Enhanced Bilevel Optimization via Bregman Distance

Feihu Huang\textsuperscript{1,2}, Junyi Li\textsuperscript{1}, Shangqian Gao\textsuperscript{1}, Heng Huang\textsuperscript{1}

\textsuperscript{1}Electrical & Computer Engineering, University of Pittsburgh, Pittsburgh, PA, United States
\textsuperscript{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 $\kappa$ and target accuracy $\epsilon$ for finding an $\epsilon$-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 \cite{37, 20, 9, 38}, meta learning \cite{9, 31, 22}, neural network architecture search \cite{30}, reinforcement learning \cite{15}, and image processing \cite{31}. In the paper, we consider solving the following nonconvex-strongly-convex bilevel optimization problem:

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

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 $\mu$-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 \cite{11, 22} that does not consider any nonsmooth regularization. Here the function $h(x)$ can be the nonsmooth regularization term such as $h(x) = \lambda \|x\|_1$.
Many recent machine learning research problems utilize the stochastic loss functions. Thus, we also consider the following stochastic bilevel optimization problem:

\[
\min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \mathbb{E}_{\xi \sim \mathcal{D}} \left[ f(x, y^*(x); \xi) \right] + h(x),
\]

(Outer)

\[
\text{s.t. } y^*(x) \in \arg \min_{y \in \mathbb{R}^d} \mathbb{E}_{\zeta \sim \mathcal{D}'} \left[ g(x, y; \zeta) \right],
\]

(Inner)

where function \( F(x) = \mathbb{E}_{\xi} [F(x; \xi)] = \mathbb{E}_{\xi} [f(x, y^*(x); \xi)] \) is smooth and possibly nonconvex, and function \( h(x) \) is convex and possibly nonsmooth, and function \( g(x, y) = \mathbb{E}_{\zeta} [g(x, y; \zeta)] : \mathcal{X} \times \mathbb{R}^d \rightarrow \mathbb{R} \) is \( \mu \)-strongly convex in \( y \in \mathbb{R}^d \). \( \xi \) and \( \zeta \) 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:

1. 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.
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).

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, [44, 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 $\ell_p$ 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 $\ell_2$-norm for vectors and spectral norm for matrices,
For stochastic case, the same assumptions hold for Assumptions 1-3 are commonly used in bilevel optimization methods [11, 22, 23]. According to the partial derivatives \( \nabla f(x; B) = \frac{1}{b} \sum_{i=1}^{b} \nabla f(x; \xi_i) \). For two sequences \( \{a_n, b_n\}_{n=1}^{\infty} \), \( a_n = O(b_n) \) denotes that \( a_n \leq Cb_n \) for some constant \( C > 0 \). The notation \( \mathcal{O}(\cdot) \) 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 \( \mu \)-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) \( \|\nabla_y f(x, y)\| \leq C_{f_y} \) and \( \|\nabla^2_{xy} g(x, y)\| \leq C_{g_{xy}} \) for any \( x \in \mathcal{X} \) and \( y \in \mathbb{R}^d \);

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 all \( x, x_1, x_2 \in \mathcal{X} \) and \( y, y_1, y_2 \in \mathbb{R}^d \),
\[
\|\nabla_x f(x_1, y) - \nabla_x f(x_2, y)\| \leq L \|x_1 - x_2\|, \quad \|\nabla_y f(x_1, y) - \nabla_y f(x_2, y)\| \leq L \|y_1 - y_2\|.
\]

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

Assumption 3. The partial derivatives \( \nabla^2_{xg}(x, y) \) and \( \nabla^2_{yg}(x, y) \) are \( L_{g_{xy}} \)-Lipschitz and \( L_{g_{yy}} \)-Lipschitz, e.g., for all \( x, x_1, x_2 \in \mathcal{X} \) and \( y, y_1, y_2 \in \mathbb{R}^d \)
\[
\|\nabla^2_{xy} g(x_1, y) - \nabla^2_{xy} g(x_2, y)\| \leq L_{g_{xy}} \|x_1 - x_2\|, \quad \|\nabla^2_{yg} g(x, y_1) - \nabla^2_{yg} g(x, y_2)\| \leq L_{g_{yy}} \|y_1 - y_2\|.
\]

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

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, \( \|f(x, y_1) - f(x, y_2)\| = \|\nabla_y f(x, y_1)(y_1 - y_2)\| \leq C_{f_y}\|y_1 - y_2\| \leq C_{f_y}\|y_1 - y_2\| \) where \( y = r y_1 + (1-r)y_2 \) and \( r \in [0, 1] \). Thus \( \|\nabla_y f(x, y)\| \leq C_{f_y} \) 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 \( \nabla^2_{xy} f(x, y) \leq M \). 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 \( \|\nabla^2_{xy} g(x_1, y) - \nabla^2_{xy} g(x_2, y)\| \leq L \|x_1 - x_2\| \), \( \|\nabla^2_{yy} g(x, y_1) - \nabla^2_{yy} g(x, y_2)\| \leq L \|y_1 - y_2\| \), where \( x, y = r_1 x_1 + (1-r_1)x_2 \) and \( r_1 \in [0, 1] \), we can let \( C_{g_{xy}} = L \) 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 \( x \) and \( y \). Throughout the paper, we let \( C_{g_{xy}} = L \). Assumption 4 is generally used for regularizing 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 \( \nabla_x g(x, y) \) or \( \nabla_y g(x, y; \zeta) \) to update variable \( y \). However, it is hard to get the (stochastic) gradient \( \nabla F(x) = \frac{\partial f(x, y^*(x))}{\partial x} + \frac{\partial f(x, y^*(x))}{\partial y} \) 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, 2]) Under the above Assumptions [1, 2, 3], we have, for any \( x \in \mathcal{X} \)
\[
\nabla F(x) = \nabla_x f(x, y^*(x)) + \nabla_y f(x, y^*(x)) = \nabla_x f(x, y^*(x)) - \nabla^2_{xy} g(x, y^*(x)) [\nabla^2_{yy} g(x, y^*(x))]^{-1} \nabla_y f(x, y^*(x)).
\]

Lemma 1 provides a natural estimator of \( \nabla F(x) \), defined as, for all \( x \in \mathcal{X}, y \in \mathbb{R}^d \)
\[
\tilde{\nabla} f(x, y) = \nabla_x f(x, y) - \nabla^2_{xy} g(x, y) [\nabla^2_{yy} g(x, y)]^{-1} \nabla_y f(x, y).
\]

Next, we show some properties of \( \nabla F(x), y^*(x) \) and \( \tilde{\nabla} f(x, y) \) in the following lemma:
Algorithm 1 Deterministic BiO-BreD Algorithm

1: Input: \( T, K \geq 1 \), learning rates \( \gamma > 0, \lambda > 0 \);
2: initialize: \( x_0 \in \mathcal{X} \) and \( y_0^{K_1} = y_0 \in \mathbb{R}^{d_2} \);
3: for \( t = 0, 1, \ldots, T - 1 \) do
4: \( y_0 = y_0^{K_1} \);
5: for \( k = 1, \ldots, K \) do
6: \( \text{Update } y_t^k = y_t^{k-1} - \lambda \nabla_y g(x_t, y_t^{k-1}) \);
7: end for
8: \( \text{Compute partial derivative } w_1 = \frac{\partial f(x_t, y_t^k)}{\partial x_t} \text{ via backpropagation w.r.t. } x_t \);
9: \( \text{Given a } \rho \text{-strongly convex mirror function } \psi_t \);
10: \( \text{Update } x_{t+1} = \arg \min_{x \in \mathcal{X}} \left\{ \langle w_t, x \rangle + h(x) + \frac{1}{2} \kappa D_{\psi_t}(x, x_t) \right\} \);
11: end for
12: Output: Uniformly and randomly choose from \( \{ x_t, y_t \}_{t=1}^T \).

Lemma 2. (Lemma 2.2 in [11]) Under the Assumptions [12][3], for all \( x, x_1, x_2 \in \mathcal{X} \) and \( y \in \mathbb{R}^{d_2} \), we have \( \| \nabla f(x, y) - \nabla F(x) \| \leq L_y \| y^*(x) - y \| \),

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

where \( L_y = L + \frac{L_y^2}{\mu} + \frac{C_{1y} C_{2y}}{\mu^2} + \frac{L_{yys} C_{1y} L}{\mu^2} \), \( \kappa = \frac{L}{\mu} \), and \( L_F = L + \frac{2L_y^2 + L_{yys} C_{1y}^2 L}{\mu} + \frac{L_{yys} C_{1y} L^2}{\mu^2} \).

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 \( \rho \)-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:

\[
x_{t+1} = \arg \min_{x \in \mathcal{X}} \left\{ \langle w_t, x \rangle + h(x) + \frac{1}{2} \kappa D_{\psi_t}(x, x_t) \right\}, \tag{5}
\]

where \( \gamma > 0 \) is stepsize, and \( w_t \) is an estimator of \( \nabla F(x_t) \). Here the mirror function \( \psi_t \) 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 \( D_{\psi_t}(x, x_t) = \frac{1}{2} \| x - x_t \|^2 \). When \( H_t \) is an approximated Hessian matrix, the above subproblem (5) is equivalent to the proximal quasi-Newton descent. When \( H_t \) is an adaptive matrix as used in [19], the above subproblem (5) is equivalent to the proximal adaptive gradient descent.

In Algorithm 1 we use gradient estimator \( w_t = \frac{\partial f(x_t, y_t^k)}{\partial x_t} \) to estimate \( \nabla F(x_t) \), where the partial derivative \( w_t = \frac{\partial f(x_t, y_t^k)}{\partial x_t} \) 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

1: Input: $T, K \geq 1$, stepsizes $\gamma > 0, \lambda > 0, \{\eta_t\}_{t=1}^T$;
2: initialize: $x_0, y_0 \in \mathbb{R}^{d_2}$;
3: for $t = 0, 1, \ldots, T - 1$ do
4: Draw randomly $b$ independent samples $B_t = \{\xi_t^i\}_{i=1}^b$, and compute stochastic partial derivatives $v_t = \nabla g(x_t, y_t; B_t)$;
5: Update $y_{t+1} = y_t - \lambda \eta_t v_t$;
6: Draw randomly $(K + 1)$ independent samples $\bar{B}_t = \{x_{t,i}, y_{0,i}, \ldots, y_{K-1,i}\}_{i=1}^b$, and compute stochastic partial derivatives $w_t = \nabla f(x_t, y_t; \bar{B}_t)$;
7: Given a $\rho$-strongly convex mirror function $\psi_t$;
8: Update $x_{t+1} = \arg\min_{x \in \mathcal{X}} \{\langle w_t, x \rangle + h(x) + \frac{\rho}{2} D_{\psi_t}(x, x_t)\}$;
9: end for
10: Output: Uniformly and randomly choose from $\{x_t, y_t\}_{t=1}^T$.

Given $K \geq 1$ and draw $K + 1$ independent samples $\bar{\xi} = \{\xi, \xi_0, \ldots, \xi_K\}$, as in [15, 23], we define a stochastic gradient estimator:

$$
\nabla f(x, y, \xi) = \nabla f(x, y; \xi) - \nabla^2_{x,y} g(x, y; \zeta_0) \left[ \frac{K}{L} \prod_{i=1}^{k} \left( Id_2 - \frac{1}{L} \nabla^2_{x,y} g(x, y; \zeta_i^0) \right) \right] \nabla_y f(x, y; \xi),
$$

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

**Lemma 3.** (Lemma 2.1 in [23]) Under the about Assumptions 7[2][3], for any $K \geq 1$, the gradient estimator in (6) satisfies

$$
\|R(x, y)\| \leq \frac{LCf(y)}{\mu} \left(1 - \frac{\mu}{L}\right)^K.
$$

Lemma 5 shows that the bias $R(x, y)$ decays exponentially fast with number $K$, and with choosing $K = \frac{\mu}{L} \log(LCf(y)/\mu)$, we have $\|R(x, y)\| \leq \frac{1}{\mu}$. Let $\frac{LCf(y)}{\mu} \left(1 - \frac{\mu}{L}\right)^K \leq \frac{1}{\mu}$, we have $K \geq \log(Cf_0) / \log(L/\mu)$. Further due to $\frac{L}{\mu} \leq \log(L/\mu)$, we have $\|R(x, y)\| \leq \frac{1}{\mu}$. Note that here we use $C_{gxy} = L$.

To simplify notations, let $\bar{\xi}_i = \{\xi_{i,0}, \xi_{i,1}, \ldots, \xi_{i,K-1}\}$. In Algorithm 2 we use mini-batch stochastic gradient estimator $w_t = \nabla f(x_{t,i}, y_{t}; \bar{B}_i) = \frac{1}{b} \sum_{i=1}^{b} \nabla f(x_{t,i}, y_{t}; \xi_i)$, where $\nabla f(x_{t,i}, y_{t}; \xi_i)$

$$
= \nabla f(x_t, y_t; \xi_{t,i}) - \nabla^2_{x,y} g(x_t, y_t; \xi_0^i) \left[ \frac{K}{L} \prod_{i=1}^{k} \left( Id_2 - \frac{1}{L} \nabla^2_{x,y} g(x_t, y_t; \zeta_i^0) \right) \right] \nabla_y f(x_t, y_t; \xi_{t,i}),
$$

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

### 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 $B_t = \{\xi_t^i\}_{i=1}^b$ and $\bar{B}_t = \{\xi_t^i\}_{i=1}^b$ to estimate our stochastic partial derivatives $v_t$ and $w_t$, respectively. When $\mod(t, q) \neq 0$, we draw a mini-batch samples $\mathcal{I}_t = \{\xi_t^i\}_{i=1}^b$ and $\bar{I}_t = \{\xi_t^i\}_{i=1}^b$ to estimate $v_t$ and $w_t$, respectively. Let $R(x_t, y_t) = \nabla f(x_t, y_t; \mathcal{I}_t) - \nabla f(x_t, y_t)$ when $\mod(t, q) \neq 0$, we have $\mathbb{E}\nabla f(x_t, y_t; \mathcal{I}_t) = R(x_t, y_t) + \nabla f(x_t, y_t)$ when $\mod(t, q) \neq 0$. We have $\mathbb{E}\nabla f(x_t, y_t; \mathcal{I}_t) = \frac{LCf(y)}{\mu} \left(1 - \frac{\mu}{L}\right)^K$. 


We begin with introducing a useful convergence metric \( \|G\| \). Assumption 6. \( F \) defined in (6) and above assumptions, given Lemma 4. (Lemma 3.1 in [23]) Under the above Assumptions (1, 2, 3), stochastic gradient estimator \( \|G\| \) in [11, 22]. When \( h \) algorithms, as in [10, 29], we define the generalized gradient at the properties of our algorithms. Given the generated parameter vector \( \BreD \), and \( \ASBiO-BreD \) under mild conditions. All proofs are provided in the Appendix B. Output: 14: Update 11: end if 7: Draw randomly \( t \) \( \{\xi_i\}_{i=1}^{b_1} \), and compute stochastic partial derivative \( v_t = \nabla g(x_t, y_t; B_t) \); 6: Draw randomly \( b(K+1) \) independent samples \( \BreD = \{\xi_i, \zeta_i^{(0)}, \ldots, \zeta_i^{(K-1)}\}_{i=1}^{b_1} \), and compute stochastic partial derivative \( w_t = \nabla f(x_t, y_t; \BreD) \); 8: Generate randomly \( b_1 \) independent samples \( \BreD = \{\zeta_i^{(0)}, \ldots, \zeta_i^{(K-1)}\}_{i=1}^{b_1} \), and compute stochastic partial derivative \( w_t = \nabla f(x_t, y_t; \BreD) = \nabla f(x_t, y_t; \BreD) + w_{t-1} \); end if 11: Update \( y_{t+1} = y_t - \lambda \eta_t v_t; \) 12: Given a \( \rho \)-strongly convex mirror function \( \psi_i; \) 13: Output \( x_{t+1} = \arg \min_{x} \{ (w_1, x) + h(x) + \frac{1}{\gamma} D_{\psi_i}(x, x_1) \} ; \) 15: 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 [5].

We begin with introducing a useful convergence metric \( \|G_t\|^2 \) or \( \mathbb{E}\|G_t\|^2 \) 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:

\[
G_t = \frac{1}{\gamma} (x_t - x_{t+1}^+), \quad x_{t+1}^+ = \arg \min_{x \in X} \left\{ \nabla F(x_t, x) + h(x) + \frac{1}{\gamma} D_{\psi_i}(x, x_1) \right\},
\]

where \( F(x) = f(x, y^*(x)) \) or \( F(x) = \mathbb{E}_\xi[f(x, y^*(x); \xi)] \). If \( \psi_i(x) = \frac{1}{2}\|x\|^2 \), \( \mathcal{X} = \mathbb{R}^{d_1} \) and \( h(x) = c \) is a constant, we have \( \|G_t\|^2 = \|\nabla F(x_t)\|^2 \), which is a common convergence metric used in [11, 22]. When \( \psi_i(x) = \frac{1}{2}\|x\|^2 \), \( \mathcal{X} \subseteq \mathbb{R}^{d_1} \) and \( h(x) = c \) is a constant, our convergence metric is \( \|G_t\|^2 = \frac{1}{2}(x_t - \mathcal{P}_\mathcal{X}(x_t - 2\nabla F(x_t)))^2 \), 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 \( \nabla f(x, y; \xi) \) is \( L_K \)-Lipschitz continuous, e.g., for \( x_1, x_2 \in \mathcal{X} \) and \( y \in \mathbb{R}^{d_2} \),

\[
\mathbb{E}_\xi \|\nabla f(x_1, y; \xi) - \nabla f(x_2, y; \xi)\|^2 \leq L_{K}^2 \|x_1 - x_2\|^2,
\]

where \( L_{K}^2 = 2L^3 + 6L^4 \frac{K}{\eta_t \mu - \mu} + 6C_{f y} L_2 \gamma y \frac{K}{2\mu - \mu} + 6L^4 \frac{K^2 \lambda t}{(L - \mu \gamma)(2\mu L - \mu)} \).

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 \( 0 < \lambda \leq \frac{1}{4\gamma} \), we have

\[
\|y_{t+1} - y^*(x_{t+1})\|^2 \leq (1 - \frac{\eta_t \mu \lambda}{4}) \|y_t - y^*(x_t)\|^2 - \frac{3\eta_t \lambda}{4} \|v_t\|^2 + \frac{25\eta_t \lambda}{6\mu} \|\nabla g(x_t, y_t) - v_t\|^2 + \frac{25\eta_t \lambda}{6\eta_t \mu} \|x_{t+1} - x_t\|^2.
\]

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

Assumption 6. The stochastic partial derivative \( \nabla_y g(x, y; \xi) \) satisfies \( \mathbb{E}\|\nabla_y g(x, y; \xi)\|^2 \leq \sigma^2 \). The estimated stochastic partial derivative \( \hat{\nabla} f(x, y; \xi) \) defined in [6] satisfies \( \mathbb{E}_\xi[\hat{\nabla} f(x, y; \xi)] = \nabla f(x, y) + R(x, y) \) and \( \mathbb{E}_\xi[\hat{\nabla} f(x, y; \xi) - \nabla f(x, y) - R(x, y)]^2 \leq \sigma^2 \).
Assumption 7. The mirror functions \( \{\psi_t(x)\}^T_{t=0} \) are \( \rho \)-strongly convex, where \( \rho > 0 \).

Assumption 6 is commonly used in stochastic bilevel optimization methods. Assumption 7 shows that the constant \( \rho \) 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 and adaptive gradient algorithms.

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 \( \{x_t, y_t\}^T_{t=1} \) be generated from Algorithm 1. Let \( 0 < \gamma \leq \frac{3\rho}{4L_F} \), \( 0 < \lambda < \frac{1}{L} \), \( K = \log(T) / \log(\frac{1}{1-\lambda\rho}) + 1 \) and \( \|y_0 - y^*(x_0)\|^2 \leq \Delta \) for all \( t \geq 0 \), we have

\[
1 \sum_{t=0}^{T-1} \|G_t\|^2 \leq \frac{16 (\Phi(x_0) - \Phi^*)}{3T \gamma \rho} + \frac{22 \Delta L_1^2}{\rho^2 T} + \frac{22 \Delta L_2^2}{\rho^2 T} + \frac{22 \Delta L_3^2}{\rho^2 T^2},
\]

where \( \kappa = \frac{L}{\mu}, \ L_1 = \frac{L(L+\mu)}{\mu}, \ L_2 = \frac{2C^{\text{ls}}(\mu L_{\text{avg}} + L L_{\text{avg}})}{\mu} \) and \( L_3 = \frac{LC^{\text{ls}}}{\mu} \).

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

5.2 Convergence Analysis of SBOi-BreD Algorithm

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

**Theorem 2.** Suppose the sequence \( \{x_t, y_t\}^T_{t=1} \) be generated from Algorithm 1. Let \( \Delta = \|y_0 - y^*(x_0)\|^2 \), \( K = \frac{L}{\mu} \log\left(\frac{L \mu T_c}{\mu T_c} + \frac{L \mu T_c}{\mu T_c}ight) \), \( 0 < \eta = \eta_t \leq 1 \), \( 0 < \gamma \leq \min\left(\frac{3\rho}{4L_F}, \frac{9 \mu \lambda}{800 \epsilon^2}, \frac{9 \mu \lambda}{4L_F^2}\right) \) and \( 0 < \lambda \leq \frac{1}{L} \), we have

\[
1 \sum_{t=1}^{T} \mathbb{E}[G_t^2] \leq \frac{32 (\Phi(x_0) - \Phi^*)}{3T \gamma \rho} + \frac{32 \Delta}{3T \gamma \rho} + \frac{752 \sigma^2}{3 \rho^2 b} + \frac{400 \eta \lambda \sigma^2}{9 \gamma \rho b} + \frac{752}{3 \rho^2 T^2}.
\]

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

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 \( \{x_t, y_t\}^T_{t=1} \) be generated from Algorithm 1. Let \( \Delta = \|y_0 - y^*(x_0)\|^2 \), \( b_1 = q \), \( K = \frac{L}{\mu} \log\left(\frac{L \mu T_c}{\mu T_c} + \frac{L \mu T_c}{\mu T_c}\right) \), \( 0 < \eta = \eta_t \leq 1 \), \( 0 < \gamma \leq \min\left(\frac{3\rho}{8L_F^2}, \frac{3 \mu \lambda}{42 \epsilon^2}, \frac{3 \mu \lambda}{64 L_F^2}, \frac{9 \mu \lambda}{100 \epsilon^2}, \frac{9 \mu \lambda}{100 \epsilon^2}\right) \) and \( 0 < \lambda \leq \min\left(\frac{1}{6L}, \frac{9 \mu \lambda}{1000 L_F^2}\right) \), we have

\[
1 \sum_{t=0}^{T-1} \mathbb{E}[G_t^2] \leq \frac{32 (\Phi(x_0) - \Phi^*)}{3T \gamma \rho} + \frac{152}{3 \gamma \rho^2} + \frac{4}{\eta} \gamma \rho \left(\frac{1}{L^2} + \frac{1}{L_F^2}\right) \frac{\sigma^2}{b}.
\]

**Remark 3.** Without loss of generality, let \( L \geq \frac{1}{\mu}, \lambda = \min\left(\frac{3\rho}{8L_F^2}, \frac{3 \mu \lambda}{42 \epsilon^2}, \frac{3 \mu \lambda}{64 L_F^2}, \frac{9 \mu \lambda}{100 \epsilon^2}, \frac{9 \mu \lambda}{100 \epsilon^2}\right) \) and \( \rho = O(L) \), we have \( \gamma \rho = O\left(\frac{1}{L^2}\right) \). It is easily verified that our ASBiO-BreD algorithm has a convergence rate of \( O\left(\frac{\kappa^4}{T^2} + \frac{\kappa^2}{b}\right) \). Let \( \frac{1}{T} = \epsilon \) and \( \frac{\kappa^4}{T} = \frac{1}{2} \) and \( \frac{\kappa^2}{b} = \frac{1}{2} \), we have
\[ T = 2\kappa^4 \epsilon^{-1} \text{ and } b = 2\kappa^2 \epsilon^{-1} \]. Due to \( K = \frac{\log \left( \frac{LC_T}{T} \right)}{\mu} \), we have \( K = O(\kappa \log(\frac{\kappa^4}{\epsilon})) = \tilde{O}(\kappa) \).

Let \( b_1 = q = \kappa^{-0.5} \). For finding \( \epsilon \)-stationary point of the problem (2), we need the gradient complexity: \( Gc(f, \epsilon) = 2\left( \frac{bT}{q} + 2b_1T \right) = O(\kappa^5 \epsilon^{-1.5}) \) and \( Gc(g, \epsilon) = \frac{bT}{q} + 2b_1T = O(\kappa^5 \epsilon^{-1.5}) \), and the Jacobian-vector and Hessian-vector product complexities: \( JV(g, \epsilon) = bT + 2b_1 = O(\kappa^5 \epsilon^{-1.5}) \) and \( HV(g, \epsilon) = K(bT + 2b_1) = \tilde{O}(\kappa^6 \epsilon^{-1.5}) \).

![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 \( \varrho \): large noise setting \( \varrho = 0.8 \) (top row) and small noise setting \( \varrho = 0.4 \) (bottom row).](image)

### 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:

\[
\begin{align*}
\min_{\lambda} & \quad l_{\text{eval}}(\lambda, w^*(\lambda)) := \frac{1}{|D_V|} \sum_{(x_i, y_i) \in D_V} l(x_i^T w^*(\lambda), y_i) \\
\text{s.t.} & \quad w^*(\lambda) = \arg \min_w l_T(\lambda, w) := \frac{1}{|D_T|} \sum_{(x_i, y_i) \in D_T} \sigma(\lambda_i) l(x_i^T w, y_i) + C \|w\|_2,
\end{align*}
\]

where \( l(\cdot) \) denotes the cross entropy loss, \( D_T \) and \( D_V \) are training and validation datasets, respectively. Here \( \lambda = \{\lambda_i\}_{i \in D_T} \) 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)

<table>
<thead>
<tr>
<th>Time</th>
<th>AID_BiO</th>
<th>ITD_BiO</th>
<th>MRBO</th>
<th>FSLA</th>
<th>VRBO</th>
<th>VR-saBiAdam</th>
<th>ASBiO-BreD</th>
</tr>
</thead>
<tbody>
<tr>
<td>20s</td>
<td>0.6509</td>
<td>0.6411</td>
<td>0.6103</td>
<td>0.6539</td>
<td>0.5951</td>
<td>0.6812</td>
<td>0.6653</td>
</tr>
<tr>
<td>40s</td>
<td>0.7365</td>
<td>0.7210</td>
<td>0.6971</td>
<td>0.7399</td>
<td>0.6805</td>
<td>0.7141</td>
<td>0.7403</td>
</tr>
<tr>
<td>60s</td>
<td>0.7762</td>
<td>0.7721</td>
<td>0.7519</td>
<td>0.7661</td>
<td>0.7429</td>
<td>0.7523</td>
<td>0.7830</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Time</th>
<th>AID_BiO</th>
<th>ITD_BiO</th>
<th>MRBO</th>
<th>FSLA</th>
<th>VRBO</th>
<th>VR-saBiAdam</th>
<th>ASBiO-BreD</th>
</tr>
</thead>
<tbody>
<tr>
<td>20s</td>
<td>0.8316</td>
<td>0.8131</td>
<td>0.8174</td>
<td>0.7993</td>
<td>0.7730</td>
<td>0.7753</td>
<td>0.8529</td>
</tr>
<tr>
<td>40s</td>
<td>0.8779</td>
<td>0.8621</td>
<td>0.8634</td>
<td>0.8485</td>
<td>0.8305</td>
<td>0.8188</td>
<td>0.8967</td>
</tr>
<tr>
<td>60s</td>
<td>0.9032</td>
<td>0.8968</td>
<td>0.8819</td>
<td>0.8824</td>
<td>0.8745</td>
<td>0.8640</td>
<td>0.9313</td>
</tr>
</tbody>
</table>

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 $\varrho$.

The detailed experimental setup is described in the Appendix A. 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:

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

subject to

$$w^*_\xi(\lambda) = \arg \min_w l_{tr}(\lambda, w; \xi) := \frac{1}{|D_{T,\xi}|} \sum_{(x_i, y_i) \in D_{T,\xi}} l\left(w^T \phi(x_i; \lambda), y_i\right) + C \|w\|^2,$$

where $l(\cdot)$ denotes the cross entropy loss, $D_{T,\xi}$ and $D_{V,\xi}$ are training and validation datasets for randomly sampled meta task $\xi$. 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. $\lambda$ 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. 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.

### 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 non-smooth 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 $\kappa$ and the target accuracy $\epsilon$ for finding an $\epsilon$-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


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 [13]. 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:

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

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

where $$l(\cdot)$$ denotes the cross entropy loss, $$D_T$$ and $$D_v$$ are training and validation dataset, respectively. Here $$\lambda = \{\lambda_i\} \in D_T$$ 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 $$K = 5$$ 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 0.1 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 SBiO-BreD; we set the stepsize as 3 for VRBO and our ASBiO-BreD.

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:

$$\min_{\lambda} l_{val}(\lambda, w^*(\lambda)) : = E_{\xi} \left[ \frac{1}{|D_{v,\xi}|} \sum_{(x_i, y_i) \in D_{v,\xi}} l\left( w^T(\lambda)^T \phi(x_i; \lambda), y_i \right) \right]$$

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

where $$l(\cdot)$$ denotes the cross entropy loss, $$D_{t,\xi}$$ and $$D_{v,\xi}$$ are training and validation dataset for randomly sampled meta task $$\xi$$. Here $$\phi(\cdot, \cdot)$$ is a four-layers convolutional neural network with maxpooling and 32 filters per layer [23], which denotes a representation mapping. $$\lambda$$ 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 $\frac{\partial f(x_t, y^K_t)}{\partial x}$ is the following analytical form:

$$\frac{\partial f(x_t, y^K_t)}{\partial x} = \nabla_x f(x_t, y^K_t) - \lambda \sum_{k=0}^{K-1} \nabla_{x,y} g(x_t, y^K_t) \prod_{j=k+1}^{K} (I_{d_2} - \lambda \nabla_{y,y} g(x_t, y^K_t)) \nabla_y f(x_t, y^K_t).$$

The above lemma shows an analytical form of $w_t$ in Algorithm [1].
Algorithm 1, and

Under the above assumptions, given \( g \in \mathbb{R}^n \) and \( 0 < \lambda < \frac{1}{2} \), we have

\[
\left\| \frac{\partial f(x_t, y_t^K)}{\partial x} - \nabla F(x_t) \right\| \leq \left( L_1 (1 - \lambda \mu) \frac{K}{2} + L_2 (1 - \lambda \mu) \frac{K - 1}{2} \right) \| y^*_t - y^*(x_t) \| + L_3 (1 - \lambda \mu)^K,
\]

where \( L_1 = \frac{L(1 - \mu)}{\mu} \), \( L_2 = \frac{2 \mu (L \chi_{\gamma} + L \chi_{\gamma})}{\mu^2} \) and \( L_3 = \frac{4 \mu}{\mu^2} \).

The above lemma 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 \( \mathbb{E}\| \frac{1}{n} \sum_{i=1}^n \zeta_i \|^2 = \frac{1}{4} \mathbb{E}\| \zeta_i \|^2 \) for any \( i \in [n] \).

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

\[
\langle w_t, \tilde{G}_t \rangle \geq \rho \| \tilde{G}_t \|^2 + \frac{1}{\gamma} (h(x_{t+1}) - h(x_t)),
\]

where \( \rho > 0 \) depends on \( \rho \)-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 \( x^+_{t+1} = \arg\min_{x \in X} \left\{ \langle \nabla F(x_t), x \rangle + h(x) + \frac{1}{\gamma} D_{\psi_t}(x, x_t) \right\} \) and let \( \tilde{G}_t = \frac{1}{\gamma} (x_t - x_{t+1}) \), we have

\[
\| \tilde{G}_t - \tilde{G}_t \| \leq \frac{1}{\rho} \| \nabla F(x_t) - w_t \|,
\]

where \( \rho > 0 \) depends on \( \rho \)-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 1, 2 and 3 Under the above assumptions, given \( 0 < \eta_t \leq 1 \) for all \( t \geq 1 \) and \( 0 < \lambda \leq \frac{1}{\eta_t} \), we have

\[
\| y_{t+1} - y^*(x_{t+1}) \|^2 \leq \left( 1 - \frac{\eta_t \lambda}{4} \right) \| y_t - y^*(x_t) \|^2 - \frac{3 \eta_t \lambda^2}{4} \| v_t \|^2 + \frac{25 \eta_t \lambda^2}{6 \mu^2} \| \nabla g(x_t, y_t) - v_t \|^2 + \frac{25 \eta_t \lambda^2}{6 \mu^2} \| x_{t+1} - x_t \|^2,
\]

where \( \kappa = \frac{L}{\mu} \).

**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 1 and step 11 in Algorithm 3 i.e., \( y_{t+1} = y_t - \lambda v_t \). This proof mainly follows the proof of Lemma 28 in \([17]\).

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

\[
g(x_t, y) \geq g(x_t, y_t) + \langle \nabla_y g(x_t, y_t), y - y_t \rangle + \frac{\mu}{2} \| y - y_t \|^2
\]

\[
- g(x_t, y_t) + \langle \nabla_y g(x_t, y_t), v_t, y - \tilde{y}_{t+1} \rangle + \langle \nabla_y g(x_t, y_t) - v_t, y - \tilde{y}_{t+1} \rangle + \frac{\mu}{2} \| y - y_t \|^2.
\]

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

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

By combining the above inequalities \([14]\) with \([15]\), we have

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

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

\[
\langle v_t, y - \tilde{y}_{t+1} \rangle = \frac{1}{\lambda} \langle \tilde{y}_{t+1} - y_t, \tilde{y}_{t+1} - y_t \rangle
\]

\[
= \frac{1}{\lambda} \| \tilde{y}_{t+1} - y_t \|^2 + \frac{1}{\lambda} \langle \tilde{y}_{t+1} - y_t, y_t - y_t \rangle.
\]
By plugging the inequalities (17) into (16), we have
\[
g(x_t, y) \geq g(x_t, y_{t+1}) + \frac{1}{\lambda} \langle \dot{y}_{t+1} - y_t - y \rangle + \frac{\mu}{4} \| y - y_{t+1} \|^2 \\
+ \langle \nabla_y g(x_t, y) - v_t, y - y_{t+1} \rangle + \frac{\mu}{2} \| y - y_t \|^2 = L \| y - y_{t+1} - y_t \|^2.
\] (18)

Let \( y = y^*(x_t) \), then we have
\[
g(x_t, y^*(x_t)) \geq g(x_t, y_{t+1}) + \frac{1}{\lambda} \langle \dot{y}_{t+1} - y_t - y^*(x_t) \rangle + \frac{\mu}{2} \| y - y_{t+1} \|^2 \\
+ \langle \nabla_y g(x_t, y) - v_t, y^*(x_t) - y_{t+1} \rangle + \frac{\mu}{2} \| y^*(x_t) - y_t \|^2.
\] (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 \( g(x_t, y^*(x_t)) \leq g(x_t, y_{t+1}) \). Thus, we obtain
\[
0 \geq \frac{1}{\lambda} \langle \dot{y}_{t+1} - y_t - y - y^*(x_t) \rangle + \langle \nabla_y g(x_t, y) - v_t, y^*(x_t) - y_{t+1} \rangle \\
+ \langle \frac{1}{\lambda} (\dot{y}_{t+1} - y_t - y^*(x_t)) + \mu g^*(x_t) - y_{t+1} \rangle + \frac{\mu}{2} \| y^*(x_t) - y_t \|^2.
\] (20)

By \( y_{t+1} = y_t + \eta_t (\tilde{y}_{t+1} - y_t) \), we have
\[
\| y_{t+1} - y^*(x_t) \|^2 = \| y_t + \eta_t (\tilde{y}_{t+1} - y_t) - y^*(x_t) \|^2 \\
= \| y_t - y^*(x_t) \|^2 + 2 \eta_t \langle \tilde{y}_{t+1} - y_t, y_t - y^*(x_t) \rangle + \eta_t^2 \| \tilde{y}_{t+1} - y_t \|^2.
\] (21)

Then we obtain
\[
\langle \tilde{y}_{t+1} - y_t, y_t - y^*(x_t) \rangle = \frac{1}{2 \eta_t} \| y_{t+1} - y^*(x_t) \|^2 - \frac{1}{2 \eta_t} \| y_t - y^*(x_t) \|^2 - \frac{\eta_t}{2} \| \tilde{y}_{t+1} - y_t \|^2.
\] (22)

Consider the upper bound of the term \( \langle \nabla_y g(x_t, y_t) - v_t, y^*(x_t) - y_{t+1} \rangle \), we have
\[
\langle \nabla_y g(x_t, y_t) - v_t, y^*(x_t) - y_{t+1} \rangle = \langle \nabla_y g(x_t, y_t) - v_t, y^*(x_t) - y_t \rangle + \langle \nabla_y g(x_t, y_t) - v_t, y_t - y_{t+1} \rangle \\
\geq -\frac{1}{\mu} \| \nabla_y g(x_t, y_t) - v_t \|^2 - \mu \| y^*(x_t) - y_t \|^2 - \frac{1}{\mu} \| \nabla_y g(x_t, y_t) - v_t \|^2 - \frac{\mu}{4} \| y_t - y_{t+1} \|^2 \\
= -\frac{2}{\mu} \| \nabla_y g(x_t, y_t) - v_t \|^2 - \frac{3}{2} \| y^*(x_t) - y_t \|^2 - \frac{\mu}{4} \| y_t - y_{t+1} \|^2.
\] (23)

By plugging the inequalities (22) and (23) into (20), we obtain
\[
\frac{1}{2 \eta_t \lambda} \| y_{t+1} - y^*(x_t) \|^2 \\
\leq \left( \frac{1}{2 \eta_t} - \frac{\mu}{4} \right) \| y_t - y^*(x_t) \|^2 + \left( \frac{\eta_t}{2 \lambda} + \frac{\mu}{4} + \frac{L}{2} - \frac{1}{\lambda} \right) \| y_{t+1} - y_t \|^2 + \frac{\mu}{2} \| \nabla_y g(x_t, y_t) - v_t \|^2 \\
\leq \left( \frac{1}{2 \eta_t} - \frac{\mu}{4} \right) \| y_t - y^*(x_t) \|^2 + \left( \frac{3 \lambda}{4} - \frac{1}{2 \lambda} \right) \| y_{t+1} - y_t \|^2 + \frac{\mu}{2} \| \nabla_y g(x_t, y_t) - v_t \|^2 \\
= \left( \frac{1}{2 \eta_t} - \frac{\mu}{4} \right) \| y_t - y^*(x_t) \|^2 - \left( \frac{3}{8 \lambda} + \frac{1}{8 \lambda} - \frac{3 \lambda}{4} \right) \| y_{t+1} - y_t \|^2 + \frac{\mu}{2} \| \nabla_y g(x_t, y_t) - v_t \|^2 \\
\leq \left( \frac{1}{2 \eta_t} - \frac{\mu}{4} \right) \| y_t - y^*(x_t) \|^2 - \frac{3}{8 \lambda} \| y_{t+1} - y_t \|^2 + \frac{\mu}{2} \| \nabla_y g(x_t, y_t) - v_t \|^2,
\] (24)

where the second inequality holds by \( L \geq \mu \) and \( 0 < \eta_t \leq 1 \), and the last inequality is due to \( 0 < \lambda \leq \frac{1}{4 \mu} \). It implies that
\[
\| y_{t+1} - y^*(x_t) \|^2 \leq \left( 1 - \frac{\eta_t \mu \lambda}{2} \right) \| y_t - y^*(x_t) \|^2 - \frac{3 \eta_t}{4} \| y_{t+1} - y_t \|^2 + \frac{4 \eta_t \lambda}{\mu} \| \nabla_y g(x_t, y_t) - v_t \|^2.
\] (25)

Next, we decompose the term \( \| y_{t+1} - y^*(x_{t+1}) \|^2 \) as follows:
\[
\| y_{t+1} - y^*(x_{t+1}) \|^2 = \| y_{t+1} - y^*(x_t) + y^*(x_t) - y^*(x_{t+1}) \|^2 \\
= \| y_{t+1} - y^*(x_t) \|^2 + 2 \langle y_{t+1} - y^*(x_t), y^*(x_t) - y^*(x_{t+1}) \rangle + \| y^*(x_t) - y^*(x_{t+1}) \|^2 \\
\leq \left( 1 + \frac{\eta_t \mu \lambda}{4} \right) \| y_{t+1} - y^*(x_t) \|^2 + \left( 1 + \frac{4 \eta_t \mu \lambda}{\eta_t \mu \lambda} \right) \| y^*(x_t) - y^*(x_{t+1}) \|^2 \\
\leq \left( 1 + \frac{\eta_t \mu \lambda}{4} \right) \| y_{t+1} - y^*(x_t) \|^2 + \left( 1 + \frac{4 \eta_t \mu \lambda}{\eta_t \mu \lambda} \right) \| x_t - x_{t+1} \|^2.
\] (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 $x_{t+1} = x_t + \eta_t(x_t - x^*)$.

By combining the above inequalities (25) and (26), we have
\[
\|y_{t+1} - y^*(x_{t+1})\|^2 \leq (1 + \frac{\eta_t \lambda}{4})(1 - \frac{\eta_t \lambda}{2})\|y_t - y^*(x_t)\|^2 - (1 + \frac{\eta_t \lambda}{4})\|\tilde{y}_{t+1} - y_t\|^2 \\
+ (1 + \frac{\eta_t \lambda}{2})\|\nabla_y g(x_t, y_t) - v_t\|^2 + (1 + \frac{4}{\eta_t \lambda})\kappa^2\|x_t - x_{t+1}\|^2.
\]

Since $0 < \eta_t \leq 1$, $0 < \lambda \leq \frac{1}{6\mu}$ and $L \geq \mu$, we have $\lambda \leq \frac{1}{6\mu} \leq \frac{1}{\eta_t \lambda}$ and $\eta_t \leq 1 \leq \frac{1}{6\mu \lambda}$. Then by using $\geq 1$, we have
\[
(1 + \frac{\eta_t \lambda}{4})(1 - \frac{\eta_t \lambda}{2}) = 1 - \frac{\eta_t \lambda}{4} + \frac{\eta_t \lambda}{2} - \frac{\eta_t \lambda^2}{8} \leq 1 - \frac{\eta_t \lambda}{4}.
\]

Thus we have
\[
\|y_{t+1} - y^*(x_{t+1})\|^2 \leq (1 - \frac{\eta_t \lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t \lambda^2}{4}\|v_t\|^2 \\
+ \frac{25\eta_t \lambda}{6\mu}\|\nabla_y g(x_t, y_t) - v_t\|^2 + \frac{25\eta_t^2 \kappa^2}{6\eta_t \lambda \mu}\|x_{t+1} - x_t\|^2.
\]

\[\blacksquare\]

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 $0 < \gamma \leq \frac{3\mu}{2\mu + L F}$, $0 < \lambda < \frac{1}{L F}$, $K = \log (T)/\log (\frac{1}{1 - \rho}) + 1$ and $\|y^0 - y^*(x_t)\|^2 \leq \Delta$ for all $t \geq 0$, we have
\[
\frac{1}{T} \sum_{t=0}^{T-1} \|G_t\|^2 \leq \frac{16(\Phi(x_0) - \Phi^*)}{3T \gamma \rho} + \frac{22L_1^2}{\rho^2 T} + \frac{22L_3^2}{\rho^2 T^2}.
\]

where $L_1 = \frac{L(L + \mu)}{\mu}$, $L_2 = \frac{2C\mu(L + \mu + L_{\text{avg}} \mu^2)}{\mu^2}$ and $L_3 = \frac{L C_1}{\mu}$.

**Proof.** According to the above Lemma 2, the function $F(x)$ has $L F$-Lipschitz continuous gradient. Let $\tilde{G}_t = \frac{1}{2} (x_t - x_{t+1})$, we have
\[
F(x_{t+1}) \leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{L F}{2} \|x_{t+1} - x_t\|^2
\]
\[
= F(x_t) - \gamma \langle \nabla F(x_t), \tilde{G}_t \rangle + \frac{\gamma^2 L F}{2} \|\tilde{G}_t\|^2
\]
\[
= F(x_t) - \gamma \langle \frac{\partial f(x_t, y_t K)}{\partial x}, \tilde{G}_t \rangle + \gamma \langle \nabla F(x_t), \tilde{G}_t \rangle + \frac{\gamma^2 L F}{2} \|\tilde{G}_t\|^2
\]
\[
\leq F(x_t) - \gamma \rho \|\tilde{G}_t\|^2 - h(x_t + h(x_t)) + h(x_t) + \gamma \|\nabla F(x_t), \tilde{G}_t \rangle + \frac{\gamma^2 L F}{2} \|\tilde{G}_t\|^2
\]
\[
\leq F(x_t) + \frac{\gamma^2 L F}{2} - \gamma \rho \|\tilde{G}_t\|^2 - h(x_t + h(x_t)) + h(x_t) + \gamma \|\nabla F(x_t), \tilde{G}_t \rangle - \|\nabla F(x_t)\|^2,
\]

where the second last inequality holds by the above Lemma 9 and the last inequality holds by the following inequality
\[
\langle \frac{\partial f(x_t, y_t K)}{\partial x}, \tilde{G}_t \rangle \leq \|\nabla F(x_t), \tilde{G}_t \rangle - \|\nabla F(x_t)\| \|\tilde{G}_t\| \leq \frac{1}{\rho} \|\nabla F(x_t)\|^2 + \frac{\rho}{4} \|\tilde{G}_t\|^2.
\]

17
According to the above Lemma[7], we have
\[ \| \frac{\partial f(x_t, y_t^k)}{\partial x} - \nabla F(x_t) \|^2 \leq 3(L_1^2(1 - \lambda \mu)^K + L_2^2(1 - \lambda \mu)^{K-1})\| y_t^0 - y^*(x_t) \|^2 + 3L_3^2(1 - \lambda \mu)^{2K} \]
\[ \leq 3\Delta \| (1 - \lambda \mu)^K + 3\Delta L_2^2(1 - \lambda \mu)^{K-1} + 3L_3^2(1 - \lambda \mu)^{2K}, \tag{31} \]
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
\[ \Phi(x_{t+1}) \leq \Phi(x_t) + \frac{\gamma}{2} \frac{L_\rho}{2} \| \tilde{G}_t \|^2 + \frac{\gamma}{\rho} \frac{\partial f(x_t, y_t^k)}{\partial x} - \nabla F(x_t) \|^2 \]
\[ \leq \Phi(x_t) - \frac{3\gamma}{8} \| \tilde{G}_t \|^2 + \frac{3\gamma L_2^2}{\rho^2} (1 - \lambda \mu)^K + \frac{3\gamma L_3^2}{\rho^2} (1 - \lambda \mu)^{K-1} + \frac{3\gamma L_3^2}{\rho} (1 - \lambda \mu)^{2K}, \tag{32} \]
where the last inequality is due to \( 0 < \gamma \leq \frac{\lambda \mu}{4L_\rho} \) and the above inequality (31). According to Lemma[10] the difference between \( \tilde{G}_t \) and \( \bar{G}_t \) are bounded, we have
\[ \| \tilde{G}_t \|^2 \leq 2\| \bar{G}_t \|^2 + 2\| \bar{G}_t - \tilde{G}_t \|^2 \]
\[ \leq 2\| \bar{G}_t \|^2 + \frac{2}{\rho^2} \| w_t - \nabla F(x_t) \|^2 \]
\[ = 2\| \bar{G}_t \|^2 + \frac{2}{\rho^2} \| \partial f(x_t, y_t^k) \|^2 - \nabla F(x_t) \|^2 \]
\[ \leq 2\| \bar{G}_t \|^2 + \frac{6L_1^2}{\rho^2} (1 - \lambda \mu)^K + \frac{6L_2^2}{\rho^2} (1 - \lambda \mu)^{K-1} + \frac{6L_3^2}{\rho^2} (1 - \lambda \mu)^{2K}. \tag{33} \]
Thus we have
\[ -\| \tilde{G}_t \|^2 \leq -\frac{1}{2} \| \tilde{G}_t \|^2 + \frac{3\gamma L_2^2}{\rho^2} (1 - \lambda \mu)^K + \frac{3\gamma L_3^2}{\rho^2} (1 - \lambda \mu)^{K-1} + \frac{3\gamma L_3^2}{\rho} (1 - \lambda \mu)^{2K}. \tag{34} \]
By plugging (34) into (29), we have
\[ \Phi(x_{t+1}) \leq \Phi(x_t) - \frac{3\gamma}{16} \| \tilde{G}_t \|^2 + \frac{3\gamma}{8} \left( \frac{3\gamma L_2^2}{\rho^2} (1 - \lambda \mu)^K + \frac{3\gamma L_3^2}{\rho^2} (1 - \lambda \mu)^{K-1} + \frac{3\gamma L_3^2}{\rho^2} (1 - \lambda \mu)^{2K} \right) \]
\[ + \frac{3\gamma L_2^2}{\rho^2} (1 - \lambda \mu)^K + \frac{3\gamma L_3^2}{\rho^2} (1 - \lambda \mu)^{K-1} + \frac{3\gamma L_3^2}{\rho^2} (1 - \lambda \mu)^{2K} \]
\[ = \Phi(x_t) - \frac{3\gamma}{16} \| \tilde{G}_t \|^2 + \frac{33\gamma L_2^2}{8\rho} (1 - \lambda \mu)^K + \frac{33\gamma L_3^2}{8\rho} (1 - \lambda \mu)^{K-1} + \frac{33\gamma L_3^2}{8\rho} (1 - \lambda \mu)^{2K}. \tag{35} \]
Thus, we have
\[ 1 - \frac{T}{T-1} \| G_t \|^2 \leq \frac{16(\Phi(x_0) - \Phi(x_T))}{3T \gamma \rho} + \frac{22\Delta L_1^2}{\rho^2} (1 - \lambda \mu)^K + \frac{22\Delta L_2^2}{\rho^2} (1 - \lambda \mu)^{K-1} + \frac{22L_3^2}{\rho^2} (1 - \lambda \mu)^{2K} \]
\[ \leq \frac{16(\Phi(x_0) - \Phi^*)}{3T \gamma \rho} + \frac{22\Delta L_1^2}{\rho^2} (1 - \lambda \mu)^K + \frac{22\Delta L_2^2}{\rho^2} (1 - \lambda \mu)^{K-1} + \frac{22L_3^2}{\rho^2} (1 - \lambda \mu)^{2K} \]
\[ \leq \frac{16(\Phi(x_0) - \Phi^*)}{3T \gamma \rho} + \frac{22\Delta L_1^2}{\rho^2 T} + \frac{22\Delta L_2^2}{\rho^2 T} + \frac{22L_3^2}{\rho^2 T}, \tag{36} \]
where the last inequality holds by \( K = \log(T)/\log(\frac{1}{1-\lambda \mu}) + 1 \) and \( \lambda < \frac{1}{T} \).

\[ \Box \]

### B.2 Convergence Analysis of the SBio-BreD Algorithm

In this subsection, we provide the convergence analysis of our SBio-BreD algorithm. Let \( R(x_t, y_t) = \nabla f(x_t, y_t, B_t) \) 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 \( K = \frac{\log(T)}{\eta} \), \( 0 < \eta = \eta_t \leq 1 \), \( 0 < \gamma \leq \min \left( \frac{3\gamma}{4L_\rho}, \frac{\eta \mu \Delta}{3 \mu \Delta}, \frac{\eta \mu \lambda}{3 \mu \lambda} \right) \) and \( 0 < \lambda < \frac{1}{6T} \), we have
\[ \frac{1}{T} \sum_{t=1}^T E \| \tilde{G}_t \|^2 \leq \frac{32(\Phi(x_0) - \Phi^*)}{3T \gamma \rho} + \frac{32\Delta}{3T \gamma \rho} + \frac{752\sigma^2}{3 \rho^2 b} + \frac{400\eta \lambda^2}{9 \gamma \mu b} + \frac{752}{3 \rho^2 T^2}, \tag{37} \]
where $\Delta = ||y_t - y^*(x_0)||^2$.

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

$$F(x_{t+1}) \leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{L_F}{2} ||x_{t+1} - x_t||^2$$

$$= F(x_t) - \gamma \langle \nabla F(x_t), \hat{G}_t \rangle + \gamma^2 \frac{L_F}{2} ||\hat{G}_t||^2$$

$$= F(x_t) - \gamma(w_t, \hat{G}_t) + \gamma(w_t - \nabla F(x_t), \hat{G}_t) + \gamma^2 \frac{L_F}{2} ||\hat{G}_t||^2$$

$$\leq F(x_t) - \gamma \rho ||\hat{G}_t||^2 - h(x_{t+1}) + h(x_t) + \gamma(w_t - \nabla F(x_t), \hat{G}_t) + \gamma^2 \frac{L_F}{2} ||\hat{G}_t||^2$$

$$\leq F(x_t) + \left( \frac{\gamma^2 L_F}{2} - \frac{3\gamma \rho}{4} \right) ||\hat{G}_t||^2 - h(x_{t+1}) + h(x_t) + \frac{\gamma}{\rho} ||w_t - \nabla F(x_t)||^2, \quad (38)$$

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

$$(w_t - \nabla F(x_t), \hat{G}_t) \leq ||w_t - \nabla F(x_t)|| ||\hat{G}_t||$$

$$\leq \frac{1}{\rho} ||w_t - \nabla F(x_t)||^2 + \frac{\rho}{4} ||\hat{G}_t||^2. \quad (39)$$

According to the above Lemma 2, we have

$$||w_t - \nabla F(x_t)||^2 = ||w_t - \nabla f(x_t, y_t) + \nabla f(x_t, y_t) - \nabla f(x_t)||^2$$

$$\leq 2 ||w_t - \nabla f(x_t, y_t)||^2 + 2 \nabla f(x_t, y_t) - \nabla F(x_t)||^2$$

$$\leq 2 ||w_t - \nabla f(x_t, y_t)||^2 + 2 L^{2}_y ||y_t - y^*(x_t)||^2. \quad (40)$$

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

$$\Phi(x_{t+1}) \leq \Phi(x_t) + \left( \frac{\gamma^2 L_F}{2} - \frac{3\gamma \rho}{4} \right) ||\hat{G}_t||^2 + \frac{2\gamma}{\rho} ||w_t - \nabla f(x_t, y_t)||^2 + \frac{2 L^{2}_y \gamma}{\rho} ||y_t - y^*(x_t)||^2$$

$$\leq \Phi(x_t) - \frac{3\gamma \rho}{8} ||\hat{G}_t||^2 + \frac{2\gamma}{\rho} ||w_t - \nabla f(x_t, y_t)||^2 + \frac{2 L^{2}_y \gamma}{\rho} ||y_t - y^*(x_t)||^2, \quad (41)$$

where the last inequality is due to $0 < \gamma \leq \frac{3\rho}{4 L^{2}_F}$. According to Lemma 10, the difference between $\hat{G}_t$ and $\hat{G}_t$ are bounded, we have

$$||\hat{G}_t||^2 \leq 2 ||\hat{G}_t||^2 + 2 ||\hat{G}_t - \hat{G}_t||^2$$

$$\leq 2 ||\hat{G}_t||^2 + \frac{2}{\rho^2} ||w_t - \nabla F(x_t)||^2$$

$$\leq 2 ||\hat{G}_t||^2 + \frac{4}{\rho^2} ||w_t - \nabla f(x_t, y_t)||^2 + \frac{4 L^{2}_y \gamma}{\rho^2} ||y_t - y^*(x_t)||^2. \quad (42)$$

Thus we have

$$-||\hat{G}_t||^2 \leq - \frac{1}{2} ||\hat{G}_t||^2 + \frac{2}{\rho^2} ||w_t - \nabla f(x_t, y_t)||^2 + \frac{2 L^{2}_y \gamma}{\rho^2} ||y_t - y^*(x_t)||^2. \quad (43)$$

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

$$\Phi(x_{t+1}) \leq \Phi(x_t) - \frac{3\gamma \rho}{16} ||\hat{G}_t||^2 + \frac{3\gamma \rho}{8} \left( \frac{2}{\rho^2} ||w_t - \nabla f(x_t, y_t)||^2 + \frac{2 L^{2}_y \gamma}{\rho^2} ||y_t - y^*(x_t)||^2 \right)$$

$$+ \frac{2\gamma}{\rho} ||w_t - \nabla f(x_t, y_t)||^2 + \frac{2 L^{2}_y \gamma}{\rho} ||y_t - y^*(x_t)||^2$$

$$= \Phi(x_t) - \frac{3\gamma \rho}{16} ||\hat{G}_t||^2 + \frac{11\gamma}{4\rho} ||w_t - \nabla f(x_t, y_t)||^2 + \frac{11 L^{2}_y \gamma}{4\rho} ||y_t - y^*(x_t)||^2. \quad (44)$$

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

$$\Omega_t = E\left[\Phi(x_t) + ||y_t - y^*(x_t)||^2\right]. \quad (45)$$
According to Lemma 11, we have

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

(46)

Then we have

\[ \Omega_{t+1} - \Omega_t = \mathbb{E} \left[ \Phi(x_{t+1}) - \Phi(x_t) + \|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2 \right] \]

\[ \leq \frac{\gamma \eta_t}{16} \mathbb{E} \|G_t\|^2 + \frac{11 \gamma}{4\rho} \mathbb{E} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{11L^2 \gamma}{4\rho} \mathbb{E} \|y_t - y^*(x_t)\|^2 - \frac{\eta_t \lambda}{4} \mathbb{E} \|y_t - y^*(x_t)\|^2 \\
- \frac{3\eta_t \lambda}{4} \mathbb{E} \|v_t\|^2 + \frac{25\eta_t \lambda}{6\mu} \mathbb{E} \|\nabla_y g(x_t, y_t) - v_t\|^2 + \frac{25\kappa^2}{6\eta_t \lambda} \mathbb{E} \|x_{t+1} - x_t\|^2 \\
= \frac{\gamma \eta_t}{16} \mathbb{E} \|G_t\|^2 + \frac{11 \gamma}{4\rho} \mathbb{E} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{11L^2 \gamma}{4\rho} \mathbb{E} \|y_t - y^*(x_t)\|^2 - \frac{\eta_t \lambda}{4} \mathbb{E} \|y_t - y^*(x_t)\|^2 \\
- \frac{3\eta_t \lambda}{4} \mathbb{E} \|v_t\|^2 + \frac{25\eta_t \lambda}{6\mu} \mathbb{E} \|\nabla_y g(x_t, y_t) - v_t\|^2 + \frac{25\kappa^2}{6\eta_t \lambda} \mathbb{E} \|x_{t+1} - x_t\|^2 \\
\leq \frac{\gamma \eta_t}{16} \mathbb{E} \|\tilde{G}_t\|^2 + \frac{47\gamma}{4\rho} \mathbb{E} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{4L^2 \gamma}{\rho^2} \mathbb{E} \|y_t - y^*(x_t)\|^2. \]  

(47)

where the last inequality holds by the following inequality

\[ \|\tilde{G}_t\|^2 \leq 2\|G_t\|^2 + 2\|\tilde{G}_t - G_t\|^2 \\
\leq 2\|G_t\|^2 + \frac{2}{\rho^2} \|w_t - \nabla F(x_t)\|^2 \\
\leq 2\|G_t\|^2 + \frac{4}{\rho^2} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{4L^2 \gamma}{\rho^2} \mathbb{E} \|y_t - y^*(x_t)\|^2. \]  

(48)

Let \( \eta = \eta_t \) for all \( t \geq 0 \). By using \( 0 < \gamma \leq \frac{9\mu \lambda}{8000\alpha} \), we have

\[ \frac{3\gamma \eta_t}{32} \geq \frac{50\kappa^2 \gamma^2}{6\eta_t \lambda \mu}, \quad \frac{3L^2 \gamma}{16\rho} \geq \frac{50\kappa^2 \gamma^2 L^2}{3\eta_t \mu \rho^2}, \quad \frac{3\gamma}{16 \rho} \geq \frac{50\kappa^2 \gamma^2}{3\eta_t \lambda \rho^2}. \]  

(49)

Let \( \frac{25\lambda}{4} \geq \frac{47L^2 \gamma}{4\rho} \), we have \( 0 < \gamma \leq \frac{25\lambda}{47L^2 \gamma} \). Given \( 0 < \gamma \leq \min \left( \frac{25\lambda}{47L^2 \gamma}, \frac{9\mu \lambda}{8000 \alpha} \right) \), we have

\[ \Omega_{t+1} - \Omega_t \leq \frac{3\gamma \eta_t}{32} \mathbb{E} \|G_t\|^2 + \frac{47\gamma}{4\rho} \mathbb{E} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{25\eta_t \lambda}{6\mu} \mathbb{E} \|v_t - \nabla_y g(x_t, y_t)\|^2. \]  

(50)

Thus, we have

\[ \mathbb{E} \|G_t\|^2 \]

\[ \leq \frac{32(\Omega_t - \Omega_{t+1})}{3\gamma \rho} + \frac{376}{3\rho^2} \mathbb{E} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{400\eta_t \lambda}{9\gamma \rho \mu} \mathbb{E} \|v_t - \nabla_y g(x_t, y_t)\|^2 \\
= \frac{32(\Omega_t - \Omega_{t+1})}{3\gamma \rho} + \frac{376}{3\rho^2} \mathbb{E} \|w_t - \nabla f(x_t, y_t) - R(x_t, y_t) + R(x_t, y_t)\|^2 + \frac{400\eta_t \lambda}{9\gamma \rho \mu} \mathbb{E} \|v_t - \nabla_y g(x_t, y_t)\|^2 \\
\leq \frac{32(\Omega_t - \Omega_{t+1})}{3\gamma \rho} + \frac{752}{3\rho^2} \mathbb{E} \|w_t - \nabla f(x_t, y_t) - R(x_t, y_t)\|^2 + \frac{752}{3\rho^2} \|R(x_t, y_t)\|^2 + \frac{400\eta_t \lambda}{9\gamma \rho \mu} \mathbb{E} \|v_t - \nabla_y g(x_t, y_t)\|^2 \\
\leq \frac{32(\Omega_t - \Omega_{t+1})}{3\gamma \rho} + \frac{752\gamma^2}{3\rho^2 \beta} + \frac{400\eta_t \lambda \sigma^2}{9\gamma \rho \mu \beta} + \frac{752}{3\rho^2} \mathbb{E} \|R(x_t, y_t)\|^2. \]  

(51)

where the last inequality holds by Assumption 6 and \( w_t = \nabla f(x_t, y_t; B_t) = \frac{1}{b} \sum_{i \in B_t} \nabla f(x_t, y_t, \xi_i), v_t = \nabla_y g(x_t, y_t; B_t) = \frac{1}{b} \sum_{i \in B_t} \nabla_y g(x_t, y_t, \xi_i). \)
Taking average over $t = 0, 2, \cdots, T - 1$ on both sides of the above inequality (51), we have

$$
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\| \bar{g}_t \|^2] \leq \frac{32(\Omega_0^2 - \Omega_T^2)}{3T\gamma \rho} + \frac{752\sigma^2}{3\rho b} + \frac{400\eta \lambda^2}{9\gamma \rho b} + \frac{752}{3\rho^2 b} \mathbb{E}[\| R(x_t, y_t) \|^2] \\
= \frac{32(\Phi(x_0) + \|g_0 - g^*(x_0)\|^2)}{3T\gamma \rho} - \frac{32(\Phi(x_T) + \|g_T - g^*(x_T)\|^2)}{3T\gamma \rho} + \frac{752\sigma^2}{3\rho b} + \frac{400\eta \lambda^2}{9\gamma \rho b} + \frac{752}{3\rho^2 b} \mathbb{E}[\| R(x_t, y_t) \|^2]
$$

where the last inequality holds by Assumption 5 and $\mathbb{E}[\| R(x_t, y_t) \|^2] \leq \frac{1}{T}$ for all $t \geq 1$ by choosing $K = \frac{\log(\frac{L\Phi(x_0)}{\mu^2})}{\rho}$.

\[\square\]

### 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 $R(x_t, y_t) = \nabla f(x_t, y_t) - \nabla f(x_t, y_t; \bar{\xi}_t)$ for all $t \geq 0$, when $\mod(t, q) = 0$, let $R(x_t, y_t) = \nabla f(x_t, y_t) - \nabla f(x_t, y_t; \bar{\xi}_t)$.

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

$$
\mathbb{E}[\| \nabla f(x_t, y_t) + R(x_t, y_t) - w_t \|^2] \leq \frac{2L^2}{b_1} \sum_{i=(t-1)q}^{t-1} \left( \mathbb{E}[\| x_{i+1} - x_i \|^2] + \mathbb{E}[\| y_{i+1} - y_i \|^2] \right) + \frac{\sigma^2}{b},
$$

(53)

$$
\mathbb{E}[\| \nabla g(x_t, y_t) - v_t \|^2] \leq \frac{2L^2}{b_1} \sum_{i=(t-1)q}^{t-1} \left( \mathbb{E}[\| x_{i+1} - x_i \|^2] + \mathbb{E}[\| y_{i+1} - y_i \|^2] \right) + \frac{\sigma^2}{b},
$$

(54)

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

$$
w_t - w_{t-1} = \nabla f(x_t, y_t; \bar{\xi}_t) - \nabla f(x_t, y_{t-1}; \bar{\xi}_t).
$$

(55)

Then we have

$$
\mathbb{E}[\| \nabla f(x_t, y_t) + R(x_t, y_t) - w_t \|^2]
\begin{align*}
&= \mathbb{E}[\| \nabla f(x_t, y_t) + R(x_t, y_t) - w_{t-1} - (w_t - w_{t-1}) \|^2] \\
&= \mathbb{E}[\| \nabla f(x_t, y_t) + R(x_t, y_t) - w_{t-1} - \nabla f(x_{t-1}, y_{t-1}; \bar{\xi}_{t-1}) \|^2] \\
&= \mathbb{E}[\| \nabla f(x_{t-1}, y_{t-1}) + R(x_{t-1}, y_{t-1}) - w_{t-1} - \nabla f(x_{t-1}, y_{t-1}) + \nabla f(x_{t-1}, y_{t-1}) - R(x_{t-1}, y_{t-1}) \|^2] \\
&= \mathbb{E}[\| \nabla f(x_{t-1}, y_{t-1}) + R(x_{t-1}, y_{t-1}) - w_{t-1} - \nabla f(x_{t-1}, y_{t-1}) + \nabla f(x_{t-1}, y_{t-1}) - R(x_{t-1}, y_{t-1}) \|^2] \\
&= \mathbb{E}[\| \nabla f(x_{t-1}, y_{t-1}) + R(x_{t-1}, y_{t-1}) - w_{t-1} \|^2] + \frac{1}{b_1} \mathbb{E}[\| \nabla f(x_{t-1}, y_{t-1}) + R(x_{t-1}, y_{t-1}) - R(x_{t-1}, y_{t-1}) \|^2] \\
&\leq \mathbb{E}[\| \nabla f(x_{t-1}, y_{t-1}) + R(x_{t-1}, y_{t-1}) - w_{t-1} \|^2] + \frac{1}{b_1} \mathbb{E}[\| \nabla f(x_{t-1}, y_{t-1}) - \nabla f(x_{t-1}, y_{t-1}) \|^2] \\
&\leq \mathbb{E}[\| \nabla f(x_{t-1}, y_{t-1}) + R(x_{t-1}, y_{t-1}) - w_{t-1} \|^2] + \frac{2L^2}{b_1} \left( \| x_{t-1} - x_{t-1} \|^2 + \| y_{t-1} - y_{t-1} \|^2 \right),
\end{align*}
$$

(56)

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

21
Throughout the paper, let \( n_t = \lfloor t/q \rfloor \) such that \((n_t - 1)q \leq t \leq n_tq - 1\). Telescoping (56) over \( t \) from \((n_t - 1)q + 1 \) to \( t \), we have
\[
\mathbb{E}[\nabla f(x_t, y_t) + R(x_t, y_t) - w_t]^2 \leq \frac{2L^2_2}{b_1} \sum_{t=(n_t-1)q}^{t-1} (\mathbb{E}||x_{i+1} - x_t||^2 + \mathbb{E}||y_{i+1} - y_i||^2)
+ \mathbb{E}[\nabla f(x_{(n_t-1)q}, y_{(n_t-1)q}) + R(x_{(n_t-1)q}, y_{(n_t-1)q}) - w_{(n_t-1)q}]^2
\leq \frac{2L^2_2}{b_1} \sum_{t=(n_t-1)q}^{t-1} (\mathbb{E}||x_{i+1} - x_t||^2 + \mathbb{E}||y_{i+1} - y_i||^2) + \frac{\sigma^2}{b},
\]
where the last inequality is due to Assumption 6 and \( w_{(n_t-1)q} = \frac{1}{t} \sum_{i \in \mathcal{F}_{(n_t-1)q}} \nabla f(x_{(n_t-1)q}, y_{(n_t-1)q}, \xi_{(n_t-1)q}). \) Similarly, we can obtain
\[
\mathbb{E}[\nabla y g(x_t, y_t) - v_t]^2 \leq \frac{2L^2_2}{b_1} \sum_{t=(n_t-1)q}^{t-1} (\mathbb{E}||x_{i+1} - x_t||^2 + \mathbb{E}||y_{i+1} - y_i||^2) + \frac{\sigma^2}{b},
\]
\( \square \)

**Theorem 6.** (Restatement of Theorem 3) Suppose the sequence \( \{x_t, y_t\}_{t=1}^T \) is generated from Algorithm 3. Let \( b_1 = q, K = \frac{L}{\rho} \log(\frac{LcT}{\rho q}) \), \( 0 < \eta = \eta_t \leq 1, 0 < \gamma \leq \min\left(\frac{3\rho}{25L_K}, \frac{3\rho}{21L^2\rho}, \frac{20\rho\gamma\lambda}{7}, \frac{\rho \gamma \lambda}{16}, \frac{9\rho \gamma}{400}\right) \) and \( 0 < \lambda \leq \min(\frac{3\rho}{25L_K}, \frac{9\rho \gamma}{400}) \).

Let \( F(x) = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\mathcal{G}_t] \), we have
\[
1 + \frac{T-1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\mathcal{G}_t] \leq \frac{32(\Phi(x_0) - \Phi^*)}{3T\gamma \rho} + \frac{32\Delta}{3T\gamma \rho} + \frac{152}{3T^2 \rho^2} + \frac{4}{\eta \rho \gamma} + \frac{1}{16L} + \frac{1}{L_K} \frac{\sigma^2}{b}.
\]
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 \( \tilde{G}_t = \frac{1}{T}(x_t - x_{t+1}) \), we have
\[
F(x_{t+1}) \leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{L_F}{2} \|x_{t+1} - x_t\|^2
= F(x_t) - \gamma \langle \nabla F(x_t), \tilde{G}_t \rangle + \frac{\gamma^2 L_F}{2} \|\tilde{G}_t\|^2
= F(x_t) - \gamma \rho \|\tilde{G}_t\|^2 - h(x_{t+1}) + h(x_t) + \gamma \rho \|\nabla F(x_t)\| \tilde{G}_t + \frac{\gamma^2 L_F}{2} \|\tilde{G}_t\|^2
\leq F(x_t) + \frac{\gamma^2 L_F}{2} - \frac{3\gamma \rho}{4} \|\tilde{G}_t\|^2 - h(x_{t+1}) + h(x_t) + \frac{\gamma}{\rho} \|w_t - \nabla F(x_t)\|^2,
\]
where the second last inequality holds by the above Lemma 5 and the last inequality holds by the following inequality
\[
\langle w_t - \nabla F(x_t), \tilde{G}_t \rangle \leq \|w_t - \nabla F(x_t)\| \|\tilde{G}_t\|
\leq \frac{1}{\rho} \|w_t - \nabla F(x_t)\|^2 + \frac{\rho}{4} \|\tilde{G}_t\|^2.
\]

According to the above Lemma 2, we have
\[
\|w_t - \nabla F(x_t)\|^2 = \|w_t - \nabla f(x_t, y_t) + \nabla f(x_t, y_t) - \nabla F(x_t)\|^2
\leq 2 \|w_t - \nabla f(x_t, y_t)\|^2 + 2 \|\nabla f(x_t, y_t) - \nabla F(x_t)\|^2
\leq 2 \|w_t - \nabla f(x_t, y_t)\|^2 + 2L^2_2 \rho \|y_t - y^*(x_t)\|^2.
\]

Let \( \Phi(x) = F(x) + h(x) \), plugging (65) into (66), we have
\[
\Phi(x_{t+1}) \leq \Phi(x_t) + \frac{\gamma^2 L_F}{2} - \frac{3\gamma \rho}{4} \|\tilde{G}_t\|^2 + \frac{2\gamma}{\rho} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{2L^2_2 \gamma}{\rho} \|y_t - y^*(x_t)\|^2
\leq \Phi(x_t) - \frac{3\rho}{8} \|\tilde{G}_t\|^2 + \frac{2\gamma}{\rho} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{2L^2_2 \gamma}{\rho} \|y_t - y^*(x_t)\|^2
\leq \Phi(x_t) - \frac{3\rho}{16} \|\tilde{G}_t\|^2 - \frac{3\rho}{16} \|x_{t+1} - x_t\|^2 + \frac{2\gamma}{\rho} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{2L^2_2 \gamma}{\rho} \|y_t - y^*(x_t)\|^2.
\]

(63)
where the second last inequality is due to $0 < \gamma \leq \frac{3\eta}{8\rho}$. By using Lemma 10, the difference between $\tilde{G}_t$ and $G_t$ are bounded, we have

$$
\|G_t\|^2 \leq 2\|\tilde{G}_t\|^2 + 2\|\tilde{G}_t - G_t\|^2
$$

$$
\leq 2\|\tilde{G}_t\|^2 + \frac{2}{\rho^2} \|w_t - \nabla F(x_t)\|^2
$$

$$
\leq 2\|\tilde{G}_t\|^2 + \frac{4}{\rho^2} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{4L_y^2}{\rho^2} \|y_t - y^*(x_t)\|^2.
$$

(64)

Thus we have

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

(65)

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

$$
\Phi(x_{t+1}) \leq \Phi(x_t) - \frac{3\gamma}{32} \|G_t\|^2 - \frac{3\gamma}{16} \left( \frac{2}{\rho^2} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{2L_y^2}{\rho^2} \|y_t - y^*(x_t)\|^2 \right)
$$

$$
- \frac{3\rho}{16\gamma} \|x_{t+1} - x_t\|^2 + \frac{2\gamma}{\rho} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{2L_y^2\gamma}{\rho} \|y_t - y^*(x_t)\|^2
$$

$$
= \Phi(x_t) - \frac{3\gamma}{32} \|G_t\|^2 - \frac{3\rho}{16\gamma} \|x_{t+1} - x_t\|^2 + \frac{19\gamma}{8\rho} \|w_t - \nabla f(x_t, y_t)\|^2 + \frac{19L_y^2\gamma}{8\rho} \|y_t - y^*(x_t)\|^2.
$$

(66)

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

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

(67)

According to Lemma 11, we have

$$
\|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2 \leq -\frac{\eta\mu\lambda}{4} \|y_t - y^*(x_t)\|^2 - \frac{3\eta\lambda^2}{4} \|v_t\|^2
$$

$$
+ \frac{25\eta\lambda}{6\mu} \|\nabla g(x_t, y_t) - v_t\|^2 + \frac{25\kappa^2}{6\eta\mu\lambda} \|x_{t+1} - x_t\|^2.
$$

(68)

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

$$
\Omega_{t+1} - \Omega_t = \mathbb{E}\left[\Phi(x_{t+1}) - \Phi(x_t) + \|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2\right]
$$

$$
\leq -\frac{3\gamma}{32} \mathbb{E}\|G_t\|^2 - \frac{3\gamma}{16\gamma} \mathbb{E}\|x_{t+1} - x_t\|^2 + \frac{19\gamma}{8\rho} \mathbb{E}\|w_t - \nabla f(x_t, y_t)\|^2 + \frac{19L_y^2\gamma}{8\rho} \mathbb{E}\|y_t - y^*(x_t)\|^2
$$

$$
- \frac{\eta\mu\lambda}{4} \mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{3\eta\lambda^2}{4} \mathbb{E}\|v_t\|^2 + \frac{25\eta\lambda}{6\mu} \mathbb{E}\|\nabla g(x_t, y_t) - v_t\|^2 + \frac{25\kappa^2}{6\eta\mu\lambda} \mathbb{E}\|x_{t+1} - x_t\|^2
$$

$$
= -\frac{3\gamma}{32} \mathbb{E}\|G_t\|^2 - \left( \frac{3\rho}{16\gamma} - \frac{25\kappa^2}{6\eta\mu\lambda}\right) \mathbb{E}\|x_{t+1} - x_t\|^2 + \frac{19\gamma}{8\rho} \mathbb{E}\|w_t - \nabla f(x_t, y_t)\|^2
$$

$$
- \left( \frac{\eta\mu\lambda}{4} - \frac{19L_y^2\gamma}{8\rho}\right) \mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{3\eta\lambda^2}{4} \mathbb{E}\|v_t\|^2 + \frac{25\eta\lambda}{6\mu} \mathbb{E}\|\nabla g(x_t, y_t) - v_t\|^2
$$

$$
\leq -\frac{3\gamma}{32} \mathbb{E}\|G_t\|^2 - \left( \frac{3\rho}{16\gamma} - \frac{25\kappa^2}{6\eta\mu\lambda}\right) \mathbb{E}\|x_{t+1} - x_t\|^2 + \frac{19\gamma}{4\rho} \mathbb{E}\|w_t - \nabla f(x_t, y_t) - R(x_t, y_t)\|^2
$$

$$
+ \frac{19\gamma}{4\rho} \mathbb{E}\|R(x_t, y_t)\|^2 - \left( \frac{\eta\mu\lambda}{4} - \frac{19L_y^2\gamma}{8\rho}\right) \mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{3\eta\lambda^2}{4} \mathbb{E}\|v_t\|^2
$$

$$
+ \frac{25\eta\lambda}{6\mu} \mathbb{E}\|\nabla g(x_t, y_t) - v_t\|^2
$$

$$
\leq -\frac{3\gamma}{32} \mathbb{E}\|G_t\|^2 - \left( \frac{3\rho}{16\gamma} - \frac{25\kappa^2}{6\eta\mu\lambda}\right) \mathbb{E}\|x_{t+1} - x_t\|^2 + \frac{19\gamma}{4\rho} \mathbb{E}\|w_t - \nabla f(x_t, y_t) - R(x_t, y_t)\|^2
$$

$$
+ \frac{19\gamma}{4\rho} \mathbb{E}\|R(x_t, y_t)\|^2 - \frac{3\eta\lambda^2}{4} \mathbb{E}\|v_t\|^2 + \frac{25\eta\lambda}{6\mu} \mathbb{E}\|\nabla g(x_t, y_t) - v_t\|^2,
$$

(69)

where the last inequality holds by $\gamma \leq \frac{2\eta\mu\lambda}{19L_y^2\gamma}$.
Summing over \( t = 0, 1, \ldots, T - 1 \) on both sides of (69), by Lemma 12, we have

\[
\frac{3\gamma \rho}{32} \sum_{t=0}^{T-1} \mathbb{E} \|G_t\|^2 \\
\leq \Omega_0 - \Omega_T - \left( \frac{3\rho}{16\gamma} - \frac{25\kappa^2}{6\mu \lambda} \right) \sum_{t=0}^{T-1} \mathbb{E} \|x_{t+1} - x_t\|^2 - \frac{3\eta^2}{4} \mathbb{E} \|v_t\|^2 + \frac{19\gamma}{4\rho} \sum_{t=0}^{T-1} \mathbb{E} \|R(x_t, y_t)\|^2 \\
+ \frac{19\gamma}{4\rho} \sum_{t=0}^{T-1} \left( \frac{L^4}{b_1} \sum_{i=(n_t-1)q}^{n_tq-1} \mathbb{E} \|x_{i+1} - x_i\|^2 + \mathbb{E} \|y_{i+1} - y_i\|^2 \right) + \frac{\sigma^2}{b} \\
+ \frac{25\eta \lambda}{6\mu} \sum_{t=0}^{T-1} \left( \frac{L^2}{b_1} \sum_{i=(n_t-1)q}^{n_tq-1} \mathbb{E} \|x_{i+1} - x_i\|^2 + \mathbb{E} \|y_{i+1} - y_i\|^2 \right) + \frac{\sigma^2}{b} \\
\leq \Omega_0 - \Omega_T - \left( \frac{3\rho}{16\gamma} - \frac{25\kappa^2}{6\mu \lambda} \right) \sum_{t=0}^{T-1} \mathbb{E} \|x_{t+1} - x_t\|^2 - \frac{3\eta^2}{4} \mathbb{E} \|v_t\|^2 + \frac{19\gamma}{4\rho} \sum_{t=0}^{T-1} \mathbb{E} \|R(x_t, y_t)\|^2 \\
+ \frac{19\gamma}{4\rho} \sum_{t=0}^{T-1} \left( \frac{L^4}{b_1} \sum_{i=(n_t-1)q}^{n_tq-1} \mathbb{E} \|x_{i+1} - x_i\|^2 + \mathbb{E} \|y_{i+1} - y_i\|^2 \right) + \frac{\sigma^2}{b} \\
+ \frac{25\eta \lambda}{6\mu} \sum_{t=0}^{T-1} \left( \frac{L^2}{b_1} \sum_{i=(n_t-1)q}^{n_tq-1} \mathbb{E} \|x_{i+1} - x_i\|^2 + \mathbb{E} \|y_{i+1} - y_i\|^2 \right) + \frac{\sigma^2}{b} \\
= \Omega_0 - \Omega_T - \left( \frac{3\rho}{16\gamma} - \frac{25\kappa^2}{6\mu \lambda} \right) \sum_{t=0}^{T-1} \mathbb{E} \|x_{t+1} - x_t\|^2 - \frac{3\eta^2}{4} \mathbb{E} \|v_t\|^2 + \left( \frac{19\gamma}{4\rho} + \frac{25\eta \lambda}{6\mu} \right) \frac{T \sigma^2}{b},
\]

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

Let \( b_1 = q, 0 < \gamma \leq \frac{3\rho}{38\kappa^2}, \) and \( 0 < \lambda \leq \frac{9\mu}{1000\kappa^2 \gamma^2} \), we have \( \frac{3\eta^2}{8} \geq \frac{19\gamma L^4}{4\rho b_1} \) and \( \frac{25\lambda^2}{8\mu} \geq \frac{25\lambda^2}{6\mu b_1} \), i.e., we obtain

\[
\frac{3\eta^2}{4} - \frac{19\gamma L^4 q^2 \lambda^2}{4\rho b_1} - \frac{25\lambda^2 L^2 q^2}{6\mu b_1} \geq 0.
\]

At the same time, we have \( \frac{3}{8\gamma} \leq \frac{19\gamma L^4 q}{4\rho b_1}, \) \( \frac{3}{8\gamma} \leq \frac{25\lambda^2 L^2 q}{6\mu b_1}, \) \( \frac{3}{8\gamma} \leq \frac{19\gamma L^4 q}{4\rho b_1} \) and \( \frac{3}{8\gamma} \leq \frac{25\lambda^2}{6\mu b_1} \). Thus we have \( \frac{3}{4\eta} \geq \frac{19\gamma L^4 q}{4\rho b_1} + \frac{25\lambda^2 L^2 q}{6\mu b_1} \). Let \( \gamma \leq \min \left( \frac{3}{4\eta}, \frac{9\mu \lambda}{4000\kappa^2} \right) \), we have

\[
\frac{3\rho}{16\gamma} \geq \frac{25\kappa^2}{6\mu \lambda} + \frac{3}{4\eta} \geq \frac{25\kappa^2}{6\mu \lambda} + \frac{19\gamma L^4 q}{8\rho b_1} + \frac{25\lambda^2 L^2 q}{6\mu b_1}.
\]

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

\[
\frac{3\gamma \rho}{32} \sum_{t=0}^{T-1} \mathbb{E} \|G_t\|^2 \leq \Omega_0 - \Omega_T + \frac{19\gamma}{4\rho} \sum_{t=0}^{T-1} \mathbb{E} \|R(x_t, y_t)\|^2 + \frac{3}{8\gamma} \left( \frac{1}{L^2} + \frac{1}{L^4} \right) \frac{T \sigma^2}{b},
\]

(73)
By using the above inequality (73), we have

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|G_t\|^2 \leq \frac{32(\Omega_0 - \Omega_T)}{3T\rho \gamma} + \frac{152}{3T\rho^2} \sum_{t=0}^{T-1} \mathbb{E}\|R(x_t, y_t)\|^2 + \frac{4}{\eta \rho \gamma} \left( \frac{1}{L^2} + \frac{1}{L_K^2} \right) \frac{\sigma^2}{b} \\
= \frac{32(\Phi(x_0) + \|y_0 - y^*(x_0)\|^2)}{3T\rho} - \frac{32\mathbb{E}(\Phi(x_T) + \|y_T - y^*(x_T)\|^2)}{3T\rho} \\
+ \frac{152}{3T\rho^2} \sum_{t=0}^{T-1} \mathbb{E}\|R(x_t, y_t)\|^2 + \frac{4}{\eta \rho \gamma} \left( \frac{1}{L^2} + \frac{1}{L_K^2} \right) \frac{\sigma^2}{b} \\
\leq \frac{32(\Phi(x_0) - \Phi^*)}{3T\rho \gamma} + \frac{32\Delta}{3T\rho \gamma} + \frac{152}{3T\rho^2} \sum_{t=0}^{T-1} \mathbb{E}\|R(x_t, y_t)\|^2 + \frac{4}{\eta \rho \gamma} \left( \frac{1}{L^2} + \frac{1}{L_K^2} \right) \frac{\sigma^2}{b},
\]

(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 \( K = \frac{\mu}{\mu} \log \left( \frac{LC_{f_T}}{\mu} \right) \).