F}{2}\|\tilde{\mathcal{G}}_t\|^2 \\ &= F(x_t) - \gamma\langle w_t, \tilde{\mathcal{G}}_t\rangle + \gamma\langle w_t - \nabla F(x_t), \tilde{\mathcal{G}}_t\rangle + \frac{\gamma^2 L_F}{2}\|\tilde{\mathcal{G}}_t\|^2 \\ &\leq F(x_t) - \gamma\rho\|\tilde{\mathcal{G}}_t\|^2 - h(x_{t+1}) + h(x_t) + \gamma\langle w_t - \nabla F(x_t), \tilde{\mathcal{G}}_t\rangle + \frac{\gamma^2 L_F}{2}\|\tilde{\mathcal{G}}_t\|^2 \\ &\leq F(x_t) + (\frac{\gamma^2 L_F}{2} - \frac{3\gamma\rho}{4})\|\tilde{\mathcal{G}}_t\|^2 - h(x_{t+1}) + h(x_t) + \frac{\gamma}{\rho}\|w_t - \nabla F(x_t)\|^2, \end{aligned} \tag{60}$$

where the second last inequality holds by the above Lemma 9, and the last inequality holds by the following inequality

$$\begin{aligned}\langle w_t - \nabla F(x_t), \tilde{\mathcal{G}}_t\rangle &\leq \|w_t - \nabla F(x_t)\|\|\tilde{\mathcal{G}}_t\| \\ &\leq \frac{1}{\rho}\|w_t - \nabla F(x_t)\|^2 + \frac{\rho}{4}\|\tilde{\mathcal{G}}_t\|^2. \end{aligned} \tag{61}$$

According to the above Lemma 2, we have

$$\begin{aligned}\|w_t - \nabla F(x_t)\|^2 &= \|w_t - \bar{\nabla} f(x_t, y_t) + \bar{\nabla} f(x_t, y_t) - \nabla F(x_t)\|^2 \\ &\leq 2\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + 2\|\bar{\nabla} f(x_t, y_t) - \nabla F(x_t)\|^2 \\ &\leq 2\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + 2L_y^2\|y_t - y^*(x_t)\|^2. \end{aligned} \tag{62}$$

Let $\Phi(x) = F(x) + h(x)$, plugging (62) into (60), we have

$$\begin{aligned}\Phi(x_{t+1}) &\leq \Phi(x_t) + (\frac{\gamma^2 L_F}{2} - \frac{3\gamma\rho}{4})\|\tilde{\mathcal{G}}_t\|^2 + \frac{2\gamma}{\rho}\|w_t - \nabla_x f(x_t, y_t)\|^2 + \frac{2L_y^2\gamma}{\rho}\|y_t - y^*(x_t)\|^2 \\ &\leq \Phi(x_t) - \frac{3\gamma\rho}{8}\|\tilde{\mathcal{G}}_t\|^2 + \frac{2\gamma}{\rho}\|w_t - \nabla_x f(x_t, y_t)\|^2 + \frac{2L_y^2\gamma}{\rho}\|y_t - y^*(x_t)\|^2 \\ &= \Phi(x_t) - \frac{3\gamma\rho}{16}\|\tilde{\mathcal{G}}_t\|^2 - \frac{3\rho}{16\gamma}\|x_{t+1} - x_t\|^2 + \frac{2\gamma}{\rho}\|w_t - \nabla_x f(x_t, y_t)\|^2 + \frac{2L_y^2\gamma}{\rho}\|y_t - y^*(x_t)\|^2, \end{aligned} \tag{63}$$

where the second last inequality is due to $0 < \gamma \leq \frac{3\rho}{4L_F}$. By using Lemma 10, the difference between $\tilde{\mathcal{G}}_t$ and $\mathcal{G}_t$ are bounded, we have

$$\begin{aligned}\|\mathcal{G}_t\|^2 &\leq 2\|\tilde{\mathcal{G}}_t\|^2 + 2\|\tilde{\mathcal{G}}_t - \mathcal{G}_t\|^2 \\ &\leq 2\|\tilde{\mathcal{G}}_t\|^2 + \frac{2}{\rho^2}\|w_t - \nabla F(x_t)\|^2 \\ &\leq 2\|\tilde{\mathcal{G}}_t\|^2 + \frac{4}{\rho^2}\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + \frac{4L_y^2}{\rho^2}\|y_t - y^*(x_t)\|^2. \end{aligned} \tag{64}$$

Thus we have

$$-\|\tilde{\mathcal{G}}_t\|^2 \leq -\frac{1}{2}\|\mathcal{G}_t\|^2 + \frac{2}{\rho^2}\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + \frac{2L_y^2}{\rho^2}\|y_t - y^*(x_t)\|^2. \tag{65}$$

By plugging (65) into (63), we have

$$\begin{aligned}\Phi(x_{t+1}) &\leq \Phi(x_t) - \frac{3\gamma\rho}{32}\|\mathcal{G}_t\|^2 + \frac{3\gamma\rho}{16}\big(\frac{2}{\rho^2}\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + \frac{2L_y^2}{\rho^2}\|y_t - y^*(x_t)\|^2\big) \\ &\quad - \frac{3\rho}{16\gamma}\|x_{t+1} - x_t\|^2 + \frac{2\gamma}{\rho}\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + \frac{2L_y^2\gamma}{\rho}\|y_t - y^*(x_t)\|^2 \\ &= \Phi(x_t) - \frac{3\gamma\rho}{32}\|\mathcal{G}_t\|^2 - \frac{3\rho}{16\gamma}\|x_{t+1} - x_t\|^2 + \frac{19\gamma}{8\rho}\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + \frac{19L_y^2\gamma}{8\rho}\|y_t - y^*(x_t)\|^2. \end{aligned} \tag{66}$$

Next, we define a useful Lyapunov function, for any $t \geq 1$

$$\Omega_t = \mathbb{E}\big[\Phi(x_t) + \|y_t - y^*(x_t)\|^2\big]. \tag{67}$$

According to Lemma 11, we have

$$\begin{aligned}\|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2 \leq &-\frac{\eta_t\mu\lambda}{4}\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t\lambda^2}{4}\|v_t\|^2 \\ &+ \frac{25\eta_t\lambda}{6\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2 + \frac{25\kappa^2}{6\eta_t\mu\lambda}\|x_{t+1} - x_t\|^2. \end{aligned} \tag{68}$$

Let $\eta = \eta_t$ for all $t \geq 0$. Then we have

$$\begin{aligned}\Omega_{t+1} - \Omega_t &= \mathbb{E}\big[\Phi(x_{t+1}) - \Phi(x_t) + \|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2\big] \\ &\leq -\frac{3\gamma\rho}{32}\mathbb{E}\|\mathcal{G}_t\|^2 - \frac{3\rho}{16\gamma}\mathbb{E}\|x_{t+1} - x_t\|^2 + \frac{19\gamma}{8\rho}\mathbb{E}\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 + \frac{19L_y^2\gamma}{8\rho}\mathbb{E}\|y_t - y^*(x_t)\|^2 \\ &\quad - \frac{\eta_t\mu\lambda}{4}\mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t\lambda^2}{4}\mathbb{E}\|v_t\|^2 + \frac{25\eta_t\lambda}{6\mu}\mathbb{E}\|\nabla_y g(x_t, y_t) - v_t\|^2 + \frac{25\kappa^2}{6\eta_t\mu\lambda}\mathbb{E}\|x_{t+1} - x_t\|^2 \\ &= -\frac{3\gamma\rho}{32}\mathbb{E}\|\mathcal{G}_t\|^2 - \big(\frac{3\rho}{16\gamma} - \frac{25\kappa^2}{6\eta_t\mu\lambda}\big)\mathbb{E}\|x_{t+1} - x_t\|^2 + \frac{19\gamma}{8\rho}\mathbb{E}\|w_t - \bar{\nabla} f(x_t, y_t)\|^2 \\ &\quad - \big(\frac{\eta_t\mu\lambda}{4} - \frac{19L_y^2\gamma}{8\rho}\big)\mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t\lambda^2}{4}\mathbb{E}\|v_t\|^2 + \frac{25\eta_t\lambda}{6\mu}\mathbb{E}\|\nabla_y g(x_t, y_t) - v_t\|^2 \\ &\leq -\frac{3\gamma\rho}{32}\mathbb{E}\|\mathcal{G}_t\|^2 - \big(\frac{3\rho}{16\gamma} - \frac{25\kappa^2}{6\eta_t\mu\lambda}\big)\mathbb{E}\|x_{t+1} - x_t\|^2 + \frac{19\gamma}{4\rho}\mathbb{E}\|w_t - \bar{\nabla} f(x_t, y_t) - R(x_t, y_t)\|^2 \\ &\quad + \frac{19\gamma}{4\rho}\mathbb{E}\|R(x_t, y_t)\|^2 - \big(\frac{\eta_t\mu\lambda}{4} - \frac{19L_y^2\gamma}{8\rho}\big)\mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t\lambda^2}{4}\mathbb{E}\|v_t\|^2 \\ &\quad + \frac{25\eta_t\lambda}{6\mu}\mathbb{E}\|\nabla_y g(x_t, y_t) - v_t\|^2 \\ &\leq -\frac{3\gamma\rho}{32}\mathbb{E}\|\mathcal{G}_t\|^2 - \big(\frac{3\rho}{16\gamma} - \frac{25\kappa^2}{6\eta\mu\lambda}\big)\mathbb{E}\|x_{t+1} - x_t\|^2 + \frac{19\gamma}{4\rho}\mathbb{E}\|w_t - \bar{\nabla} f(x_t, y_t) - R(x_t, y_t)\|^2 \\ &\quad + \frac{19\gamma}{4\rho}\mathbb{E}\|R(x_t, y_t)\|^2 - \frac{3\eta\lambda^2}{4}\mathbb{E}\|v_t\|^2 + \frac{25\eta\lambda}{6\mu}\mathbb{E}\|\nabla_y g(x_t, y_t) - v_t\|^2, \end{aligned} \tag{69}$$

where the last inequality holds by $\gamma \leq \frac{2\rho\eta\mu\lambda}{19L_y^2}$.

Summing over $t = 0, 1, \cdots, T-1$ on both sides of (69), by Lemma 12, we have

$$
\begin{aligned}
&\frac{3\gamma\rho}{32}\sum_{t=0}^{T-1}\mathbb{E}\|\mathcal{G}_t\|^2 \\
&\leq \Omega_0 - \Omega_T - \big(\frac{3\rho}{16\gamma} - \frac{25\kappa^2}{6\eta\mu\lambda}\big)\sum_{t=0}^{T-1}\mathbb{E}\|x_{t+1}-x_t\|^2 - \frac{3\eta\lambda^2}{4}\mathbb{E}\|v_t\|^2 + \frac{19\gamma}{4\rho}\sum_{t=0}^{T-1}\mathbb{E}\|R(x_t,y_t)\|^2 \\
&\quad + \frac{19\gamma}{4\rho}\sum_{t=0}^{T-1}\big(\frac{L_K^2}{b_1}\sum_{i=(n_t-1)q}^{t-1}\big(\mathbb{E}\|x_{i+1}-x_i\|^2 + \mathbb{E}\|y_{i+1}-y_i\|^2\big) + \frac{\sigma^2}{b}\big) \\
&\quad + \frac{25\eta\lambda}{6\mu}\sum_{t=0}^{T-1}\big(\frac{L^2}{b_1}\sum_{i=(n_t-1)q}^{t-1}\big(\mathbb{E}\|x_{i+1}-x_i\|^2 + \mathbb{E}\|y_{i+1}-y_i\|^2\big) + \frac{\sigma^2}{b}\big) \\
&\leq \Omega_0 - \Omega_T - \big(\frac{3\rho}{16\gamma} - \frac{25\kappa^2}{6\eta\mu\lambda}\big)\sum_{t=0}^{T-1}\mathbb{E}\|x_{t+1}-x_t\|^2 - \frac{3\eta\lambda^2}{4}\mathbb{E}\|v_t\|^2 + \frac{19\gamma}{4\rho}\sum_{t=0}^{T-1}\mathbb{E}\|R(x_t,y_t)\|^2 \\
&\quad + \frac{19\gamma}{4\rho}\sum_{t=0}^{T-1}\big(\frac{L_K^2 q}{b_1}(\mathbb{E}\|x_{t+1}-x_t\|^2 + \mathbb{E}\|y_{t+1}-y_t\|^2) + \frac{\sigma^2}{b}\big) \\
&\quad + \frac{25\eta\lambda}{6\mu}\sum_{t=0}^{T-1}\big(\frac{L^2 q}{b_1}(\mathbb{E}\|x_{t+1}-x_t\|^2 + \mathbb{E}\|y_{t+1}-y_t\|^2) + \frac{\sigma^2}{b}\big) \\
&= \Omega_0 - \Omega_T - \big(\frac{3\rho}{16\gamma} - \frac{25\kappa^2}{6\eta\mu\lambda} - \frac{19\gamma L_K^2 q}{4\rho b_1} - \frac{25\eta\lambda L^2 q}{6\mu b_1}\big)\sum_{t=0}^{T-1}\mathbb{E}\|x_{t+1}-x_t\|^2 + \frac{19\gamma}{4\rho}\sum_{t=0}^{T-1}\mathbb{E}\|R(x_t,y_t)\|^2 \\
&\quad - \big(\frac{3\eta\lambda^2}{4} - \frac{19\gamma L_K^2 q\eta^2\lambda^2}{4\rho b_1} - \frac{25\lambda^3 L^2 q\eta^3}{6\mu b_1}\big)\sum_{t=0}^{T-1}\mathbb{E}\|v_t\|^2 + \big(\frac{19\gamma}{4\rho} + \frac{25\eta\lambda}{6\mu}\big)\frac{T\sigma^2}{b},
\end{aligned} \tag{70}
$$

where the second inequality holds by $\sum_{t=0}^{T-1}\sum_{i=(n_t-1)q}^{t-1}\big(\mathbb{E}\|x_{i+1}-x_i\|^2 + \mathbb{E}\|y_{i+1}-y_i\|^2\big) \leq q\sum_{t=0}^{T-1}\big(\mathbb{E}\|x_{t+1}-x_t\|^2 + \mathbb{E}\|y_{t+1}-y_t\|^2\big)$.

Let $b_1 = q$, $0 < \gamma \leq \frac{3\rho}{38L_K^2\eta}$ and $0 < \lambda \leq \frac{9\mu}{100\eta^2L^2}$, we have $\frac{3\eta\lambda^2}{8} \geq \frac{19\gamma L_K^2 q\eta^2\lambda^2}{4\rho b_1}$ and $\frac{3\eta\lambda^2}{8} \geq \frac{25\lambda^3L^2q\eta^3}{6\mu b_1}$, i.e., we obtain

$$
\frac{3\eta\lambda^2}{4} - \frac{19\gamma L_K^2 q\eta^2\lambda^2}{4\rho b_1} - \frac{25\lambda^3 L^2 q\eta^3}{6\mu b_1} \geq 0. \tag{71}
$$

At the same time, we have $\frac{3}{8\eta} \leq \frac{19\gamma L_K^2 q}{4\rho b_1}$, $\frac{3}{8\eta} \leq \frac{25\eta\lambda L^2 q}{6\mu b_1}$, $\frac{3}{8\eta L_K^2} \leq \frac{19\gamma}{4\rho}$ and $\frac{3}{8\eta L^2} \leq \frac{25\eta\lambda}{6\mu}$. Thus we have $\frac{3}{4\eta} \geq \frac{19\gamma L_K^2 q}{4\rho b_1} + \frac{25\eta\lambda L^2 q}{6\mu b_1}$. Let $\gamma \leq \min\big(\frac{\rho\eta}{8}, \frac{9\rho\eta\mu\lambda}{400\kappa^2}\big)$, we have

$$
\frac{3\rho}{16\gamma} \geq \frac{25\kappa^2}{6\eta\mu\lambda} + \frac{3}{4\eta} \geq \frac{25\kappa^2}{6\eta\mu\lambda} + \frac{19\gamma L_f^2 q}{8\rho b_1} + \frac{25\eta\lambda L_f^2 q}{6\mu b_1}. \tag{72}
$$

Based on the above inequalities (71) and (72), we have

$$
\frac{3\gamma\rho}{32}\sum_{t=0}^{T-1}\mathbb{E}\|\mathcal{G}_t\|^2 \leq \Omega_0 - \Omega_T + \frac{19\gamma}{4\rho}\sum_{t=0}^{T-1}\mathbb{E}\|R(x_t,y_t)\|^2 + \frac{3}{8\eta}\big(\frac{1}{L^2} + \frac{1}{L_K^2}\big)\frac{T\sigma^2}{b}, \tag{73}
$$

By using the above inequality (73), we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathcal{G}_t\|^2 &\leq \frac{32(\Omega_0-\Omega_T)}{3\gamma\rho T}+\frac{152}{3T\rho^2}\sum_{t=0}^{T-1}\mathbb{E}\|R(x_t,y_t)\|^2+\frac{4}{\eta\rho\gamma}\big(\frac{1}{L^2}+\frac{1}{L_K^2}\big)\frac{\sigma^2}{b}\\
&= \frac{32(\Phi(x_0)+\|y_0-y^*(x_0)\|^2)}{3T\gamma\rho}-\frac{32\mathbb{E}(\Phi(x_T)+\|y_T-y^*(x_T)\|^2)}{3T\gamma\rho}\\
&\quad+\frac{152}{3T\rho^2}\sum_{t=0}^{T-1}\mathbb{E}\|R(x_t,y_t)\|^2+\frac{4}{\eta\rho\gamma}\big(\frac{1}{L^2}+\frac{1}{L_K^2}\big)\frac{\sigma^2}{b}\\
&\leq \frac{32(\Phi(x_0)-\Phi^*)}{3T\gamma\rho}+\frac{32\Delta}{3T\gamma\rho}+\frac{152}{3T\rho^2}\sum_{t=0}^{T-1}\mathbb{E}\|R(x_t,y_t)\|^2+\frac{4}{\eta\rho\gamma}\big(\frac{1}{L^2}+\frac{1}{L_K^2}\big)\frac{\sigma^2}{b}\\
&\leq \frac{32(\Phi(x_0)-\Phi^*)}{3T\gamma\rho}+\frac{32\Delta}{3T\gamma\rho}+\frac{152}{3T^2\rho^2}+\frac{4}{\eta\rho\gamma}\big(\frac{1}{L^2}+\frac{1}{L_K^2}\big)\frac{\sigma^2}{b},
\end{aligned}\tag{74}$$

where the last inequality is due to $\mathbb{E}\|R(x_t,y_t)\|\leq\frac{1}{T}$ for all $t\geq 0$ by choosing $K=\frac{L}{\mu}\log(\frac{LC_{fy}T}{\mu})$.

□