Blockwise Stochastic Variance-Reduced Methods with Parallel Speedup for Multi-Block Bilevel Optimization

Quanqi Hu [1] **Zi-Hao Qiu** [2 3] **Zhishuai Guo** [1] **Lijun Zhang** [2] **Tianbao Yang** [1]

Abstract

In this paper, we consider non-convex multi-block bilevel optimization (MBBO) problems, which involve $m \gg 1$ lower level problems and have important applications in machine learning. Designing a stochastic gradient and controlling its variance is more intricate due to the hierarchical sampling of blocks and data and the unique challenge of estimating hyper-gradient. We aim to achieve three nice properties for our algorithm: (a) matching the state-of-the-art complexity of standard BO problems with a single block; (b) achieving parallel speedup by sampling I blocks and sampling B samples for each sampled block per-iteration; (c) avoiding the computation of the inverse of a high-dimensional Hessian matrix estimator. However, it is non-trivial to achieve all of these by observing that existing works only achieve one or two of these properties. To address the involved challenges for achieving (a, b, c), we propose two stochastic algorithms by using advanced blockwise variance-reduction techniques for tracking the Hessian matrices (for low-dimensional problems) or the Hessian-vector products (for high-dimensional problems), and prove an iteration complexity of $\mathcal{O}(\frac{m\epsilon^{-3}\mathbb{I}(I<m)}{I\sqrt{I}} + \frac{m\epsilon^{-3}}{I\sqrt{B}})$ for finding an ϵ-stationary point under appropriate conditions. We also conduct experiments to verify the effectiveness of the proposed algorithms comparing with existing MBBO algorithms.

[1]Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA [2]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China [3]Most work of Z.H. Qiu was done when visiting the OptMAI lab at TAMU. Correspondence to: Tianbao Yang, Quanqi Hu <tianbao-yang, quanqi-hu@tamu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1. Introduction

This paper considers solving the following generalized bilevel optimization problem with multi-block structure:

$$\min_{\mathbf{x}\in\mathbb{R}^{d_x}} F(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} \underbrace{f_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x}))}_{F_i(\mathbf{x})}, \quad \mathbf{y}_i(\mathbf{x}) = \arg\min_{\mathbf{y}_i\in\mathbb{R}^{d_{y,i}}} g_i(\mathbf{x}, \mathbf{y}_i), i = 1, \ldots, m. \tag{1}$$

where f_i, g_i are continuously differentiable functions in expectation forms and $g_i(\mathbf{x}, \mathbf{y}_i)$ is strongly convex with respect to $\mathbf{y}_i$. To be specific, f_i, g_i are defined as $f_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x})) := \mathbb{E}_{\xi\sim\mathcal{P}_i}[f_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x}); \xi)]$ and $g_i(\mathbf{x}, \mathbf{y}_i) = \mathbb{E}_{\zeta\sim\mathcal{Q}_i}[g_i(\mathbf{x}, \mathbf{y}_i; \zeta)]$. The number of blocks m is considered to be greatly larger than 1. We refer to the above problem as **multi-block bilevel optimization** (MBBO). When $m = 1$, the MBBO problem reduces to the standard BO problem. The MBBO problem has found many interesting applications in machine learning and AI, e.g., multi-task compositional AUC maximization (Hu et al., 2022), top-K normalized discounted cumulative gain (NDCG) optimization for learning to rank (Qiu et al., 2022), and meta-learning (Rajeswaran et al., 2019). Recently, Yang (2022) uses MBBO to formulate a family of risk functions for optimizing performance at the top.

The theoretical study of MBBO was initiated by (Guo et al., 2021). In their paper, the authors proposed a randomized stochastic variance-reduced method (RSVRB) for solving MBBO aiming to achieve a state-of-the-art (SOTA) iteration complexity in the order of $O(1/\epsilon^3)$ for finding an ϵ-stationary solution. However, RSVRB and its analysis suffer from several drawbacks: (i) RSVRB requires computing the inverse of the Hessian matrix estimator, which is prohibited for high-dimensional lower-level problems; (ii) the Jacobian estimators maintained for each block could be memory consuming and slow down the algorithm in practice for problems with high-dimensional $\mathbf{x}$; (iii) RSVRB does not achieve a parallel speed-up when using a mini-batch of samples to estimate the gradients, Jacobians and Hessians. While these issues have been tackled for the standard BO problems, e.g., the Hessian matrix can be estimated by the Neumann series and there are works achieving SOTA

complexity without maintainng Jacobian estimator (Yang et al., 2021a; Khanduri et al., 2021), they become trickier for MBBO problems due to extra noise caused by sampling blocks. Although some later studies for particular MBBO problems have achieved parallel speed-up and eschewed computing the inverse of a Hessian estimator (Hu et al., 2022), they do not match the SOTA complexity of $O(1/\epsilon^3)$.

In this paper, we aim to achieve three nice properties for solving MBBO problems: (a) matching the SOTA $O(1/\epsilon^3)$ complexity of standard BO problems with a single block; (b) achieving parallel speedup by sampling multiple blocks and multiple samples for each sampled block per-iteration; (c) avoiding the computation of the inverse of a Hessian matrix estimator for high-dimensional lower level problems. To the best of our knowledge, this is the first work that enjoys all of these three properties for solving MBBO problems. We propose two algorithms named BSVRB$^{\text{v1}}$ and BSVRB$^{\text{v2}}$ for low-dimensional and high-dimensional lower-level problems, respectively. For BSVRB$^{\text{v1}}$, we propose to use an advanced blockwise stochastic variance-reduced estimator namely MSVR (Jiang et al., 2022) to track and estimate the Hessian matrices and the partial gradients of the lower level problems. To further achieve (c) in BSVRB$^{\text{v2}}$, we explore the idea of converting the inverse of the Hessian matrix multiplied by a partial gradient for each block into solving another lower level problem using matrix-vector products. To maintain the same iteration complexity of BSVRB$^{\text{v1}}$, we update the estimators of Hessian-vector products of all blocks without compromising the sample complexity per-iteration. At the end, we manage to prove the same iteration complexity of $\mathcal{O}(\frac{m\epsilon^{-3}\mathbb{I}(I<m)}{I\sqrt{I}} + \frac{m\epsilon^{-3}}{I\sqrt{B}})$ for both algorithms, which reduces to the SOTA complexity $O(\epsilon^{-3}/\sqrt{B})$ of the standard BO with one block.

Our contributions are summarized as following:

- We propose two efficient algorithms by using blockwise stochastic variance reduction for solving MBBO problems with low-dimensional and high-dimensional lower-level problems, respectively.
- We prove the iteration complexity of the two algorithms, which not only matches the SOTA complexity of existing algorithms for solving the standard BO but also achieves parallel speed-up of using multiple blocks and multiple samples of sampled blocks.
- We conduct experiments on both algorithms for low-dimensional and high-dimensional lower problems and demonstrate the effectiveness of the proposed algorithms against existing algorithms of MBBO.

2. Related Work

Stochastic Bilevel Optimization (SBO). SBO algorithms have garnered increasing attention recently. The first non-asymptotic convergence analysis for non-convex SBO with strongly convex lower level problem was given by (Ghadimi & Wang, 2018). The authors proposed a double-loop stochastic algorithm, where the inner loop solves the lower level problem and the outer loop solves the upper level, and established a sample complexity of $\mathcal{O}(\epsilon^{-6})$ for finding an ϵ-stationary point of $F(\mathbf{x})$, i.e., a point $\mathbf{x}$ such that $\|\nabla F(\mathbf{x})\| \leq \epsilon$ in expectation. With a large mini-batch size, (Ji et al., 2020a) improved the sample complexity to $\mathcal{O}(\epsilon^{-4})$. A single-loop two timescale algorithm (TTSA) based on SGD was proposed in (Hong et al., 2020), but suffers from a worse sample complexity of $\mathcal{O}(\epsilon^{-5})$. By utilizing variance-reduction method (STORM) to estimate second-order gradients, i.e., Jacobian $\nabla^2_{xy}g(\mathbf{x},\mathbf{y})$ and Hessian $\nabla^2_{yy}g(\mathbf{x},\mathbf{y})$, (Chen et al., 2021) proposed a single-loop single timescale algorithm (STABLE) that enjoys a sample complexity of $\mathcal{O}(\epsilon^{-4})$ without large mini-batch. Recently, (Khanduri et al., 2021; Yang et al., 2021a; Guo et al., 2021) further improved the sample complexity to $\mathcal{O}(\epsilon^{-3})$ by fully utilizing variance-reduced estimator for gradients of both upper and lower level objectives. (Huang et al., 2021) proposed Bregman distance-based algorithms for solving nonsmooth BO with and without variance reduction.

One of the difficulties for solving SBO problems lies at how to efficiently compute the Hessian inverse in the gradient estimation. To avoid such potentially expensive matrix inverse operation, many existing works have employed the Neumann series approximation with independent mini-batches following (Ghadimi & Wang, 2018). Another method is to transfer the product of the Hessian inverse and a vector to the solution to a quadratic problem (Li et al., 2021; Dagréou et al., 2022; Rajeswaran et al., 2019) and to solve it by using deterministic methods (e.g., conjugate gradient) or stochastic methods that only involve matrix-vector products. However, these methods are tailored to single-block BO problems, and their direct applications to MBBO may suffer from per iteration computation inefficiency. Thus, with the potential efficiency issue in consideration, it is trickier to achieve faster rates for MBBO problems (Hu et al., 2022).

MBBO. Besides (Guo et al., 2021), two recent works have considered MBBO and their applications in ML (Qiu et al., 2022; Hu et al., 2022). In particular, Qiu et al. (2022) formulated top-K NDCG optimization for learning-to-rank as a MBBO problem with a compositional objective function, which can be formulated as our MBBO problem. There are many lower-level problems with each having only an one-dimensional variable for optimization. They proposed a stochastic algorithm (K-SONG) that uses blockwise sampling and moving average estimators for tracking gradients and Hessians, and proved an iteration complexity of $\mathcal{O}(\max\{\frac{m}{IB\epsilon^4}, \frac{1}{\min\{I,B\}\epsilon^4}\})$. Hu et al. (2022) considered a MBBO problem with a min-max objective which includes our considered MBBO problem as a special case. They pro-

Table 1. Comparison of iteration complexity and the three properties of different methods for solving MBBO and FCCO problems. We use MMBO-v2 to refer to the second algorithm proposed in (Hu et al., 2022) for solving a MBBO problem with a min-max objective. The iteration complexity only considers the case $I < m$, where m is the total number of blocks, I is the number of sampled blocks per-iteration and B is the number of sampled data for each sampled block per-iteration. (c) is not applicable to FCCO problems.

Method	Objective	Iteration Complexity	Satisfying (a), (b), (c)
MSVR-v2 (Jiang et al., 2022)	FCCO	$\mathcal{O}(\frac{m\epsilon^{-3}}{I\sqrt{I}} + \frac{m\epsilon^{-3}}{I\sqrt{B}})$	(a), (b)
MMBO-v2 (Hu et al., 2022)	MBBO (min-max)	$\mathcal{O}(\max\left\{\frac{m}{IB}, \frac{1}{\min\{I,B\}}\right\}\epsilon^{-4})$	(b), (c)
K-SONG (Qiu et al., 2022)	MBBO	$\mathcal{O}(\max\left\{\frac{m}{IB}, \frac{1}{\min\{I,B\}}\right\}\epsilon^{-4})$	(b)
RSVRB (Guo et al., 2021)	MBBO	$\mathcal{O}(m\epsilon^{-3})$	(a)
BSVRB^{v1} (this work)	MBBO	$\mathcal{O}(\frac{m\epsilon^{-3}}{I\sqrt{I}} + \frac{m\epsilon^{-3}}{I\sqrt{B}})$	(a), (b)
BSVRB^{v2} (this work)	MBBO	$\mathcal{O}(\frac{m\epsilon^{-3}}{I\sqrt{I}} + \frac{m\epsilon^{-3}}{I\sqrt{B}})$	(a), (b), (c)

posed two algorithms that use moving average estimators for tracking gradients and Hessians or Hessian-vector products for lower-dimensional and high-dimensional lower-level problems, respectively, and established a similar iteration complexity of $\mathcal{O}(\max\{\frac{m}{IB\epsilon^4}, \frac{1}{\min\{I,B\}\epsilon^4}\})$. In their second algorithm, they avoided computing the inverse of the Hessian matrix estimator by using SGD to solve a quadratic problem. It is notable that the iteration complexities of these two works do not match the SOTA result for the standard BO. As discussed before and later, achieving (a), (b) and (c) simultaneously is not just applying variance-reduction techniques such as SPIDER/SARAH/STORM, etc. (Fang et al., 2018; Nguyen et al., 2017; Cutkosky & Orabona, 2019; Zhang et al., 2013), as done in (Guo et al., 2021).

Finally, we would like to point out a related work (Jiang et al., 2022) that considered the finite-sum coupled compositional optimization (FCCO) problem, which is a special case of MBBO with the lower problems being quadratic problems with an identity Hessian matrix. They proposed multi-block-Single-probe Variance Reduced (MSVR) estimator for tracking the inner functional mappings in a blockwise stochastic manner. MSVR helps achieve both the SOTA complexity and the parallel speed-up, which is also leveraged in this work. However, since MBBO is more general than FCCO and involves estimating the hyper-gradient, our algorithmic design and analysis face a new challenge for tracking the Hessian-vector-products, which is not present in their work. We make a comparison between different works for solving MBBO and FCCO problems in Table 1.

3. Preliminaries

Notations. Let $\|\cdot\|$ denote the ℓ_2 norm of a vector or the spectral norm of a matrix. Let $\Pi_\Omega[\cdot]$ denote Euclidean projection onto a convex set Ω for a vector and $\mathbf{S}_\lambda[X]$ denotes a projection onto the set $\{X \in \mathbb{R}^{d\times d} : X \succeq \lambda I\}$. The matrix projection operation $\mathbf{S}_\lambda[X]$ can be implemented by using singular value decomposition and thresholding the singular values. For multi-block structured vectors, we use vector name with subscript i to denote its i-th block. For a twice differentiable function $f : X \times Y \to \mathbb{R}$, let $\nabla_x f(x, y)$ and $\nabla_y f(x, y)$ denote its partial gradients taken with respect to x and y respectively, and let $\nabla^2_{xy} f(x, y)$ and $\nabla^2_{yy} f(x, y)$ denote the Jacobian and the Hessian matrix w.r.t y respectively. We use $f(\cdot; \mathcal{B})$ to represent an unbiased stochastic estimator of $f(\cdot)$ depending on a sampled mini-batch $\mathcal{B}$. An unbiased stochastic estimator using one sample ξ is said to have bounded variance σ^2 if $\mathbb{E}_\xi[\|f(\cdot;\xi) - f(\cdot)\|^2] \leq \sigma^2$. A mapping $f : X \to \mathbb{R}$ is C-Lipschitz continuous if $\|f(x) - f(x')\| \leq C\|x - x'\|\ \forall x, x' \in X$. Function f is L-smooth if its gradient $\nabla f(\cdot)$ is L-Lipschitz continuous. A function $g : X \to \mathbb{R}$ is λ-strongly convex if $\forall x, x' \in X$, $g(x) \geq g(x') + \nabla g(x')^T(x - x') + \frac{\lambda}{2}\|x - x'\|^2$. A point $\mathbf{x}$ is called ϵ-stationary of $F(\cdot)$ if $\|\nabla F(\mathbf{x})\| \leq \epsilon$.

In order to understand the proposed algorithms, we first present following proposition about the (hyper-)gradient of $F(\mathbf{x})$, which follows from the standard result in the literature of bilevel optimization (Ghadimi & Wang, 2018).

Proposition 3.1. *When $g_i(\mathbf{x}, \mathbf{y}_i)$ is strongly convex w.r.t. $\mathbf{y}_i$, we have*

$$\nabla F(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}\big\{\nabla_x f_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x})) - \nabla^2_{xy} g_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x}))[\nabla^2_{yy} g_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x}))]^{-1}\nabla_y f_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x}))\big\}.$$

There are three sources of computational costs involved in the above gradient: (i) the sum over all m blocks; (ii) the costs for computing the partial gradients, Jacobians and Hessian matrices of individual blocks, which usually depend on many samples; and (iii) the inverse of Hessian matrices. The last two have been tackled in the existing literature of BO. The first cost can be alleviated by sampling a mini-batch of blocks. However, due to the compositional structure of the hyper-gradient, designing

a variance-reduced stochastic gradient estimator is complicated due to the existence of multiple blocks (Jiang et al., 2022). In particular, we need to track multiple Hessian matrices $\nabla^2_{yy} g_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x}))$ or Hessian-vector products $[\nabla^2_{yy} g_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x}))]^{-1} \nabla_y f_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x}))$. To this end, we will leverage the MSVR estimator (Jiang et al., 2022), which is described below.

MSVR estimator. Consider multiple functional mappings $(h_1(\mathbf{e}), \ldots, h_m(\mathbf{e}))$, at the t-th iteration we need to estimate their values by an estimator $\mathbf{h}_t = (\mathbf{h}_{1,t}, \ldots, \mathbf{h}_{m,t})$. Given the constraint that only a few blocks of mappings $h_i(\mathbf{e})$ are sampled for assessing their stochastic values, the MSVR update is given by (Jiang et al., 2022):

$$\mathbf{h}_{i,t+1} = \begin{cases} \Big[(1-\alpha)\mathbf{h}_{i,t} + \alpha h_i(\mathbf{e}_t; \mathcal{B}_i^t) \\ \quad + \underbrace{\gamma(h_i(\mathbf{e}_t; \mathcal{B}_i^t) - h_i(\mathbf{e}_{t-1}; \mathcal{B}_i^t))}_{\text{error correction}}\Big], \ i \in \mathcal{I}_t \\ \mathbf{h}_{i,t}, \ \text{o.w.} \end{cases}$$

The update for the sampled $I = |\mathcal{I}_t|$ blocks have a customized error correction term, which is inspired by previous variance reduced estimator STORM (Cutkosky & Orabona, 2019) but has a subtle difference in setting the value of γ. Different from the setting of STORM, i.e., $\gamma = 1 - \alpha$, MSVR sets $\gamma = \frac{m-I}{I(1-\alpha)} + (1-\alpha)$ to account for the randomness and noise induced from block sampling. Due to the need of tracking individual $\mathbf{y}_i$ for each block and the boundedness in our analysis, we extend the above MSVR estimator with two changes: (i) adding a projection onto a convex domain Ω for the update $\mathbf{h}_{i,t+1}$ of sampled blocks whenever boundedness is required, (ii) the input argument $\mathbf{e}_t$ is changed to individual input $\mathbf{e}_{i,t}$.

4. Algorithms

Due to the compositional structure in terms of $\mathbf{y}_i(\mathbf{x})$ and $\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}))$ in the hyper-gradient as shown in Proposition 3.1, we need to maintain and estimate variance-reduced estimators for these variables. Below, we present two algorithms for low-dimensional and high-dimensional lower-level problems, respectively. For low-dimensional lower-level problems, we directly estimate the Hessian matrices and compute their inverse if needed. For high-dimensional lower-level problems, we propose to estimate the Hessian-vector products $[\nabla^2_{yy} g_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x})]^{-1} \nabla_y f_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x}))$.

4.1. For low-dimensional lower-level problems

We first discuss updates for estimators of the (partial) gradients and the Hessian matrices as they are the major costs per-iteration. Then we discuss the updates of $\mathbf{x}$ and $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$, and finally compare with RSVRB.

Updates for Gradient/Hessian Estimators. We need to estimate $\nabla_y g_i(\mathbf{x}_t, \mathbf{y}_{i,t})$ for updating $\mathbf{y}_{i,t}$, to estimate $\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_{i,t})$ for updating $\mathbf{x}_t$. To this end, at each iteration t, we randomly sample a subset of blocks $\mathcal{I}_t \subset [m]$. For each sampled block $i \in \mathcal{I}_t$, we sample a mini-batch $\widetilde{\mathcal{B}}_i^t \sim \mathcal{Q}_i$ for the lower-level problem, and a mini-batch $\mathcal{B}_i^t \sim \mathcal{P}_i$ for the upper-level problem. We update the following MSVR estimators of $\nabla_y g_i(\mathbf{x}_t, \mathbf{y}_{i,t})$ and $\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_{i,t})$ for $i \in \mathcal{I}_t$ and keep their other coordinates unchanged:

$$\begin{aligned} \mathbf{s}_{i,t+1} &= (1-\alpha_t)\mathbf{s}_{i,t} + \alpha_t \nabla_y g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t) \\ &\quad + \gamma_t(\nabla_y g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t) - \nabla_y g_i(\mathbf{x}_{t-1}, \mathbf{y}_{i,t-1}; \widetilde{\mathcal{B}}_i^t)) \\ H_{i,t+1} &= \mathbf{S}_\lambda \Big[(1-\bar{\alpha}_t)H_{i,t} + \bar{\alpha}_t \nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t) \\ &\quad + \bar{\gamma}_t \left(\nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t) - \nabla^2_{yy} g_i(\mathbf{x}_{t-1}, \mathbf{y}_{i,t-1}; \widetilde{\mathcal{B}}_i^t)\right)\Big], \end{aligned} \tag{2}$$

where $\gamma_t = \frac{m-I}{I(1-\alpha_t)} + (1-\alpha_t)$ and $\bar{\gamma}_t = \frac{m-I}{I(1-\bar{\alpha}_t)} + (1-\bar{\alpha}_t)$, λ is the lower bound of the Hessian matrix (cf. Assumption 5.1) and $\mathbf{S}_\lambda$ is a projection operator to ensure the eigenvalue of $H_{i,t+1}$ is lower bounded so that its inverse can be appropriately bounded.

To compute the variance-reduced estimator of $\nabla F(\mathbf{x}_t)$, we compute the stochastic gradient estimations at two iterations:

$$\begin{aligned} G_t &= \frac{1}{I}\sum_{i\in\mathcal{I}_t} \big[\nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t) \\ &\quad - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t)[H_{i,t}]^{-1} \nabla_y f_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t)\big], \\ \widetilde{G}_t &= \frac{1}{I}\sum_{i\in\mathcal{I}_t} \big[\nabla_x f_i(\mathbf{x}_{t-1}, \mathbf{y}_{i,t-1}; \mathcal{B}_i^t) \\ &\quad - \nabla^2_{xy} g_i(\mathbf{x}_{t-1}, \mathbf{y}_{i,t-1}; \widetilde{\mathcal{B}}_i^t)[H_{i,t-1}]^{-1} \nabla_y f_i(\mathbf{x}_{t-1}, \mathbf{y}_{i,t-1}; \mathcal{B}_i^t)\big]. \end{aligned} \tag{3}$$

Then the STORM gradient estimator $\mathbf{z}_{t+1}$ of $\nabla F(\mathbf{x}_t)$ is updated by $\mathbf{z}_{t+1} = (1-\beta_t)(\mathbf{z}_t - \widetilde{G}_t) + G_t$. Note that in the above updates, only stochastic partial gradients, Jacobians, and Hessians based on *two mini-batches of data* $\mathcal{B}_i^t$ and $\widetilde{\mathcal{B}}_i^t$ for the sampled blocks $i \in \mathcal{I}_t$ are computed. This is in sharp contrast with the previous SOTA variance-reduced methods (Khanduri et al., 2021; Yang et al., 2021a) that require *three or four independent* mini-batches due to the use of the Neumann series for estimating the Hessian inverse.

Algorithm 1 Blockwise Stochastic Variance-Reduced Bilevel Method (version 1): BSVRB$^{\text{v1}}$

Initialization: $\mathbf{x}_0 = \mathbf{x}_1, \mathbf{y}_0 = \mathbf{y}_1, \mathbf{s}_1, H_1, \mathbf{z}_1$
for $t = 1, 2, ..., T$ **do**
  Sample a subset of lower problems $\mathcal{I}_t$
  Sample two batches $\mathcal{B}_i^t \sim \mathcal{P}_i, \widetilde{\mathcal{B}}_i^t \sim \mathcal{Q}_i$ for $i \in \mathcal{I}_t$.
  Update $\mathbf{s}_{i,t+1}$ and $H_{i,t+1}$ according to (2) for $i \in \mathcal{I}_t$.
  Compute $G_t, \widetilde{G}_t$ according to (3).
  Update $\mathbf{z}_{t+1} = (1-\beta_t)(\mathbf{z}_t - \widetilde{G}_t) + G_t$
  Update $\mathbf{y}_{t+1} = \mathbf{y}_t - \tau\tau_t \mathbf{s}_t$
  Update $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{z}_{t+1}$
end for
return $(\mathbf{x}_{\tilde{t}}, \mathbf{y}_{\tilde{t}}, \mathbf{s}_{\tilde{t}}, H_{\tilde{t}}, \mathbf{z}_{\tilde{t}})$ for a randomly selected $\tilde{t}$

It is also notable that we use the Hessian estimator $H_{i,t}$ from the previous iteration in computing G_t to decouple its dependence from $\nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t)$ due to using the same mini-batch of data $\widetilde{\mathcal{B}}_i^t$; otherwise we need two independent mini-batches (Wang & Yang, 2022; Hu et al., 2022).

Updates for $\mathbf{x}_{t+1}$ and $\mathbf{y}_{t+1}$. While the update for $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{z}_{t+1}$ is simple, the update of $\mathbf{y}_{t+1}$ is trickier as there are multiple blocks $\mathbf{y}_{i,t+1}, i \in [m]$. A simple approach is to only update $\mathbf{y}_{i,t+1}$ for $i \in \mathcal{I}_t$ as only their gradient estimators $\mathbf{s}_{i,t+1}$ are updated. This is adopted by (Hu et al., 2022). However, since we use MSVR estimators $s_{i,t+1}$ for deriving a fast rate, additional error terms of MSVR estimators will emerge and cause a blow-up on the dependence of m/I. In particular, if we only update $\mathbf{y}_{i,t+1}$ for $i \in \mathcal{I}_t$ and keep other blocks unchanged, we will have an iteration complexity of $T = \mathcal{O}\left(\max\{\frac{m\mathbb{I}(I<m)}{I\sqrt{I}}, \frac{m^{1.5}}{I^{1.5}\sqrt{B}}\}\epsilon^{-3}\right)$, which has an additional scaling $\sqrt{m/I}$ compared that in (Hu et al., 2022) albeit with an improved order on ϵ^{-1}. To avoid this unnecessary blow up, a simple remedy will work by updating all blocks of $\mathbf{y}_{i,t+1}$, i.e., $\mathbf{y}_{t+1} = \mathbf{y}_t - \tau\tau_t \mathbf{s}_t$, where $\mathbf{s}_t = (\mathbf{s}_{1,t}, \ldots, \mathbf{s}_{m,t})$, τ is a parameter and τ_t is scaled stepsize. In fact, such all-block updates can be avoided by using a lazy update strategy. Since the unsampled blocks $\mathbf{y}_{i,t}, \mathbf{y}_{i,t-1}$ are not used in computing gradient/Hessian estimators until i is sampled again, one may accumulate the $\mathbf{y}_{i,t}$ updates and leave it to the future. In particular, at iteration t, we replace the full-block updates of $\mathbf{y}_{i,t+1}$ with the following updates for sampled blocks $i \in \mathcal{I}_t$ at the beginning of the iteration:

$$\begin{aligned} \mathbf{y}_{i,t-1} &= \mathbf{y}_{i,t-1} - (K_{i,t} - 1)\tau\tau_t \mathbf{s}_{i,t}, \\ \mathbf{y}_{i,t} &= \mathbf{y}_{i,t-1} - \tau\tau_t \mathbf{s}_{i,t}, \end{aligned} \tag{4}$$

where $K_{i,t}$ denotes the number of iterations passed since the last time i was sampled. For unsampled blocks $i \notin \mathcal{I}_t$, no update is needed. Finally, we present the detailed steps in Algorithms 1, to which is referred as BSVRB$^{\text{v1}}$.

Comparison with RSVRB. BSVRB$^{\text{v1}}$ is different from RSVRB regarding both algorithm design and theoretical analysis. We summarize the key differences in algorithm design below, and leave the differences of theoretical analysis to section 5. First of all, RSVRB keeps variance-reduced estimators for all partial gradients, Jacobians and Hessians involved in $\nabla F(\mathbf{x})$, while BSVRB$^{\text{v1}}$ only need it for the Hessians and partial gradients of lower-level problems. In fact, estimators for $\nabla_x f_i, \nabla^2_{xy} g_i, \nabla_y f_i$ do not need variance reduction because they all have unbiased estimator and we keep a variance-reduced estimator $\mathbf{z}_t$ for $\nabla F(\mathbf{x})$. Secondly, in the updates of variance-reduced estimators, RSVRB requires scaling updates for non-sampled blocks, while BSVRB$^{\text{v1}}$ requires none. Thirdly, at each iteration RSVRB requires twice independent blocking samplings, one for updates of partial gradient estimators and the other for the update of STORM estimator of $\nabla F(\mathbf{x})$. Lastly, RSVRB involves projection operation for the updates of $\mathbf{y}_{i,t}$ while BSVRB$^{\text{v1}}$ does not. These improvements simplify the algorithm without sacrificing its convergence rate.

4.2. For high-dimensional lower-level problems

One limitation of BSVRB$^{\text{v1}}$ is that computing the inverse of the Hessian estimator $H_{i,t}$ is not suitable for high-dimensional lower-level problems. To address this issue, we propose our second method BSVRB$^{\text{v2}}$. The main idea is to treat $[\nabla^2_{yy} g_i(\mathbf{x}, \mathbf{y}_i)]^{-1} \nabla_y f_i(\mathbf{x}, \mathbf{y}_i)$ as the solution to a quadratic function minimization problem. As a result, $[\nabla^2_{yy} g_i(\mathbf{x}, \mathbf{y}_i)]^{-1} \nabla_y f_i(\mathbf{x}, \mathbf{y}_i)$ can be approximated in a similar way as $\mathbf{y}_i$ in Algorithm 1. This strategy has been studied for solving BO problems in recent works (Hu et al., 2022; Dagréou et al., 2022; Li et al., 2021). However, none of them directly applies to variance reduction methods for MBBO, which incurs additional challenge to be discussed shortly.

Let us define m quadratic problems and their solutions:

$$\begin{aligned} \phi_i(\mathbf{v}, \mathbf{x}, \mathbf{y}_i) &:= \frac{1}{2}\mathbf{v}^T \nabla^2_{yy} g_i(\mathbf{x}, \mathbf{y}_i)\mathbf{v} - \mathbf{v}^T \nabla_y f_i(\mathbf{x}, \mathbf{y}_i) \\ \mathbf{v}_i(\mathbf{x}, \mathbf{y}_i) &:= \arg\min_{\mathbf{v} \in \mathbb{R}^{d_y}} \phi_i(\mathbf{v}, \mathbf{x}, \mathbf{y}_i). \end{aligned} \tag{5}$$

It is not difficult to show that $\mathbf{v}_i(\mathbf{x}, \mathbf{y}_i)$ is equal to $[\nabla^2_{yy} g_i(\mathbf{x}, \mathbf{y}_i)]^{-1} \nabla_y f_i(\mathbf{x}, \mathbf{y}_i)$. Since $\mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})$ can be viewed as solution to another layer of lower-level problem, we conduct similar updates for $\mathbf{v}_{i,t}$ to that for $\mathbf{y}_{i,t}$. Define a stochastic estimator $\nabla_v \phi_i(\mathbf{v}, \mathbf{x}, \mathbf{y}_i; \mathcal{B}_i, \widetilde{\mathcal{B}}_i) := \nabla^2_{yy} g_i(\mathbf{x}, \mathbf{y}_i; \widetilde{\mathcal{B}}_i)\mathbf{v} - \nabla_y f_i(\mathbf{x}, \mathbf{y}_i; \mathcal{B}_i)$. Then an MSVR estimator $\mathbf{u}_{i,t+1}$ for gradient $\nabla_v \phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t})$ is given by

$$\begin{aligned} \mathbf{u}_{i,t+1} = (1 - \bar{\alpha}_t)\mathbf{u}_{i,t} &+ \bar{\alpha}_t \nabla_v \phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t, \widetilde{\mathcal{B}}_i^t) \\ &+ \bar{\gamma}_t \big[\nabla_v \phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t, \widetilde{\mathcal{B}}_i^t) \\ &- \nabla_v \phi_i(\mathbf{v}_{i,t-1}, \mathbf{x}_t, \mathbf{y}_{i,t-1}; \mathcal{B}_i^t, \widetilde{\mathcal{B}}_i^t)\big], i \in \mathcal{I}_t, \end{aligned} \tag{6}$$

and then an update $\mathbf{v}_{i,t+1} = [\mathbf{v}_{i,t} - \bar{\tau}_t \mathbf{u}_{i,t}]$ for the sampled blocks can be conducted. Then we compute STORM gradient estimator of $\nabla F(\mathbf{x}_t)$ using the following two stochastic gradient estimations:

$$\begin{aligned} G_t &= \frac{1}{I}\sum_{i \in \mathcal{I}_t} \big[\nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t) - \nabla^2_{xy} g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t)\mathbf{v}_{i,t}\big], \\ \widetilde{G}_t &= \frac{1}{I}\sum_{i \in \mathcal{I}_t} \big[\nabla_x f_i(\mathbf{x}_{t-1}, \mathbf{y}_{i,t-1}; \mathcal{B}_i^t) \\ &\qquad - \nabla^2_{xy} g_i(\mathbf{x}_{t-1}, \mathbf{y}_{i,t-1}; \widetilde{\mathcal{B}}_i^t)\mathbf{v}_{i,t-1}\big]. \end{aligned} \tag{7}$$

Updates for $\mathbf{x}_{t+1}, \mathbf{y}_{t+1}$ and $\mathbf{v}_{t+1}$. The updates of $\mathbf{x}_{t+1}$ and $\mathbf{y}_{t+1}$ will be conducted similarly as before. However, the update for $\mathbf{v}_{t+1}$ is more subtle. First, the stochastic estimator $\nabla_v \phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i, \widetilde{\mathcal{B}}_i) = \nabla^2_{yy} g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i)\mathbf{v}_{i,t} - \nabla_y f_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i)$ has no bounded variance unless $\mathbf{v}_{i,t}$ is bounded. To this end, we derive an upper bound $V = \frac{C_{fy}}{\lambda}$ so that $\mathbf{v}_i(\mathbf{x}, \mathbf{y}_i) \in \mathcal{V} = \{\mathbf{v}_i : \|\mathbf{v}_i\|_2^2 \leq V^2\}$ under the Assumption 5.1, 5.2 and 5.3 (cf. Appendix B.2). Then the

Algorithm 2 Block-wise Stochastic Variance-Reduced Bilevel Method (version 2): BSVRB$^{\text{v2}}$

Initialization: $\mathbf{x}_0 = \mathbf{x}_1, \mathbf{y}_0 = \mathbf{y}_1, \mathbf{v}_0 = \mathbf{v}_1, \mathbf{s}_1, \mathbf{u}_1, \mathbf{z}_1$
for $t = 1, 2, ..., T$ **do**
  Sample a subset of lower problems $\mathcal{I}_t$
  Sample two batches $\mathcal{B}_i^t \sim \mathcal{P}_i, \widetilde{\mathcal{B}}_i^t \sim \mathcal{Q}_i$ for $i \in \mathcal{I}_t$.
  Update $\mathbf{s}_{i,t+1}$ and $\mathbf{u}_{i,t+1}$ according to (2), (6).
  Compute $G_t, \widetilde{G}_t$ according to (7).
  $\mathbf{z}_{t+1} = (1-\beta_t)(\mathbf{z}_t - \widetilde{G}_t) + G_t$
  Update $\mathbf{y}_{t+1} = \mathbf{y}_t - \tau\tau_t\mathbf{s}_t$
  Update $\mathbf{v}_{t+1} = \Pi_{\mathcal{V}^m}\left[\mathbf{v}_t - \bar{\tau}_t\mathbf{u}_t\right]$
  Update $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t\mathbf{z}_{t+1}$
end for
return $(\mathbf{x}_{\tilde{t}}, \mathbf{y}_{\tilde{t}}, \mathbf{v}_{\tilde{t}}, \mathbf{s}_{\tilde{t}}, \mathbf{u}_{\tilde{t}}, \mathbf{z}_{\tilde{t}})$ for a randomly selected $\tilde{t}$

update of $\mathbf{v}_{i,t+1}$ is modified as $\mathbf{v}_{i,t+1} = \Pi_{\mathcal{V}}\left[\mathbf{v}_{i,t} - \bar{\tau}_t\mathbf{u}_{i,t}\right]$. A similar approach has been used in (Hu et al., 2022). Second, similar to the problem of updating $\mathbf{y}_{i,t+1}$ only for $i \in \mathcal{I}_t$ in BSVRB$^{\text{v1}}$, updating $\mathbf{v}_{i,t+1}$ only for $i \in \mathcal{I}_t$ in BSVRB$^{\text{v2}}$ will lead to worse scaling factor in iteration complexity. To avoid this, we update all blocks of $\mathbf{v}_{i,t+1}$ using its MSVR gradient estimators $\mathbf{u}_{i,t+1}$. With these changes, we present detailed steps in Algorithm 2 for BSVRB$^{\text{v2}}$. Finally, we remark that (i) the sample complexity per-iteration of BSVRB$^{\text{v2}}$ is the same as BSVRB$^{\text{v1}}$, i.e., only two mini-batches $\mathcal{B}_i^t$ and $\widetilde{\mathcal{B}}_i^t$ are required for each sampled block; (ii) similar to (4), the updates of non-sampled blocks $\mathbf{v}_{i,t+1}$ can be delayed until they are sampled again due to that their corresponding gradient estimators do not change and they are not used for computing the gradient estimator $\mathbf{z}_{t+1}$ until their corresponding blocks are sampled again.

5. Convergence Analysis of BSVRB

In this section we provide the convergence analysis of the proposed algorithms and highlight how it is different from the analysis of RSVRB. First of all, we make the following assumptions regarding problem (1).

Assumption 5.1. For any $\mathbf{x}$, $g_i(\mathbf{x}, \cdot)$ is L_g-smooth and λ-strongly convex, i.e., $L_g I \succeq \nabla_{yy}^2 g_i(\mathbf{x}, \mathbf{y}_i) \succeq \lambda I$.

Assumption 5.2. Assume the following conditions hold

- $\nabla_x f_i(\mathbf{x}, \mathbf{y}_i; \xi)$ is L_{fx}-Lipschitz continuous, $\nabla_y f_i(\mathbf{x}, \mathbf{y}_i; \xi)$ is L_{fy}-Lipschitz continuous, $\nabla_y g_i(\mathbf{x}, \mathbf{y}_i; \zeta)$ is L_{gy}-Lipschitz continuous, $\nabla_{xy}^2 g_i(\mathbf{x}, \mathbf{y}_i; \zeta)$ is L_{gxy}-Lipschitz continuous, $\nabla_{yy}^2 g_i(\mathbf{x}, \mathbf{y}_i; \zeta)$ is L_{gyy}-Lipschitz continuous, all with respect to $(\mathbf{x}, \mathbf{y}_i)$.
- $\|\nabla_x f_i(\mathbf{x}, \mathbf{y}_i)\|^2 \le C_{fx}^2$, $\|\nabla_y f_i(\mathbf{x}, \mathbf{y}_i)\|^2 \le C_{fy}^2$.
- All stochastic estimators $\nabla_x f_i(\mathbf{x}, \mathbf{y}_i; \xi)$, $\nabla_y f_i(\mathbf{x}, \mathbf{y}_i; \xi)$, $\nabla_y g_i(\mathbf{x}, \mathbf{y}_i; \zeta)$, $\nabla_{xy}^2 g_i(\mathbf{x}, \mathbf{y}_i; \zeta)$, $\nabla_{yy}^2 g_i(\mathbf{x}, \mathbf{y}_i; \zeta)$ have bounded variance σ^2.

Assumption 5.3. $\|\nabla_{yy}^2 g_i(\mathbf{x}, \mathbf{y}_i, \zeta)\|^2 \preceq \tilde{C}_{gyy}^2 I$.

Assumption 5.1 is made in many existing works for SBO (Chen et al., 2021; Ghadimi & Wang, 2018; Hong et al., 2020; Ji et al., 2020a). Assumption 5.2 ii) iii) are also standard in the literature (Ji et al., 2020b; Ghadimi & Wang, 2018; Hong et al., 2020). To employ variance reduction technique, Lipchitz continuity of stochastic gradients, i.e., Assumption 5.2 i), is required (Yang et al., 2021b; Cutkosky & Orabona, 2019). Note that the assumption $\nabla_y g_i(\mathbf{x}, \mathbf{y}_i; \zeta)$ is L_{gy}-Lipschitz continuous implies that $\|\nabla_{xy}^2 g_i(\mathbf{x}, \mathbf{y}_i)\|^2 \le C_{gxy}^2$ and $\|\nabla_{yy}^2 g_i(\mathbf{x}, \mathbf{y}_i)\|^2 \le C_{gyy}^2$ with $C_{gxy} = C_{gyy} = L_{gy}$. Assumption 5.3 is only required by BSVRB$^{\text{v2}}$ to ensure the Lipschitz continuity of $\nabla_v \phi_i(\mathbf{v}_i, \mathbf{x}, \mathbf{y}_i, \xi, \zeta)$. It is notable that an even stronger assumption $\tilde{C}_{gyy}^2 I \succeq \nabla_{yy}^2 g(\mathbf{x}, \mathbf{y}; \zeta) \succeq \lambda I$ is made in (Ghadimi & Wang, 2018; Hong et al., 2020; Yang et al., 2021b) due to the use of the Neumann series (cf. the proof of Lemma 3.2 in (Ghadimi & Wang, 2018), Assumption 1&2 in (Yang et al., 2021b)).

Comparing to the assumptions made for RSVRB, BSVRB no longer requires the boundedness of $\mathbf{y}_i(\mathbf{x})$ and the expectation of the stochastic gradients norms $\nabla_x f_i(\mathbf{x}, \mathbf{y}_i; \xi)$, $\nabla_y f_i(\mathbf{x}, \mathbf{y}_i; \xi)$, $\nabla_y g_i(\mathbf{x}, \mathbf{y}_i; \zeta)$, $\nabla_{xy}^2 g_i(\mathbf{x}, \mathbf{y}_i; \zeta)$, $\nabla_{yy}^2 g_i(\mathbf{x}, \mathbf{y}_i; \zeta)$. The latter boundedness requirements in RSVRB come from the error bound analysis of the randomized coordinate STORM estimators, which uses $(0, \ldots, m\nabla_y g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \zeta_t), \ldots, 0)$ as an unbiased estimator of $\nabla_y g(\mathbf{x}_t, \mathbf{y}_t)$. It is also the reason for not having parallel speed-up of using multiple samples for each block (Wang & Yang, 2022).

Next, we present our main result about the convergence of BSVRB$^{\text{v1}}$ and BSVRB$^{\text{v2}}$ unified in the following theorem.

Theorem 5.4. *Under Assumptions 5.1, 5.2 and 5.3 (for BSVRB$^{\text{v2}}$), with* $|\mathcal{B}_i^t| = |\widetilde{\mathcal{B}}_i^t| = B$, $\tau \le \frac{2}{3L_g}$, $\alpha_t = \mathcal{O}(B\epsilon^2)$, $\bar{\alpha}_t, \beta_t = \mathcal{O}((\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{-1}\epsilon^2)$, $\tau_t, \bar{\tau}_t = \mathcal{O}(\sqrt{\frac{I}{m}}(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{-1/2}\epsilon)$, $\eta_t = \mathcal{O}(\frac{I}{m}(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{-1/2}\epsilon)$, *and by using a large mini-batch size of* $\mathcal{O}(1/\epsilon)$ *at the initial iteration, both Algorithm 1 and 2 give* $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \|\nabla F(\mathbf{x}_t)\|^2\right] \le \epsilon^2$ *with an iteration complexity* $T = \mathcal{O}(\frac{m\epsilon^{-3}\mathbb{I}(I<m)}{I\sqrt{I}} + \frac{m\epsilon^{-3}}{I\sqrt{B}})$.

Remark: The achieved iteration complexity (i) matches the SOTA results for standard BO problems with only one block when $I = m = 1$ (Khanduri et al., 2021; Yang et al., 2021a; Guo et al., 2021); (ii) has a parallel speed-up by using multiple blocks I and multiple samples in the mini-batches $\mathcal{B}_i^t$ and $\widetilde{\mathcal{B}}_i^t$. It is worth mentioning that the above theorem requires using a large batch size at the initial iteration. This is mainly because that we use fixed small parameters for $\alpha_t, \bar{\alpha}_t, \beta_t, \eta_t$ for simplicity of exposition and for proving faster convergence under a Polyak-Łojasiewicz (PL) condition in next section, for which we do not require the large batch size at the initial iteration. We can also

Algorithm 3 RE-BSVRB

Initialize the set of variables $\Theta_0 = \{\mathbf{x}_0 = \mathbf{x}_1, \mathbf{y}_0 = \mathbf{y}_1, \mathbf{v}_0 = \mathbf{v}_1, \mathbf{s}_1, H_1, \mathbf{u}_1, \mathbf{z}_1\}$
Define parameters $\{\Xi_k\}_{k=1}^K$ according to Theorem 6.1
for $k = 1, \ldots, K$ **do**
  Θ_k = BSVRB(Θ_{k-1}, Ξ_k)
end for
Return: $\mathbf{x}_K$

use decreasing parameters as in previous works to remove the large batch size at the initial iteration for finding an ϵ-stationary point.

6. Faster Convergence for Gradient-Dominant Functions

In this section, we use a standard restarting trick to improve the convergence of BSVRB under the gradient dominant condition (aka. Polyak-Łojasiewicz (PL) condition), i.e.,

$$\mu(F(\mathbf{x}) - \min_{\mathbf{x}'} F(\mathbf{x}')) \leq \|\nabla F(\mathbf{x})\|^2.$$

The procedure is described in Algorithm 3 and its convergence is stated below.

Theorem 6.1. *Suppose Assumptions 5.1, 5.2 and 5.3 (for RE-BSVRB^{v2}) hold and the PL condition holds. Set appropriate initial parameters $\Xi_1 = (\beta_1, \alpha_1, \bar{\alpha}_1, \bar{\tau}_1, \tau_1, \eta_1, T_1)$. Define proper constant $\epsilon_1 = \mathcal{O}(\frac{1}{\mu}(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}))$ and $\epsilon_k = \epsilon_1/2^{k-1}$. For $k \geq 2$, set parameter Ξ_k such that $\beta_k, \alpha_k, \bar{\alpha}_k = \mathcal{O}(\mu\epsilon_k(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{-1})$, $\bar{\tau}_k, \tau_k, \eta_k = \mathcal{O}(\frac{I\sqrt{\mu\epsilon_k}}{m}(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{-1/2})$ and $T_k = \mathcal{O}(\max\left\{\frac{1}{\mu\eta_k}, \frac{1}{\tau_k}, \frac{1}{\alpha_k}\right\})$ (for RE-BSVRB^{v1}), or $T_k = \mathcal{O}(\max\left\{\frac{1}{\mu\eta_k}, \frac{1}{\beta_k}, \frac{1}{\tau_k}, \frac{1}{\bar{\tau}_k}\right\})$ (for RE-BSVRB^{v2}), then after $K = \mathcal{O}(\log(\epsilon_1/\epsilon))$ stages, the output of RE-BSVRB satisfies $\mathbb{E}[F(\mathbf{x}_K) - \min_{\mathbf{x}} F(\mathbf{x})] \leq \epsilon$.*

Remark: For both RE-BSVRB^{v1} and RE-BSVRB^{v2}, it follows from Theorem 6.1 that the total sample complexity is $\sum_{k=1}^K T_k = O(\frac{m}{I\mu^{3/2}\epsilon^{1/2}}(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{1/2} + \frac{m}{I\mu\epsilon})$. If $\mu \geq \epsilon$, the dependence on μ and ϵ matches the optimal rate of $O(\frac{1}{\mu\epsilon})$ to minimize a strongly convex problem, which is a stronger condition than PL condition. In the existing works, RE-RSVRB (Guo et al., 2021) has a similar result. Other than that, to get $\mathbb{E}[\|\mathbf{x} - \mathbf{x}^*\|^2] \leq \epsilon$, STABLE (Chen et al., 2021) takes a complexity of $O(\frac{m}{\mu^4\epsilon})$, TTSA (Hong et al., 2020) takes $\widetilde{O}(\frac{m}{\mu^3\epsilon^{3/2}})$, and BSA (Ghadimi & Wang, 2018) takes $O(\frac{m}{\mu^5\epsilon^2})$, all under the strong convexity.

7. Experiments

7.1. Hyper-parameter Optimization

In this subsection, we consider solving MBBO with high-dimensional lower-level problems. In particular, we consider a hyper-parameter optimization problem for classification with imbalanced data and noisy labels. For handling data imbalance, we assign the j-th training data ζ_j a weight $\sigma(p_j) \in (0, 1)$, where $\sigma(\cdot)$ is a sigmoid function, and p_j is a decision variable which will be learned by a bilevel optimization. In order to tackle noisy labels in the training data, we consider using a robust loss function given by $\mathcal{L}_\tau(\mathbf{w}; x, y) := \log\left(1 + \exp\left(-y(\mathbf{w}^T x + w_0)/\tau\right)\right)$, where $x \in \mathbb{R}^d$ is input feature, $y \in \{1, -1\}$ denotes its label, $\tau > 0$ is a temperature parameter. This loss function has been shown to be robust to label noise by tuning the τ in (Zhu et al., 2023). In our experiment, instead of tuning τ, we consider multiple values of them and learn a model that is robust for different temperature values. In particular, for each temperature value τ_i, we learn a model $\mathbf{w}_i(\mathbf{p})$ following the weighted empirical risk minimization using the i-th loss $\mathcal{L}_i(\mathbf{w}; x, y) = \mathcal{L}_{\tau_i}(\mathbf{w}; x, y)$.

As a result, a MBBO problem is imposed as:

$$\min_{\mathbf{p}\in\mathbb{R}^n} F(\mathbf{p}) := \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\xi_j\sim\mathcal{D}_{val}}\left[\mathcal{L}_i(\mathbf{w}_i(\mathbf{p}); \xi_j)\right]$$

$$\mathbf{w}_i(\mathbf{p}) = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \mathbb{E}_{\zeta_j\sim\mathcal{D}_{tr}}\left[\sigma(p_j)\mathcal{L}_i(\mathbf{w}; \zeta_j)\right] + \frac{\lambda}{2}\|\mathbf{w}\|^2, \quad \text{for all } i = 1, \ldots, m,$$

where $\mathcal{D}_{tr}$ contains n training data points and $\mathcal{D}_{val}$ is a validation set, $\mathbb{E}_{\zeta_j\sim\mathcal{D}_{tr}}$ denotes an average of data from the given set.

In the first experiment on hyper-parameter optimization, we aim to compare BSVRB and RSVRB, compare BSVRB^{v2} with BSVRB^{v1} for high-dimensional lower-level problems, and to verify the parallel speedup of both BSVRB^{v1} and BSVRB^{v2} with respect to block sampling size I and the batch size B of samples.

Data. We use two binary classification datasets, UCI Adult benchmark dataset *a8a* (Platt, 1999) and web page classification dataset *w8a* (Dua & Graff, 2017). *a8a* and *w8a* have a feature dimensionality of 123 and 300 respectively, and contain 22696 and 49749 training samples. For both *a8a* and *w8a*, we follow 80%/20% training/validation split.

Setup. We set the number of loss functions to be 100 using randomly generated $\{\tau_i\}_{i=1}^m$ in the range of $[1, 11)$. For methods comparison, we sample 10 blocks at each iteration and set the sample batch size to be 32. The regularization parameter λ is chosen from $\{0.00001, 0.0001, 0.001, 0.01\}$. For all methods, we tune the upper-level problem learning rate η_t from $\{0.001, 0.01, 0.1\}$ and the lower-level problem learning rates $\tau_t, \bar{\tau}_t$ from $\{0.01, 0.1, 0.5, 1, 5, 10\}$. Parameters $\alpha_t = \bar{\alpha}_t$ and $\gamma_t = \bar{\gamma}_t$ in MSVR estimator are tuned from $\{0.5, 0.9, 1, 10, 100\}$ and $\{0.001, 0.01, 0.1, 1, 10, 100\}$ respectively. In RSVRB, the STORM parameter β is chosen from $\{0.1, 0.5, 0.9, 1\}$. We runs 4 trails for each setting and plot the average curves. This experiment is performed

on a computing node with Intel Xeon 8352Y (Ice Lake) processor and 64GB memory.

Results. We plot the curves of validation loss for BSVRB and RSVRB in Figure 2. For both datasets, all methods perform similarly in terms of epochs. However, in terms of running time, both BSVRB$^{\text{v1}}$ and BSVRB$^{\text{v2}}$ have better performance than RSVRB. For dataset $w8a$, that has a higher lower-level problem dimension 300, BSVRB$^{\text{v2}}$ shows its greater advantage against BSVRB$^{\text{v1}}$ and RSVRB. This is consistent with our theory that BSVRB$^{\text{v2}}$ is more suitable for high-dimensional lower-level problems. Note that one of the major issues that slows down RSVRB is maintaining the Jacobian estimators, e.g. a matrix of size $100 * 300 * 39799$ for *w8a*, which is avoided by BSVRB$^{\text{v1}}$ and BSVRB$^{\text{v2}}$. In Figure 3, we compare the loss curves of BSVRB$^{\text{v1}}$ and BSVRB$^{\text{v2}}$ with different values of I (# of sampled blocks) and B (# of sampled data per sampled block) on a8a. It shows that the convergence speed increases as I and B increases, which verifies the parallel speedup of our algorithms.

Classification with Imbalanced Data with Noisy Labels. To further demonstrate the benefit of our multi-block bilevel optimization formulation for classification with imbalanced data and noisy labels, we artificially construct an imbalanced a8a data with varied label noise. We remove 70% of the positive samples in training data to produce an imbalanced version. Moreover, we add noise by flipping the labels of the remaining training data with a certain probability, i.e., the noise level from 0 to 0.4.

We compare three methods, i) logistic regression with the standard logistic loss as the baseline, ii) BSVRB$^{\text{v1}}$ for solving the bilevel formulation with only one lower level problem ($m = 1$ and using the standard logistic loss), and iii) our method for solving multi-block bilevel formulation with $m = 100$ blocks corresponding to 100 settings of the scaling factor τ_i in the logistic loss. Since these methods optimize different objectives, we use the accuracy on a separate testing data as the performance measure for comparison. For our method, we have multiple models learned with different loss functions. We select the best model on the validation data and measure its accuracy on testing data. In terms of parameter tuning, for logistic regression we tune the step size in the range $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$. For BSVRB$^{\text{v1}}$ we follow the same parameter tuning strategy described in the previous experiment. For each setting, we repeat the experiment 3 times by changing the random seeds. We present the results in Figure 1 as a bar graph. We defer the numeric results to Appendix E. As we can see from the results, our multi-block bilevel optimization formulation of hyper-parameter optimization has superior performance, especially with high noise level.

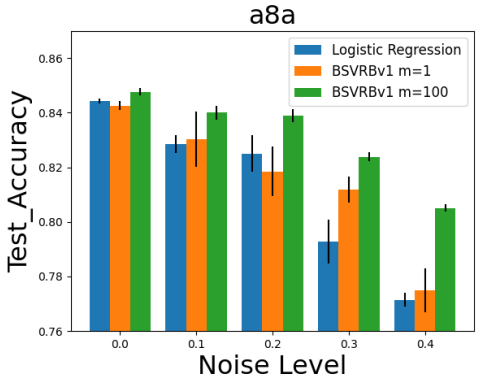

Figure 1. Comparison of testing accuracy of models learned by regular logistic regression, BSVRB$^{\text{v1}}$ with $m = 1$ lower-level problem, and BSVRB$^{\text{v1}}$ with $m = 100$ lower-level problems on a corrupted dataset *a8a* with various noise levels.

7.2. Top-K NDCG Optimization

In this experiment, we consider the top-K NDCG optimization proposed in (Qiu et al., 2022), and reformulate it into an equivalent MBBO problem. Let $q \in \mathcal{Q}$ denote a query, $\mathcal{S}_q = \{(\mathbf{x}_i^q, y_i^q)\}_{i=1}^{N_q}$ denote a set of items and their relevance scores w.r.t to q, $\mathcal{S}$ denote the set of relevant query-item pairs, and $h_q(\cdot,\cdot)$ denote the predictive model for query q. Then the MBBO formulation of this problem is:

$$\min \frac{1}{|\mathcal{S}|} \sum_{(q,\mathbf{x}_i^q)\in\mathcal{S}} \psi(h_q(\mathbf{x}_i^q;\mathbf{w}) - \lambda_q(\mathbf{w})) f_{q,i}(g(\mathbf{w};\mathbf{x}_i^q)),$$

$$\text{where} \quad \lambda_q(\mathbf{w}) = \arg\min_{\lambda} \frac{K+\varepsilon}{N_q}\lambda + \frac{\tau_2}{2}\lambda^2$$

$$+ \frac{1}{N_q} \sum_{\mathbf{x}_i\in\mathcal{S}_q} \tau_1 \ln(1 + \exp((h_q(\mathbf{x}_i;\mathbf{w}) - \lambda)/\tau_1)),$$

$$g(\mathbf{w};\mathbf{x}_i^q) = \arg\min_{g} \frac{1}{2}(g - g(\mathbf{w};\mathbf{x}_i^q,\mathcal{S}_q))^2, \forall (q,\mathbf{x}_i^q) \in \mathcal{S},$$

where $f_{q,i}(g) = \frac{1}{Z_q^K}\frac{1-2^{y_i^q}}{\log_2(N_q g+1)}$, $g(\mathbf{w};\mathbf{x}_i^q,\mathcal{S}_q) = \frac{1}{|\mathcal{S}_q|}\sum_{\mathbf{x}'\in\mathcal{S}_q} \ell(h_q(\mathbf{x}';\mathbf{w}) - h_q(\mathbf{x}_i^q;\mathbf{w}))$, $\ell(\cdot) = (\cdot + c)_+^2$ with a margin parameter c, $\psi(\cdot)$ is sigmoid function, and Z_q^K is the top-K DCG score of the perfect ranking. We refer the readers to (Qiu et al., 2022) for more detailed description of the problem which is omitted due to limite of space.

We follow the exactly same experimental settings as (Qiu et al., 2022). Specifically, we adopt two movie recommendation datasets, i.e., MovieLens20M (Harper & Konstan, 2015) and Netflix Prize dataset (Bennett et al., 2007), employ the same evaluation protocols, model architectures, and hyper-parameters for training. For our method, we tune $\alpha, \bar{\alpha}$ and $\gamma, \bar{\gamma}$ in the ranges of $\{0.7, 0.8, 0.9\}$ and $\{0.001, 0.005, 0.01, 0.1, 1, 10\}$, respectively. Details of data and experimental setups are presented in Appendix A.1.

Since all lower-level problems have one-dimensional vari-

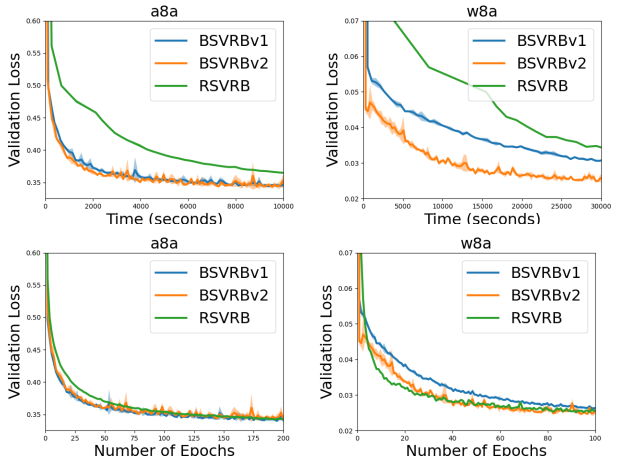

Figure 2. Comparison of convergence curves of different methods in terms of validation loss on datasets a8a and w8a.

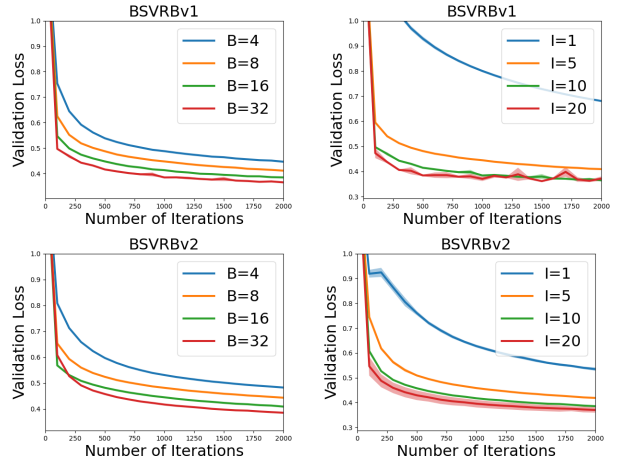

Figure 3. Comparison of convergence curves of BSVRB algorithms with different values of I and and B on a8a.

able for optimization, we only compare RE-BSVRB$^{\text{v1}}$ with K-SONG and other methods reported in (Qiu et al., 2022). We plot the convergence curves for optimizing top-10 NDCG on two datasets in Figure 4, and note that our RE-BSVRB$^{\text{v1}}$ converges faster than other methods. We also provide NDCG@10 scores on the test data for all methods in Table 2 and more results in Table 3 in Appendix A.1. We observe that our method is better for top-K NDCG optimization than other methods. Specifically, our method improves upon K-SONG by 5.24% and 6.49% on NDCG@10 for Movielens data and Netflix data, respectively.

The code for reproducing the experimental results in this section is available at https://github.com/Optimization-AI/ICML2023_BSVRB.

8. Conclusions

In this paper, we have proposed novel stochastic algorithms for solving MBBO problems. We have established the state-of-the-art complexity with a parallel speed-up. Our experiments on both algorithms for low-dimensional and high-dimensional lower problems demonstrate the effectiveness of our algorithms against existing algorithms of MBBO.

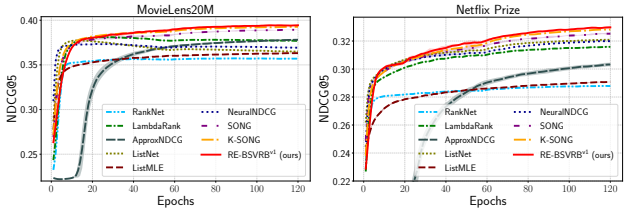

Figure 4. Comparison of convergence of different methods in terms of validation NDCG@5 on two movie recommendation datasets.

Table 2. The test NDCG@10 scores on two movie recommendation datasets averaged over 5 trials. More results for other metrics are in Table 3 in Appendix A.1

METHOD	MOVIELENS	NETFLIX
RANKNET	0.0538±0.0011	0.0362±0.0002
LISTNET	0.0660±0.0003	0.0532±0.0002
LISTMLE	0.0588±0.0001	0.0376±0.0003
LAMBDARANK	0.0697±0.0001	0.0531±0.0002
APPROXNDCG	0.0735±0.0005	0.0434±0.0005
NEURALNDCG	0.0692±0.0003	0.0554±0.0002
SONG	0.0748±0.0002	0.0571±0.0002
K-SONG	0.0747±0.0002	0.0573±0.0003
RE-BSVRB$^{\text{v1}}$	**0.0749**±0.0003	**0.0585**±0.0004

Acknowledgements

We thank anonymous reviewers for constructive comments. Q. Hu and T. Yang were partially supported by NSF Career Award 2246753, NSF Grant 2246757 and NSF Grant 2246756.

References

Bennett, J., Lanning, S., et al. The netflix prize. In *Proceedings of KDD Cup and Workshop*, volume 2007, pp. 35, 2007.

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

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pp. 15236–15245, 2019.

Dagréou, M., Ablin, P., Vaiter, S., and Moreau, T. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms, 2022. URL https://arxiv.org/abs/2201.13409.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. SPIDER: near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 687–697, 2018.

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

Guo, Z., Hu, Q., Zhang, L., and Yang, T. Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization, 2021. URL https://arxiv.org/abs/2105.02266.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T.-S. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 173–182, 2017.

He, X., He, Z., Song, J., Liu, Z., Jiang, Y.-G., and Chua, T.-S. Nais: Neural attentive item similarity model for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 30(12):2354–2366, 2018.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 639–648, 2020.

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

Hu, Q., Zhong, Y., and Yang, T. Multi-block min-max bilevel optimization with applications in multi-task deep auc maximization, 2022. URL https://arxiv.org/abs/2206.00260.

Huang, F., Li, J., Gao, S., and Huang, H. Enhanced bilevel optimization via bregman distance, 2021. URL https://arxiv.org/abs/2107.12301.

Ji, K., Yang, J., and Liang, Y. Provably faster algorithms for bilevel optimization and applications to meta-learning. *arXiv preprint arXiv:2010.07962*, 2020a.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design, 2020b. URL https://arxiv.org/abs/2010.07962.

Jiang, W., Li, G., Wang, Y., Zhang, L., and Yang, T. Multi-block-single-probe variance reduced estimator for coupled compositional optimization, 2022. URL https://arxiv.org/abs/2207.08540.

Khanduri, P., Zeng, S., Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A near-optimal algorithm for stochastic bilevel optimization via double-momentum, 2021. URL https://arxiv.org/abs/2102.07367.

Li, J., Gu, B., and Huang, H. A fully single loop algorithm for bilevel optimization without hessian inverse, 2021. URL https://arxiv.org/abs/2112.04660.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takávc, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2613–2621. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/nguyen17b.html.

Platt, J. C. Fast training of support vector machines using sequential minimal optimization. 1999.

Qiu, Z.-H., Hu, Q., Zhong, Y., Zhang, L., and Yang, T. Large-scale stochastic optimization of ndcg surrogates for deep learning with provable convergence, 2022. URL https://arxiv.org/abs/2202.12183.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. *CoRR*, abs/1909.04630, 2019. URL http://arxiv.org/abs/1909.04630.

Wang, B. and Yang, T. Finite-sum compositional stochastic optimization: Theory and applications. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 23292–23317, 2022.

Wang, C., Zhang, M., Ma, W., Liu, Y., and Ma, S. Modeling item-specific temporal dynamics of repeat consumption for recommender systems. In *Proceedings of the 28th International Conference on World Wide Web*, pp. 1977–1987, 2019a.

Wang, C., Zhang, M., Ma, W., Liu, Y., and Ma, S. Make it a chorus: knowledge-and time-aware item modeling for sequential recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 109–118, 2020.

Wang, X., He, X., Wang, M., Feng, F., and Chua, T.-S. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 165–174, 2019b.

Yang, J., Ji, K., and Liang, Y. Provably faster algorithms for bilevel optimization. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13670–13682. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/file/71cc107d2e0408e60a3d3c44f47507bd-Paper.pdf.

Yang, J., Ji, K., and Liang, Y. Provably faster algorithms for bilevel optimization, 2021b. URL https://arxiv.org/abs/2106.04692.

Yang, T. Algorithmic foundation of deep x-risk optimization. *arXiv*, 2022. doi: 10.48550/ARXIV.2206.00439. URL https://arxiv.org/abs/2206.00439.

Zhang, L., Mahdavi, M., and Jin, R. Linear convergence with condition number independent access of full gradients. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/37f0e884fbad9667e38940169d0a3c95-Paper.pdf.

Zhu, D., Ying, Y., and Yang, T. Label distributionally robust losses for multi-class classification: Consistency, robustness and adaptivity. In *Proceedings of International Conference on Machine Learning*, volume abs/2112.14869, 2023. URL https://arxiv.org/abs/2112.14869.

Table 3. The full results test NDCG on two movie recommendation datasets. We report the average NDCG@k ($k \in [10, 20, 50]$) and standard deviation over 5 runs with different random seeds.

METHOD	MOVIELENS20M			NETFLIX PRIZE DATASET		
	NDCG@10	NDCG@20	NDCG@50	NDCG@10	NDCG@20	NDCG@50
RANKNET	0.0538±0.0011	0.0744±0.0013	0.1086±0.0013	0.0362±0.0002	0.0489±0.0003	0.0730±0.0003
LISTNET	0.0660±0.0003	0.0875±0.0004	0.1227±0.0003	0.0532±0.0002	0.0700±0.0002	0.0992±0.0002
LISTMLE	0.0588±0.0001	0.0799±0.0001	0.1137±0.0001	0.0376±0.0003	0.0508±0.0004	0.0753±0.0001
LAMBDARANK	0.0697±0.0001	0.0913±0.0002	0.1259±0.0001	0.0531±0.0002	0.0693±0.0002	0.0976±0.0003
APPROXNDCG	0.0735±0.0005	0.0938±0.0003	0.1284±0.0002	0.0434±0.0005	0.0592±0.0009	0.0873±0.0012
NEURALNDCG	0.0692±0.0003	0.0901±0.0003	0.1232±0.0007	0.0554±0.0002	0.0718±0.0003	0.1003±0.0002
SONG	0.0748±0.0002	0.0969±0.0002	0.1326±0.0001	0.0571±0.0002	0.0749±0.0002	0.1050±0.0003
K-SONG	0.0747±0.0002	**0.0973**±0.0003	**0.1340**±0.0001	0.0573±0.0003	0.0743±0.0003	0.1042±0.0001
RE-BSVRB$^{\mathbf{v1}}$	**0.0749**±0.0003	0.0963±0.0002	0.1314±0.0003	**0.0585**±0.0004	**0.0760**±0.0003	**0.1061**±0.0002

A. Top-K NDCG Optimization

A.1. Details of data and experimental setups

Data. We use two large-scale movie recommendation datasets: MovieLens20M (Harper & Konstan, 2015) and Netflix Prize dataset (Bennett et al., 2007). Both datasets contain large numbers of users and movies, which are represented with integer IDs. All users have rated several movies, with ratings range from 1 to 5. To create training/validation/test sets, we use the most recent rated item of each user for testing, the second recent item for validation, and the remaining items for training, which is widely-used in the literature (He et al., 2018; Wang et al., 2020). When evaluating models, we need to collect irrelevant (unrated) items and rank them with the relevant (rated) item to compute NDCG metrics. During training, inspired by Wang et al. (2019a), we randomly sample 1000 unrated items to save time. When testing, however, we adopt the all ranking protocol (Wang et al., 2019b; He et al., 2020) — all unrated items are used for evaluation.

Setup. We choose NeuMF (He et al., 2017) as the backbone network, which is commonly used in recommendation tasks. For all methods, models are first pre-trained by our initial warm-up method for 100 epochs with the learning rate 0.001 and a batch size of 256. Then the last layer is randomly re-initialized and the network is fine-tuned by different methods. At the fine-tuning stage, the initial learning rate and weight decay are set to 0.0004 and 1e-7, respectively. We train the models for 120 epochs with the learning rate multiplied by 0.25 at 60 epochs. The hyper-parameters of all methods are individually tuned for fair comparison, e.g., we tune α_* and γ_* for our method in ranges of $\{0.7, 0.8, 0.9\}$ and $\{0.001, 0.005, 0.01\}$, respectively.

B. Convergence Analysis of BSVRB

In this section, we present the convergence analysis of BSVRB. We let $\mathbf{y}_t = (\mathbf{y}_{1,t}, \ldots, \mathbf{y}_{m,t})$, $\mathbf{v}_t = (\mathbf{v}_{1,t}, \ldots, \mathbf{v}_{m,t})$, $\mathbf{u}_t = (\mathbf{u}_{1,t}, \ldots, \mathbf{u}_{m,t})$, $\mathbf{s}_t = (\mathbf{s}_{1,t}, \ldots, \mathbf{s}_{m,t})$, $H_t = (H_{1,t}, \ldots, H_{m,t})$, $\mathbf{y}(\mathbf{x}) = (\mathbf{y}_1(\mathbf{x}), \ldots, \mathbf{y}_m(\mathbf{x}))$, $\mathbf{v}(\mathbf{x}, \mathbf{y}) = (\mathbf{v}_1(\mathbf{x}, \mathbf{y}_1), \ldots, \mathbf{v}_m(\mathbf{x}, \mathbf{y}_m))$.

For simplicity, we define the following notations.

$$\delta_{z,t} := \|\mathbf{z}_{t+1} - \Delta_t\|^2, \quad \delta_{y,t} := \sum_{i=1}^m \|\mathbf{y}_{i,t} - \mathbf{y}_i(\mathbf{x}_t)\|^2, \quad \delta_{v,t} := \sum_{i=1}^m \|\mathbf{v}_{i,t} - \mathbf{v}(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2,$$

$$\delta_{s,t} := \sum_{i=1}^m \|\mathbf{s}_{i,t} - \nabla_y g_i(\mathbf{x}_{t-1}, \mathbf{y}_{i,t-1})\|^2, \quad \tilde{\delta}_{s,t} := \sum_{i=1}^m \|\mathbf{s}_{i,t} - \nabla_y g_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2,$$

$$\delta_{u,t} := \sum_{i=1}^m \|\mathbf{u}_{i,t} - \nabla_v \phi_i(\mathbf{v}_{i,t-1}, \mathbf{x}_{t-1}, \mathbf{y}_{i,t-1})\|^2, \quad \tilde{\delta}_{u,t} := \sum_{i=1}^m \|\mathbf{u}_{i,t} - \nabla_v \phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t})\|^2,$$

$$\delta_{H,t} := \sum_{i=1}^m \|H_{i,t} - \nabla_{yy}^2 g_i(\mathbf{x}_{t-1}, \mathbf{y}_{i,t-1})\|^2, \quad \tilde{\delta}_{H,t} := \sum_{i=1}^m \|H_{i,t} - \nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2.$$

Note that under Assumption 5.1, 5.2, 5.3, we have

$$\begin{aligned}
&\tilde{\delta}_{H,t} \leq 2\delta_{H,t} + 2L_{gyy}^2(m\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2) \\
&\tilde{\delta}_{s,t} \leq 2\delta_{s,t} + 2L_{gy}^2(m\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2) \\
&\tilde{\delta}_{u,t} \leq 2\delta_{u,t} + 2L_{\phi v}^2(\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 + m\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2)
\end{aligned} \tag{8}$$

We initialize $\mathbf{x}_0 = \mathbf{x}_1$, $\mathbf{y}_0 = \mathbf{y}_1$ and $\mathbf{v}_0 = \mathbf{v}_1$, so that we have

$$\begin{aligned}
&\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \leq \sum_{t=1}^{T}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2, \quad \sum_{t=1}^{T}\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2 \leq \sum_{t=1}^{T}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2, \\
&\sum_{t=1}^{T}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 \leq \sum_{t=1}^{T}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2.
\end{aligned} \tag{9}$$

We first present some standard results from non-convex optimization and bilevel optimization literature.

Lemma B.1 (Lemma 2.2 in (Ghadimi & Wang, 2018))**.** *$F(\mathbf{x})$ is L_F-smooth and $\mathbf{y}_i(\mathbf{x})$ is L_y-Lipschitz continuous for all $i = 1, \ldots, m$, where L_y and L_F are appropriate constants.*

Lemma B.2. *Let $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{z}_{t+1}$. Under Assumptions 5.1, 5.2, with $\eta_t L_F \leq 1/2$, we have*

$$F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \frac{\eta_t}{2}\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2 - \frac{\eta_t}{2}\|\nabla F(\mathbf{x}_t)\|^2 - \frac{\eta_t}{4}\|\mathbf{z}_{t+1}\|^2.$$

Lemma B.3 (Lemma 6 in (Guo et al., 2021))**.** *Let $\mathbf{y}_{t+1} = \mathbf{y}_t - \tau_t \tau \mathbf{s}_t$ with $\tau \leq 2/(3L_g)$, we have*

$$\begin{aligned}
&\|\mathbf{y}_{t+1} - \mathbf{y}(\mathbf{x}_{t+1})\|^2 \leq (1 - \frac{\tau_t \tau \lambda}{4})\|\mathbf{y}_t - \mathbf{y}(\mathbf{x}_t)\|^2 + \frac{8\tau_t \tau}{\lambda}\|\nabla_y g(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{s}_t\|^2 \\
&+ \frac{8L_y^2\gamma^2}{\tau_t \tau \lambda}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - \frac{2\tau}{\tau_t}(1 + \frac{\tau_t \tau \lambda}{4})(\frac{1}{2\tau} - \frac{3L_g}{4})\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2.
\end{aligned}$$

Lemma B.4. *Let Ω be a convex set. Suppose mapping $h_i(\mathbf{e}_i; \xi)$ is L-Lipschitz, $h_i(\mathbf{e}) = \mathbb{E}_\xi[h_i(\mathbf{e}_i; \xi)]$, $h_i(\mathbf{e}) \in \Omega$ and $\mathbb{E}_\xi[\|h_i(\mathbf{e}_i) - h_i(\mathbf{e}_i; \xi)\|^2] \leq \sigma^2$ for all $i = 1, \ldots, m$. Consider the MSVR update:*

$$\mathbf{h}_{i,t+1} = \begin{cases} \Pi_\Omega\Big[(1-\alpha)\mathbf{h}_{i,t} + \alpha h_i(\mathbf{e}_{i,t}; \mathcal{B}_i^t) \\ \quad +\gamma(h_i(\mathbf{e}_{i,t}; \mathcal{B}_i^t) - h_i(\mathbf{e}_{i,t-1}; \mathcal{B}_i^t))\Big], \ i \in \mathcal{I}_t \\ \mathbf{h}_{i,t}, \ \text{o.w.} \end{cases} \tag{10}$$

Denote $\delta_{h,t} := \sum_{i=1}^m \|\mathbf{h}_{i,t} - h_i(\mathbf{e}_{i,t-1})\|^2$. By setting $\gamma = \frac{m-I}{I(1-\alpha)} + (1-\alpha)$, for $\alpha \leq \frac{1}{2}$, with batch sizes $I = |\mathcal{I}_t|$ and $B = |\mathcal{B}_i^t|$, we have

$$\begin{aligned}
\mathbb{E}\left[\delta_{h,t+1}\right] \leq (1 - \frac{I\alpha}{m})\mathbb{E}\left[\delta_{h,t}\right] + \frac{2I\alpha^2\sigma^2}{B} \\
+ \frac{8m^2L^2}{I}\mathbb{E}\left[\sum_{i=1}^m \|\mathbf{e}_{i,t-1} - \mathbf{e}_{i,t}\|^2\right]
\end{aligned} \tag{11}$$

With $\Omega = \mathbb{R}^d$ the above lemma is Lemma 1 in (Jiang et al., 2022). We refer the detailed proof to Appendix D.2

B.1. Convergence Analysis of BSVRB$^{\text{v1}}$

We first present a formal statement of Theorem 5.4 for BSVRB$^{\text{v1}}$.

Theorem B.5. *Under Assumptions 5.1 and 5.2, with $\tau \leq \frac{2}{3L_g}$, $\bar{\gamma}_t = \frac{m-I}{I(1-\bar{\alpha}_t)} + (1-\bar{\alpha}_t)$, $\gamma_t = \frac{m-I}{I(1-\alpha_t)} + (1-\alpha_t)$, $\alpha_t \leq \min\left\{\frac{1}{2}, \frac{B\epsilon^2}{12C_{10}}\right\}$, $\beta_t \leq \frac{\min\{I,B\}\epsilon^2}{12C_{10}}$, $\bar{\alpha}_t \leq \min\left\{\frac{1}{2}, \frac{\epsilon^2}{12C_{10}}(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{-1}\right\}$, $\tau_t \leq \sqrt{\frac{C_8}{12C_{10}}}\frac{\sqrt{I}\epsilon}{\sqrt{m}}(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{-1/2}$, $\eta_t \leq \min\left\{\frac{1}{2L_F}, \sqrt{C_{11}}\frac{I\epsilon}{m}(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{-1/2}\right\}$, where C_{10}, C_{11} are constants specified in the proof, and by using a large mini-batch size of $\mathcal{O}(1/\epsilon)$ at the initial iteration for computing $\mathbf{z}_1, \mathbf{s}_1, H_1$ and computing an accurate solution $\mathbf{y}_1$ such that $\delta_{y,1} \leq \mathcal{O}(1)$, Algorithm 1 gives $\mathbb{E}\left[\frac{1}{T}\sum_{i=1}^m \|\nabla F(\mathbf{x}_t)\|^2\right] \leq \epsilon^2$ with sample complexity $T = \mathcal{O}\left(\frac{m\epsilon^{-3}\mathbb{I}(I<m)}{I\sqrt{I}} + \frac{m\epsilon^{-3}}{I\sqrt{B}}\right)$.*

Define

$$\Delta_t = \frac{1}{m}\sum_{i=1}^{m} \nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t})\mathbb{E}_t[[H_{i,t}]^{-1}]\nabla_y f_i(\mathbf{x}_t, \mathbf{y}_{i,t}) \tag{12}$$

so that $\mathbb{E}_t[G_t] = \Delta_t$, $\mathbb{E}_t[\widetilde{G}_t] = \Delta_{t-1}$.

Lemma B.6. *With constants C_1, C_2 defined in the proof, we have*

$$\|\Delta_t - \nabla F(\mathbf{x}_t)\|^2 \le C_1\delta_{y,t} + C_2\tilde{\delta}_{H,t}. \tag{13}$$

Lemma B.7. *Consider the updates in Algorithm 1, we have*

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{z}_{t+1} - \Delta_t\|^2] \le (1-\beta_t)\|\mathbf{z}_t - \Delta_{t-1}\|^2 + 2C_3\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \frac{2C_3}{m}\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2 \\ + \frac{2C_4}{m}\|H_t - H_{t-1}\|^2 + 2\beta_t^2 C_5\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right)\end{aligned} \tag{14}$$

Lemma B.8. *With MSVR updates for H_{t+1}, if $\bar{\alpha}_t \le \frac{1}{2}$, we have*

$$\mathbb{E}\left[\|H_{t+1} - H_t\|^2\right] \le \frac{2I\bar{\alpha}_t^2\sigma^2}{B} + \frac{8I\bar{\alpha}_t^2}{m}\mathbb{E}[\delta_{H,t}] + \frac{9m^2L_{gyy}^2}{I}\mathbb{E}[\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2] + \frac{9mL_{gyy}^2}{I}E[\|\mathbf{y}_{t-1} - \mathbf{y}_t\|^2] \tag{15}$$

By applying Lemma B.4 to $\mathbf{s}_t, H_t$, we have

$$\mathbb{E}\left[\delta_{H,t+1}\right] \le (1 - \frac{I\bar{\alpha}_t}{m})\mathbb{E}\left[\delta_{H,t}\right] + \frac{2I\bar{\alpha}_t^2\sigma^2}{B} + \frac{8m^2L_{gyy}^2}{I}\mathbb{E}\left[\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2 + \frac{1}{m}\|\mathbf{y}_{t-1} - \mathbf{y}_t\|^2\right]$$

and

$$\mathbb{E}\left[\delta_{s,t+1}\right] \le (1 - \frac{I\alpha_t}{m})\mathbb{E}\left[\delta_{s,t}\right] + \frac{2I\alpha_t^2\sigma^2}{B} + \frac{8m^2L_{gy}^2}{I}\mathbb{E}\left[\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2 + \frac{1}{m}\|\mathbf{y}_{t-1} - \mathbf{y}_t\|^2\right]$$

Take summation over $t = 1, \ldots, T$, then we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}\delta_{H,t}\right] \le \mathbb{E}\left[\frac{m}{I\bar{\alpha}_t}\delta_{H,1} + \frac{2m\bar{\alpha}_t\sigma^2T}{B} + \frac{8m^3L_{gyy}^2}{\bar{\alpha}_tI^2}\sum_{t=1}^{T}\left[\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2 + \frac{1}{m}\|\mathbf{y}_{t-1} - \mathbf{y}_t\|^2\right]\right] \tag{16}$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T}\delta_{s,t}\right] \le \mathbb{E}\left[\frac{m}{I\alpha_t}\delta_{s,1} + \frac{2m\alpha_t\sigma^2T}{B} + \frac{8m^3L_{gy}^2}{\alpha_tI^2}\sum_{t=1}^{T}\left[\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2 + \frac{1}{m}\|\mathbf{y}_{t-1} - \mathbf{y}_t\|^2\right]\right] \tag{17}$$

B.1.1. Proof of Theorem B.5

Proof. By Lemma B.2, we have

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) \le \frac{\eta_t}{2}\|\mathbf{z}_{t+1} - \nabla F(\mathbf{x}_t)\|^2 - \frac{\eta_t}{2}\|\nabla F(\mathbf{x}_t)\|^2 - \frac{\eta_t}{4}\|\mathbf{z}_{t+1}\|^2. \tag{18}$$

The first term on the right hand side can be divided into two terms.

$$\|\mathbf{z}_{t+1} - \nabla F(\mathbf{x}_t)\|^2 \le 2\|\mathbf{z}_{t+1} - \Delta_t\|^2 + 2\|\Delta_t - \nabla F(\mathbf{x}_t)\|^2 \tag{19}$$

where we have recursion for the first term on the right hand side in Lemma B.7 and the second term is bounded by Lemma B.6. Combining inequalities 18,19 and Lemma B.6 gives

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) \le \eta_t\delta_{z,t} + \frac{\eta_tC_1}{m}\delta_{y,t} + \frac{\eta_tC_2}{m}\tilde{\delta}_{H,t} - \frac{\eta_t}{2}\|\nabla F(\mathbf{x}_t)\|^2 - \frac{\eta_t}{4}\|\mathbf{z}_{t+1}\|^2. \tag{20}$$

Taking summation over $t = 1, \ldots, T$ yields

$$\sum_{t=1}^{T}\|\nabla F(\mathbf{x}_t)\|^2 \le \frac{2}{\eta_t}(F(\mathbf{x}_1) - F(\mathbf{x}^*)) + 2\sum_{t=1}^{T}\delta_{z,t} + \frac{2C_1}{m}\sum_{t=1}^{T}\delta_{y,t} + \frac{2C_2}{m}\sum_{t=1}^{T}\tilde{\delta}_{H,t} - \frac{1}{2}\sum_{t=1}^{T}\|\mathbf{z}_{t+1}\|^2 \tag{21}$$

We enlarge the values of constants C_1 so that

$$\begin{aligned}&\sum_{t=1}^{T}\|\nabla F(\mathbf{x}_t)\|^2 + \sum_{t=1}^{T}\delta_{z,t} + \frac{1}{m}\sum_{t=1}^{T}\delta_{y,t} \\ &\le \frac{2}{\eta_t}(F(\mathbf{x}_1) - F(\mathbf{x}^*)) + 3\sum_{t=1}^{T}\delta_{z,t} + \frac{C_1}{m}\sum_{t=1}^{T}\delta_{y,t} + \frac{C_2}{m}\sum_{t=1}^{T}\tilde{\delta}_{H,t} - \frac{1}{2}\sum_{t=1}^{T}\|\mathbf{z}_{t+1}\|^2\end{aligned} \tag{22}$$

It follows from Lemma B.7 that

$$\mathbb{E}\left[\sum_{t=1}^{T}\delta_{z,t}\right] \leq \mathbb{E}\Bigg[\frac{\delta_{z,1}}{\beta_t} + \frac{2C_3}{\beta_t}\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \frac{2C_3}{m\beta_t}\sum_{t=1}^{T}\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2 + \frac{2C_4}{m\beta_t}\sum_{t=1}^{T}\|H_t - H_{t+1}\|^2 + 2\beta_t C_5 T\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right)\Bigg] \tag{23}$$

Combing with Lemma B.8, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\delta_{z,t}\right] \leq \mathbb{E}\Bigg[\frac{\delta_{z,1}}{\beta_t} + \left(\frac{2C_3}{\beta_t} + \frac{18C_4 m L_{gyy}^2}{I\beta_t}\right)\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \left(\frac{2C_3}{m\beta_t} + \frac{18L_{gyy}^2}{I\beta_t}\right)\sum_{t=1}^{T}\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2 + 2\beta_t C_5 T\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right) + \frac{4C_4 I\bar{\alpha}_{t+1}^2\sigma^2 T}{m\beta_t B} + \frac{8C_4 I\bar{\alpha}_{t+1}^2}{m^2\beta_t}\sum_{t=1}^{T}\delta_{H,t}\Bigg] \tag{24}$$

Following from Lemma B.3, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\delta_{y,t}\right] \leq \mathbb{E}\Bigg[\frac{4}{\tau_t\tau\lambda}\delta_{y,1} + \frac{32}{\lambda^2}\sum_{t=1}^{T}\tilde{\delta}_{s,t} - \frac{8}{\tau_t^2\lambda I}\left(\frac{1}{2\tau} - \frac{3L_g}{4}\right)\sum_{t=1}^{T}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 + \frac{32mL_y^2}{\tau_t^2\tau^2\lambda^2}\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2\Bigg] \tag{25}$$

Following from inequalities (25), (24) and (22), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla F(\mathbf{x}_t)\|^2 + \sum_{t=1}^{T}\delta_{z,t} + \frac{1}{m}\sum_{t=1}^{T}\delta_{y,t}\right] \\
&\leq \mathbb{E}\Bigg[\frac{2}{\eta_t}(F(\mathbf{x}_1) - F(\mathbf{x}^*)) + \frac{3\delta_{z,1}}{\beta_t} + 6\beta_t C_5 T\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right) + \frac{12C_4 I\bar{\alpha}_{t+1}^2\sigma^2 T}{m\beta_t B} \\
&\quad + \frac{C_1}{m}\sum_{t=1}^{T}\delta_{y,t} + \frac{C_2}{m}\sum_{t=1}^{T}\tilde{\delta}_{H,t} + \frac{24C_4 I\bar{\alpha}_{t+1}^2}{m^2\beta_t}\sum_{t=1}^{T}\delta_{H,t} \\
&\quad + \left(\frac{6C_3}{\beta_t} + \frac{54C_4 m L_{gyy}^2}{I\beta_t}\right)\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 - \frac{1}{2}\sum_{t=1}^{T}\|\mathbf{z}_{t+1}\|^2 + \left(\frac{6C_3}{m\beta_t} + \frac{54L_{gyy}^2}{I\beta_t}\right)\sum_{t=1}^{T}\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2 \\
&\leq \mathbb{E}\Bigg[\frac{2}{\eta_t}(F(\mathbf{x}_1) - F(\mathbf{x}^*)) + \frac{3\delta_{z,1}}{\beta_t} + \frac{4C_1}{\tau_t\tau\lambda m}\delta_{y,1} + 6\beta_t C_5 T\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right) + \frac{12C_4 I\bar{\alpha}_{t+1}^2\sigma^2 T}{m\beta_t B} \\
&\quad + \frac{32C_1}{m\lambda^2}\sum_{t=1}^{T}\tilde{\delta}_{s,t} + \frac{C_2}{m}\sum_{t=1}^{T}\tilde{\delta}_{H,t} + \frac{24C_4 I\bar{\alpha}_{t+1}^2}{m^2\beta_t}\sum_{t=1}^{T}\delta_{H,t} \\
&\quad + \left(\frac{6C_3}{\beta_t} + \frac{54C_4 m L_{gyy}^2}{I\beta_t} + \frac{32L_y^2 C_1}{\tau_t^2\tau^2\lambda^2}\right)\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 - \frac{1}{2}\sum_{t=1}^{T}\|\mathbf{z}_{t+1}\|^2 \\
&\quad + \left(\frac{6C_3}{m\beta_t} + \frac{54L_{gyy}^2}{I\beta_t} - \frac{8C_1}{\tau_t^2\lambda m}\left(\frac{1}{2\tau} - \frac{3L_g}{4}\right)\right)\sum_{t=1}^{T}\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2
\end{aligned}$$

Then we replace $\tilde{\delta}_{s,t}, \tilde{\delta}_{H,t}$ by $\delta_{s,t}, \delta_{H,t}$ following inequality (8) and (9).

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \|\nabla F(\mathbf{x}_t)\|^2 + \sum_{t=1}^T \delta_{z,t} + \frac{1}{m}\sum_{t=1}^T \delta_{y,t} + \frac{1}{m}\sum_{t=1}^T \delta_{H,t} + \frac{1}{m}\sum_{t=1}^T \delta_{s,t}\right] \\
&\leq \mathbb{E}\Bigg[\frac{2}{\eta_t}(F(\mathbf{x}_1) - F(\mathbf{x}^*)) + \frac{3\delta_{z,1}}{\beta_t} + \frac{4C_1}{\tau_t\tau\lambda m}\delta_{y,1} + 6\beta_t C_5 T\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right) + \frac{12C_4 I\bar{\alpha}_{t+1}^2\sigma^2 T}{m\beta_t B} + \frac{C_6}{m}\sum_{t=1}^T \delta_{s,t} \\
&\quad + \frac{C_7}{m}\sum_{t=1}^T \delta_{H,t} + \left(\frac{6C_3}{\beta_t} + \frac{54C_4 m L_{gyy}^2}{I\beta_t} + \frac{32L_y^2 C_1}{\tau_t^2\tau^2\lambda^2} + \frac{64C_1 L_{gy}^2}{\lambda^2} + 2C_2 L_{gyy}^2\right)\sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 - \frac{1}{2}\sum_{t=1}^T \|\mathbf{z}_{t+1}\|^2 \\
&\quad + \left(\frac{6C_3}{m\beta_t} + \frac{54L_{gyy}^2}{I\beta_t} - \frac{8C_1}{\tau_t^2\lambda m}(\frac{1}{2\tau} - \frac{3L_g}{4}) + \frac{64C_1 L_{gy}^2}{m\lambda^2} + \frac{2C_2 L_{gyy}^2}{m}\right)\sum_{t=1}^T \|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2
\end{aligned}
$$

where $\frac{C_6}{m} \geq \frac{64C_1}{m\lambda^2} + \frac{1}{m}$ and $\frac{C_7}{m} \geq \frac{2C_2+1}{m} + \frac{24C_4 I\bar{\alpha}_{t+1}^2}{m^2\beta_t}$

Then we plug in $\sum_{t=1}^T \delta_{s,t}$ and $\sum_{t=1}^T \delta_{H,t}$ to the right hand side following (16) and (17),

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \|\nabla F(\mathbf{x}_t)\|^2 + \sum_{t=1}^T \delta_{z,t} + \frac{1}{m}\sum_{t=1}^T \delta_{y,t} + \frac{1}{m}\sum_{t=1}^T \delta_{H,t} + \frac{1}{m}\sum_{t=1}^T \delta_{s,t}\right] \\
&\leq \mathbb{E}\Bigg[\frac{2}{\eta_t}(F(\mathbf{x}_1) - F(\mathbf{x}^*)) + \frac{3\delta_{z,1}}{\beta_t} + \frac{4C_1}{\tau_t\tau\lambda m}\delta_{y,1} + \frac{C_6}{I\alpha_t}\delta_{s,1} + \frac{C_7}{I\bar{\alpha}_t}\delta_{H,1} + 6\beta_t C_5 T\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right) \\
&\quad + \frac{12C_4 I\bar{\alpha}_{t+1}^2\sigma^2 T}{m\beta_t B} + \frac{2C_6\alpha_t\sigma^2 T}{B} + \frac{2C_7\bar{\alpha}_t\sigma^2 T}{B} + \left(\frac{6C_3}{\beta_t} + \frac{54C_4 m L_{gyy}^2}{I\beta_t} + \frac{32L_y^2 C_1}{\tau_t^2\tau^2\lambda^2} + \frac{64C_1 L_{gy}^2}{\lambda^2} + 2C_2 L_{gyy}^2\right. \\
&\quad \left. + \frac{8m^2 C_6 L_{gy}^2}{\alpha_t I^2} + \frac{8m^2 C_7 L_{gyy}^2}{\bar{\alpha}_t I^2}\right)\sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 - \frac{1}{2}\sum_{t=1}^T \|\mathbf{z}_{t+1}\|^2 + \left(\frac{6C_3}{m\beta_t} + \frac{54L_{gyy}^2}{I\beta_t}\right. \\
&\quad \left. - \frac{8C_1}{\tau_t^2\lambda m}(\frac{1}{2\tau} - \frac{3L_g}{4}) + \frac{64C_1 L_{gy}^2}{m\lambda^2} + \frac{2C_2 L_{gyy}^2}{m} + \frac{8mC_6 L_{gy}^2}{\alpha_t I^2} + \frac{8mC_7 L_{gyy}^2}{\bar{\alpha}_t I^2}\right)\sum_{t=1}^T \|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2
\end{aligned}
$$

To ensure the coefficient of $\sum_{t=1}^T \|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2$ is non positive, we need

$$
\begin{aligned}
\tau_t^2 &= C_8 \min\left\{\frac{I\beta_t}{m}, \frac{\bar{\alpha}_t I^2}{m^2}, \frac{\alpha_t I^2}{m^2}\right\} \\
&\leq \frac{48C_1}{\lambda m}(\frac{1}{2\tau} - \frac{3L_g}{4}) \min\left\{\frac{m\beta_t}{6C_3}, \frac{I\beta_t}{54L_{gyy}^2}, \frac{m\lambda^2}{64C_1 L_{gy}^2}, \frac{m}{2C_2 L_{gyy}^2}, \frac{\alpha_t I^2}{8mC_6 L_{gy}^2}, \frac{\bar{\alpha}_t I^2}{8mC_7 L_{gyy}^2}\right\}
\end{aligned}
$$

where $C_8 := \frac{48C_1}{\lambda I}(\frac{1}{2\tau} - \frac{3L_g}{4}) \min\left\{\frac{1}{6C_3}, \frac{1}{54L_{gyy}^2}, \frac{\lambda^2}{64C_1 L_{gy}^2}, \frac{1}{2C_2 L_{gyy}^2}, \frac{1}{8C_6 L_{gy}^2}, \frac{1}{8C_7 L_{gyy}^2}\right\}$.

To ensure the coefficient of $\sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2$ is non positive, we need

$$
\begin{aligned}
\eta_t^2 &= C_9 \min\left\{\frac{I\beta_t}{m}, \frac{\bar{\alpha}_t I^2}{m^2}, \tau_t^2, \frac{\alpha_t I^2}{m^2}\right\} \\
&\leq \min\{\frac{\beta_t}{84C_3}, \frac{I\beta_t}{756C_4 m L_{gyy}^2}, \frac{\tau_t^2\tau^2\lambda^2}{448L_y^2 C_1}, \frac{\lambda^2}{896C_1 L_{gy}^2}, \frac{1}{28C_2 L_{gyy}^2}, \frac{\alpha_t I^2}{112m^2 C_6 L_{gy}^2}, \frac{\bar{\alpha}_t I^2}{112m^2 C_7 L_{gyy}^2}\}
\end{aligned}
$$

where $C_9 := \min\{\frac{1}{84C_3}, \frac{1}{756C_4 L_{gyy}^2}, \frac{\tau^2\lambda^2}{448L_y^2 C_1}, \frac{\lambda^2}{896C_1 L_{gy}^2}, \frac{1}{28C_2 L_{gyy}^2}, \frac{1}{112C_6 L_{gy}^2}, \frac{1}{112C_7 L_{gyy}^2}\}$

Then it follows

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla F(\mathbf{x}_t)\|^2+\sum_{t=1}^{T}\delta_{z,t}+\frac{1}{m}\sum_{t=1}^{T}\delta_{y,t}+\frac{1}{m}\sum_{t=1}^{T}\delta_{H,t}+\frac{1}{m}\sum_{t=1}^{T}\delta_{s,t}\right]\\
&\le \mathbb{E}\Bigg[\frac{2}{\eta_t}(F(\mathbf{x}_1)-F(\mathbf{x}^*))+\frac{3\delta_{z,1}}{\beta_t}+\frac{4C_1}{\tau_t\tau\lambda m}\delta_{y,1}+\frac{C_6}{I\alpha_t}\delta_{s,1}+\frac{C_7}{I\bar{\alpha}_t}\delta_{H,1}\\
&\quad+6\beta_t C_5 T\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)+\frac{12C_4 I\bar{\alpha}_{t+1}^2\sigma^2 T}{m\beta_t B}+\frac{2C_6\alpha_t\sigma^2 T}{B}+\frac{2C_7\bar{\alpha}_t\sigma^2 T}{B}\Bigg]\\
&\le \mathbb{E}\left[\frac{2\Delta}{\eta_t}+C_{10}\left(\frac{1}{\beta_t}\delta_{z,1}+\frac{1}{I\alpha_t}\delta_{s,1}+\frac{1}{\tau_t I}\delta_{y,1}+\frac{1}{I\bar{\alpha}_t}\delta_{H,1}+\frac{\bar{\alpha}_t T}{B}+\frac{\alpha_t T}{B}+\beta_t T\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)+\frac{I\bar{\alpha}_{t+1}^2 T}{m\beta_t B}\right)\right]
\end{aligned}
$$

where $C_{10}:=\max\left\{3,\frac{4C_1}{\tau\lambda},C_6,C_7,6C_5,12C_4\sigma^2,2C_6\sigma^2,2C_7\sigma^2\right\}$.

Set $\alpha_t\le\frac{B\epsilon^2}{12C_{10}}$, $\bar{\alpha}_t,\beta_t\le\frac{\epsilon^2}{12C_{10}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)}$, so that

$$
C_{10}\left(\frac{\bar{\alpha}_t}{B}+\frac{\alpha_t}{B}+\beta_t\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)+\frac{I\bar{\alpha}_{t+1}^2}{m\beta_t B}\right)=(\frac{\epsilon^2}{12B}+\frac{I\epsilon^2}{12mB})\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1}+\frac{\epsilon^2}{12}+\frac{\epsilon^2}{12}\le\frac{\epsilon^2}{3}
$$

As a result, we have

$$
\begin{aligned}
\tau_t^2&\le C_8\min\left\{\frac{I\beta_t}{m},\frac{\bar{\alpha}_t I^2}{m^2},\frac{\alpha_t I^2}{m^2}\right\}\\
&=\frac{C_8}{12C_{10}}\min\left\{\frac{I^2\epsilon^2}{m^2}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1},\frac{I^2B\epsilon^2}{m^2}\right\}\\
&=\frac{C_8}{12C_{10}}\frac{I^2\epsilon^2}{m^2}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1}
\end{aligned}
$$

and

$$
\begin{aligned}
\eta_t^2&\le C_9\min\left\{\frac{I\beta_t}{m},\frac{\bar{\alpha}_t I^2}{m^2},\tau_t^2,\frac{\alpha_t I^2}{m^2}\right\}\\
&=C_{11}\min\left\{\frac{I\epsilon^2}{m}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1},\frac{I^2\epsilon^2}{m^2}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1},\frac{I^2B\epsilon^2}{m^2}\right\}\\
&=C_{11}\frac{I^2\epsilon^2}{m^2}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1}
\end{aligned}
$$

where $C_{11}=\frac{C_8}{12C_{10}}$.

Thus, with $T=c_T\epsilon^{-3}:=6\Delta\sqrt{C_{11}}\frac{m}{I}\left(\frac{\mathbb{I}(I<m)}{\sqrt{I}}+\frac{1}{\sqrt{B}}\right)\ge 6\Delta\sqrt{C_{11}}\frac{m}{I}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{1/2}\epsilon^{-3}$, we have

$$
\frac{2\Delta}{\eta_t T}=2\Delta\sqrt{C_{11}}\frac{m}{I}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{1/2}\epsilon^{-1}\frac{1}{T}\le\frac{\epsilon^2}{3}
$$

Note that $\frac{C_{10}}{T}\mathbb{E}\left[\frac{1}{\beta_t}\delta_{z,1}+\frac{1}{\tau_t I}\delta_{y,1}+\frac{1}{\alpha_t I}\delta_{s,1}+\frac{1}{\bar{\alpha}_t I}\delta_{H,1}\right]\le\frac{\epsilon^2}{3}$ can be achieved by processing all lower problems at the beginning and finding good initial solutions $\delta_{z,1},\delta_{s,1},\delta_{H,1}=\mathcal{O}(\epsilon)$, with complexity $\mathcal{O}(\epsilon^{-1})$, and $\delta_{y,1}=\mathcal{O}(1)$ with complexity $\mathcal{O}(1)$. Denote the iteration number for initialization as $T_0=\mathcal{O}(\epsilon^{-1})$. Then the total iteration complexity is $\mathcal{O}\left(\frac{m\epsilon^{-3}\mathbb{I}(I<m)}{I\sqrt{I}}+\frac{m\epsilon^{-3}}{I\sqrt{B}}\right)$.

□

B.2. Convergence Analysis of BSVRB$^{\text{v2}}$

We first present the formal statement of Theorem 5.4 for BSVRB$^{\text{v2}}$.

Theorem B.9. *Under Assumptions 5.1, 5.2 and 5.3, with* $\tau \leq 2/(3L_g)$, $\bar{\gamma}_t = \frac{m-I}{I(1-\bar{\alpha}_t)} + (1-\bar{\alpha}_t)$, $\gamma_t = \frac{m-I}{I(1-\alpha_t)} + (1-\alpha_t)$, $\tau_t \leq \sqrt{\frac{C_8' \min\{1, C_{\bar{\tau}}\}}{18C_9'}} \frac{I}{m}(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{-1/2}\epsilon$, $\beta_t \leq \frac{1}{18C_9'}(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{-1}\epsilon^2$, $\bar{\alpha}_t \leq \min\left\{\frac{1}{2}, \frac{1}{18C_9'}(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{-1}\epsilon^2\right\}$, $\alpha_t \leq \min\left\{\frac{1}{2}, \frac{B}{18C_9'}\epsilon^2\right\}$, $\bar{\tau}_t \leq \min\left\{\frac{\lambda}{8L_{\phi v}^2}, \frac{\lambda}{2}, \frac{1}{\lambda}, \frac{1}{2\sqrt{C_6'}}, \frac{\sqrt{C_{\bar{\tau}}}I}{m}\sqrt{\bar{\alpha}_t}\right\}$, $\eta_t \leq \min\left\{\frac{1}{2L_F}, \frac{\sqrt{C_{11}'}I}{m}(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B})^{-1/2}\epsilon\right\}$, *where* $C_8', C_9', C_{\bar{\tau}}, C_{11}'$ *are constants specified in the proof, and by using a large mini-batch size of* $\mathcal{O}(1/\epsilon)$ *at the initial iteration for computing* $\mathbf{z}_1, \mathbf{s}_1, \mathbf{u}_1$ *and computing an accurate solution* $\mathbf{y}_1, \mathbf{v}_1$ *such that* $\delta_{y,1} \leq \mathcal{O}(m)$, *Algorithm 1 gives* $\mathbb{E}\left[\frac{1}{T}\sum_{i=1}^m \|\nabla F(\mathbf{x}_t)\|^2\right] \leq \epsilon^2$ *with sample complexity* $T = \mathcal{O}\left(\frac{m\epsilon^{-3}\mathbb{I}(I<m)}{I\sqrt{I}} + \frac{m\epsilon^{-3}}{I\sqrt{B}}\right)$.

First, we note that the bounded variance of $\nabla_v \phi_i(v, \mathbf{x}, \mathbf{y}_i; \mathcal{B}_i)$ can be derived as

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B}_i^t}[\|\nabla_v \phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t, \widetilde{\mathcal{B}}_i^t) - \nabla_v \phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t})\|^2] \\
&= \mathbb{E}_{\mathcal{B}_i^t, \widetilde{\mathcal{B}}_i^t}[\|\nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t)\mathbf{v}_{i,t} - \nabla_y f_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t) - \nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t})\mathbf{v}_i^t + \nabla_y f_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2] \\
&\leq \mathbb{E}_{\mathcal{B}_i^t, \widetilde{\mathcal{B}}_i^t}[2\|\nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t)\mathbf{v}_{i,t} - \nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t})\mathbf{v}_{i,t}\|^2 + 2\|\nabla_y f_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \nabla_y f_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t)\|^2] \\
&\leq \frac{2\sigma^2}{B}\|\mathbf{v}_{i,t}\|^2 + \frac{2\sigma^2}{B} \leq (1+\mathcal{V}^2)\frac{2\sigma^2}{B}.
\end{aligned}$$

Moreover, to achieve the variance-reduced estimation error bound, we need the stochastic gradient $\nabla_v \phi_i(\mathbf{v}_i, \mathbf{x}, \mathbf{y}_i; \xi, \zeta)$ to be $L_{\phi v}$-Lipschitz with some constant $L_{\phi v}$. The value of $L_{\phi v}$ can be derived as following. Assume that $(\mathbf{v}_i, \mathbf{x}, \mathbf{y}_i)$ and $(\mathbf{v}_i', \mathbf{x}', \mathbf{y}_i')$ are parameters from some iterations in algorithm 1, then under Assumptions 5.2 and 5.3 we have

$$\begin{aligned}
&\|\nabla_v \phi_i(\mathbf{v}_i, \mathbf{x}, \mathbf{y}_i; \xi, \zeta) - \nabla_v \phi_i(\mathbf{v}_i', \mathbf{x}', \mathbf{y}_i'; \xi, \zeta)\|^2 \\
&\leq 4\|\nabla_{yy}^2 g_i(\mathbf{x}, \mathbf{y}_i; \zeta)\mathbf{v}_{i,t} - \nabla_{yy}^2 g_i(\mathbf{x}', \mathbf{y}_i'; \zeta)\mathbf{v}_{i,t}\|^2 + 4\|\nabla_{yy}^2 g_i(\mathbf{x}', \mathbf{y}_i'; \zeta)\mathbf{v}_{i,t} - \nabla_{yy}^2 g_i(\mathbf{x}', \mathbf{y}_i'; \zeta)\mathbf{v}_{i,t}'\| \\
&\quad + 2\|\nabla_y f_i(\mathbf{x}, \mathbf{y}_i; \xi) - \nabla_y f_i(\mathbf{x}', \mathbf{y}_i'; \xi)\|^2 \\
&\leq (4L_{gyy}^2\mathcal{V}^2 + 2L_{fy}^2)(\|\mathbf{x} - \mathbf{x}'\|^2 + \|\mathbf{y}_i - \mathbf{y}_i'\|^2) + 4\tilde{C}_{gyy}^2\|\mathbf{v}_{i,t} - \mathbf{v}_{i,t}'\|^2 \\
&\leq L_{\phi v}^2(\|\mathbf{v}_{i,t} - \mathbf{v}_{i,t}'\|^2 + \|\mathbf{x} - \mathbf{x}'\|^2 + \|\mathbf{y}_i - \mathbf{y}_i'\|^2)
\end{aligned}$$

where $L_{\phi v} := \max\{4L_{gyy}^2\mathcal{V}^2 + 2L_{fy}^2, 4\tilde{C}_{gyy}^2\}$.

Lemma B.10 ((Ghadimi & Wang, 2018)(Lemma 2.2))**.** *For all* $i = 1, \ldots, m$, $\mathbf{v}_i(\mathbf{x}, \mathbf{y}_i)$ *is* L_v*-Lipschitz continuous with* $L_v = \frac{L_{fy}C_{gyy} + C_{fy}L_{gyy}}{\lambda^2}$.

Define

$$\Delta_t := \frac{1}{m}\sum_{i=1}^m \nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t})\mathbf{v}_{i,t} \tag{26}$$

Note that $\mathbb{E}_t[G_t] = \Delta_t$, $\mathbb{E}_t[\widetilde{G}_t] = \Delta_{t-1}$. Then we have the following two lemmas.

Lemma B.11. *For all* $t > 0$, *we have*

$$\|\Delta_t - \nabla F(\mathbf{x}_t)\|^2 \leq \frac{C_1'}{m}\delta_{y,t} + \frac{C_2'}{m}\delta_{v,t} \tag{27}$$

Lemma B.12. *For all* $t > 0$, *we have*

$$\begin{aligned}
\mathbb{E}_t[\|\mathbf{z}_{t+1} - \Delta_t\|^2] \leq (1-\beta_t)\|\mathbf{z}_t - \Delta_{t-1}\|^2 + 2C_3'\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \frac{2C_3'}{m}\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2 \\
+ \frac{2C_4'}{m}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 + 2\beta_t^2 C_5'\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right)
\end{aligned} \tag{28}$$

Following from Lemma B.3, with update $\mathbf{y}_{i,t+1} = \mathbf{y}_{i,t} - \tau_t\tau\mathbf{s}_{i,t}$ for all $i = 1, \ldots, m$, with $\tau \leq 2/(3L_g)$, we have

$$\mathbb{E}[\delta_{y,t+1}] \leq (1 - \frac{\tau_t\tau\lambda}{4})\mathbb{E}[\delta_{y,t}] + \frac{8\tau_t\tau}{\lambda}\mathbb{E}[\tilde{\delta}_{s,t}] - \frac{2\tau}{\tau_t}(\frac{1}{2\tau} - \frac{3L_g}{4})\mathbb{E}[\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2] + \frac{8mL_y^2}{\tau_t\tau\lambda}\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2] \tag{29}$$

Lemma B.13. *Consider the update* $\mathbf{v}_{i,t+1} = \Pi_{\mathcal{V}}[\mathbf{v}_{i,t} - \bar{\tau}_t\mathbf{u}_{i,t}]$ *for all* $i = 1, \ldots, m$, *with* $\bar{\tau}_t \leq \min\left\{\frac{\lambda}{8L_{\phi v}^2}, \frac{\lambda}{2}, \frac{1}{\lambda}\right\}$, *we have*

$$\mathbb{E}[\delta_{v,t+1}] \leq (1 - \frac{\lambda\bar{\tau}_t}{4})\mathbb{E}[\delta_{v,t}] + 10\lambda\bar{\tau}_t\mathbb{E}[\tilde{\delta}_{u,t}] + \frac{5mL_v^2}{\lambda\bar{\tau}_t}E[\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2] + \frac{5L_v^2}{\lambda\bar{\tau}_t}\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2] \tag{30}$$

By applying Lemma B.4 to $\mathbf{s}_t$, $\mathbf{u}_t$, we have

$$\mathbb{E}\left[\delta_{s,t+1}\right] \leq (1-\frac{I\alpha_t}{m})\mathbb{E}\left[\delta_{s,t}\right] + \frac{2I\alpha_t^2\sigma^2}{B} + \frac{8m^2L_{gy}^2}{I}\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \frac{1}{m}\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2\right]$$

and

$$\mathbb{E}\left[\delta_{u,t+1}\right] \leq (1-\frac{I\bar{\alpha}_t}{m})\mathbb{E}\left[\delta_{u,t}\right] + \frac{2I\bar{\alpha}_t^2\sigma^2}{B} + \frac{8m^2L_{\phi v}^2}{I}\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \frac{1}{m}\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2 + \frac{1}{m}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2\right]$$

which implies

$$\mathbb{E}\left[\sum_{t=1}^T \delta_{s,t}\right] \leq \mathbb{E}\left[\frac{m}{I\alpha_t}\delta_{s,1} + \frac{2m\alpha_t\sigma^2T}{B} + \frac{8m^3L_{gy}^2}{\alpha_t I^2}\sum_{t=1}^T\left(\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \frac{1}{m}\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2\right)\right] \tag{31}$$

and

$$\mathbb{E}\left[\sum_{t=1}^T \delta_{u,t}\right] \leq \mathbb{E}\left[\frac{m}{I\bar{\alpha}_t}\delta_{u,1} + \frac{2m\bar{\alpha}_t\sigma^2T}{B} + \frac{8m^3L_{\phi v}^2}{\bar{\alpha}_t I^2}\sum_{t=1}^T\left(\frac{1}{m}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2 + \|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \frac{1}{m}\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2\right)\right] \tag{32}$$

B.2.1. Proof of Theorem B.9

Proof. By Lemma B.2, we have

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) \leq \frac{\eta_t}{2}\|\mathbf{z}_{t+1} - \nabla F(\mathbf{x}_t)\|^2 - \frac{\eta_t}{2}\|\nabla F(\mathbf{x}_t)\|^2 - \frac{\eta_t}{4}\|\mathbf{z}_{t+1}\|^2. \tag{33}$$

The first term on the right hand side can be divided into two terms.

$$\|\mathbf{z}_{t+1} - \nabla F(\mathbf{x}_t)\|^2 \leq 2\|\mathbf{z}_{t+1} - \Delta_t\|^2 + 2\|\Delta_t - \nabla F(\mathbf{x}_t)\|^2 \tag{34}$$

where we have recursion for the first term on the right hand side in Lemma B.12 and the second term is bounded by Lemma B.11. Combining inequalities 33,34 and Lemma B.11 gives

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) \leq \eta_t\delta_{z,t} + \frac{\eta_t C_1'}{m}\delta_{y,t} + \frac{\eta_t C_2'}{m}\delta_{v,t} - \frac{\eta_t}{2}\|\nabla F(\mathbf{x}_t)\|^2 - \frac{\eta_t}{4}\|\mathbf{z}_{t+1}\|^2. \tag{35}$$

Taking summation over $t = 1, \ldots, T$ yields

$$\sum_{t=1}^T \|\nabla F(\mathbf{x}_t)\|^2 \leq \frac{2}{\eta_t}(F(\mathbf{x}_1) - F(\mathbf{x}^*)) + 2\sum_{t=1}^T\delta_{z,t} + \frac{2C_1'}{m}\sum_{t=1}^T\delta_{y,t} + \frac{2C_2'}{m}\sum_{t=1}^T\delta_{v,t} - \frac{1}{2}\sum_{t=1}^T\|\mathbf{z}_{t+1}\|^2 \tag{36}$$

We enlarge the values of constants C_1', C_2' so that

$$\begin{aligned}&\sum_{t=1}^T \|\nabla F(\mathbf{x}_t)\|^2 + \sum_{t=1}^T\delta_{z,t} + \frac{1}{m}\sum_{t=1}^T\delta_{y,t} + \frac{1}{m}\sum_{t=1}^T\delta_{v,t}\\ &\leq \frac{2}{\eta_t}(F(\mathbf{x}_1) - F(\mathbf{x}^*)) + 3\sum_{t=1}^T\delta_{z,t} + \frac{C_1'}{m}\sum_{t=1}^T\delta_{y,t} + \frac{C_2'}{m}\sum_{t=1}^T\delta_{v,t} - \frac{1}{2}\sum_{t=1}^T\|\mathbf{z}_{t+1}\|^2\end{aligned} \tag{37}$$

Following from inequality (29) and Lemma B.13 we have

$$\mathbb{E}\left[\sum_{t=1}^T \delta_{y,t}\right] \leq \mathbb{E}\left[\frac{4}{\tau_t\tau\lambda}\delta_{y,1} + \frac{32}{\lambda^2}\sum_{t=1}^T\tilde{\delta}_{s,t} - \frac{8}{\tau_t^2\lambda}(\frac{1}{2\tau} - \frac{3L_g}{4})\sum_{t=1}^T\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2 + \frac{32mL_y^2}{\tau_t^2\tau^2\lambda^2}\sum_{t=1}^T\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right] \tag{38}$$

$$\mathbb{E}\left[\sum_{i=1}^m \delta_{v,t}\right] \leq \mathbb{E}\left[\frac{4}{\lambda\bar{\tau}_t}\mathbb{E}[\delta_{v,1}] + 40\sum_{t=1}^T\tilde{\delta}_{u,t} + \frac{20L_v^2}{\lambda^2\bar{\tau}_t^2}\sum_{t=1}^T\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \frac{20mL_v^2}{\lambda^2\bar{\tau}_t^2}\sum_{t=1}^T\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2\right] \tag{39}$$

It follows from Lemma B.12 that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\delta_{z,t}\right] \leq \mathbb{E}\Bigg[&\frac{\delta_{z,1}}{\beta_t} + \frac{2C_3'}{\beta_t}\sum_{t=1}^{T}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2 + \frac{2C_3'}{m\beta_t}\sum_{t=1}^{T}\|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2 \\ &+ \frac{2C_4'}{m\beta_t}\sum_{t=1}^{T}\|\mathbf{v}_t-\mathbf{v}_{t+1}\|^2 + 2\beta_t C_5' T\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)\Bigg]\end{aligned} \tag{40}$$

Note that

$$\begin{aligned}\|\mathbf{v}_t-\mathbf{v}_{t+1}\|^2 &\leq \sum_{i=1}^{m}\bar{\tau}_t^2\|\mathbf{u}_{i,t}\|^2 \\ &= \sum_{i=1}^{m}\bar{\tau}_t^2\|\mathbf{u}_{i,t}-\nabla_v\phi_i(\mathbf{v}_{i,t},\mathbf{x}_t,\mathbf{y}_{i,t})+\nabla_v\phi_i(\mathbf{v}_{i,t},\mathbf{x}_t,\mathbf{y}_{i,t})\|^2 \\ &\leq \sum_{i=1}^{m}2\bar{\tau}_t^2\|\mathbf{u}_{i,t}-\nabla_v\phi_i(\mathbf{v}_{i,t},\mathbf{x}_t,\mathbf{y}_{i,t})\|^2 + 2\bar{\tau}_t^2\|\nabla_v\phi_i(\mathbf{v}_{i,t},\mathbf{x}_t,\mathbf{y}_{i,t})-\nabla_v\phi_i(\mathbf{v}_i(\mathbf{x}_t,\mathbf{y}_{i,t}),\mathbf{x}_t,\mathbf{y}_{i,t})\|^2 \\ &\leq \sum_{i=1}^{m}2\bar{\tau}_t^2\|\mathbf{u}_{i,t}-\nabla_v\phi_i(\mathbf{v}_{i,t},\mathbf{x}_t,\mathbf{y}_{i,t})\|^2 + 2\bar{\tau}_t^2L_{\phi v}^2\|\mathbf{v}_{i,t}-\mathbf{v}_i(\mathbf{x}_t,\mathbf{y}_{i,t})\|^2 \\ &= 2\bar{\tau}_t^2\tilde{\delta}_{u,t} + 2\bar{\tau}_t^2L_{\phi v}^2\delta_{v,t}\end{aligned} \tag{41}$$

Taking summation over all iterations and expectation, and combining with inequality (39) yields

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{v}_t-\mathbf{v}_{t+1}\|^2\right] \\ &\leq \mathbb{E}\left[2\bar{\tau}_t^2\sum_{t=1}^{T}\tilde{\delta}_{u,t} + 2\bar{\tau}_t^2L_{\phi v}^2\sum_{t=1}^{T}\delta_{v,t}\right] \\ &\leq \mathbb{E}\Bigg[\left(2\bar{\tau}_t^2+80L_{\phi v}^2\bar{\tau}_t^2\right)\sum_{t=1}^{T}\tilde{\delta}_{u,t} + \frac{8L_{\phi v}^2\bar{\tau}_t}{\lambda}\delta_{v,1} + \frac{40mL_{\phi v}^2L_v^2}{\lambda^2}\sum_{t=1}^{T}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2 + \frac{40L_v^2L_{\phi v}^2}{\lambda^2}\sum_{t=1}^{T}\|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2\Bigg] \\ &\overset{(a)}{\leq} \mathbb{E}\Bigg[\left(4\bar{\tau}_t^2+160L_{\phi v}^2\bar{\tau}_t^2\right)\sum_{t=1}^{T}\delta_{u,t} + \frac{8L_{\phi v}^2\bar{\tau}_t}{\lambda}\delta_{v,1} + \left(4\bar{\tau}_t^2L_{\phi v}^2+160L_{\phi v}^4\bar{\tau}_t^2\right)\sum_{t=1}^{T}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2 \\ &\quad+\left(\frac{40mL_{\phi v}^2L_v^2}{\lambda^2}+4m\bar{\tau}_t^2L_{\phi v}^2+160mL_{\phi v}^4\bar{\tau}_t^2\right)\sum_{t=1}^{T}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2 \\ &\quad+\left(\frac{40L_v^2L_{\phi v}^2}{\lambda^2}+4\bar{\tau}_t^2L_{\phi v}^2+160L_{\phi v}^4\bar{\tau}_t^2\right)\sum_{t=1}^{T}\|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2\Bigg] \\ &\overset{(b)}{\leq} \mathbb{E}\Bigg[C_6'\bar{\tau}_t^2\sum_{t=1}^{T}\delta_{u,t} + \frac{8L_{\phi v}^2\bar{\tau}_t}{\lambda}\delta_{v,1} + C_6'\bar{\tau}_t^2\sum_{t=1}^{T}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2 + \left(\frac{40mL_{\phi v}^2L_v^2}{\lambda^2}+mC_6'\bar{\tau}_t^2\right)\sum_{t=1}^{T}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2 \\ &\quad+\left(\frac{40L_v^2L_{\phi v}^2}{\lambda^2}+C_6'\bar{\tau}_t^2\right)\sum_{t=1}^{T}\|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2\Bigg]\end{aligned}$$

where inequality (a) follows from inequality (8) and (9), and in (b) we denote $C_6' = (4+160L_{\phi v}^2)\max\{1, L_{\phi v}^2\}$.

Combining with inequality (32), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{v}_t-\mathbf{v}_{t+1}\|^2\right]\\
&\leq \mathbb{E}\Bigg[\frac{mC_6'\bar{\tau}_t^2}{I\bar{\alpha}_t}\delta_{u,1}+\frac{8L_{\phi v}^2\bar{\tau}_t}{\lambda}\delta_{v,1}+\frac{2C_6'\bar{\tau}_t^2 m\bar{\alpha}_t\sigma^2 T}{B}+\left(C_6'\bar{\tau}_t^2+\frac{8C_6'\bar{\tau}_t^2 m^2 L_{\phi v}^2}{\bar{\alpha}_t I^2}\right)\sum_{t=1}^{T}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2\\
&\quad+\left(\frac{40mL_{\phi v}^2L_v^2}{\lambda^2}+mC_6'\bar{\tau}_t^2+\frac{8C_6'\bar{\tau}_t^2m^3L_{\phi v}^2}{\bar{\alpha}_tI^2}\right)\sum_{t=1}^{T}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2\\
&\quad+\left(\frac{40L_v^2L_{\phi v}^2}{\lambda^2}+C_6'\bar{\tau}_t^2+\frac{8C_6'\bar{\tau}_t^2m^2L_{\phi v}^2}{\bar{\alpha}_tI^2}\right)\sum_{t=1}^{T}\|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2\Bigg]
\end{aligned}$$

Setting $\bar{\tau}_t^2\leq\min\{\frac{1}{4C_6'},C_{\bar{\tau}}\frac{I^2}{m^2}\bar{\alpha}_t\}$ where $C_{\bar{\tau}}:=\frac{1}{32C_6'L_{\phi v}^2}$, i.e. $\frac{8C_6'm^2L_{\phi v}^2\bar{\tau}_t^2}{\bar{\alpha}_tI^2}\leq\frac{1}{4}$ and $C_6'\bar{\tau}^2\leq\frac{1}{4}$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{v}_t-\mathbf{v}_{t+1}\|^2\right]\leq\mathbb{E}\Bigg[&\frac{2mC_6'\bar{\tau}_t^2}{I\bar{\alpha}_t}\delta_{u,1}+\frac{16L_{\phi v}^2\bar{\tau}_t}{\lambda}\delta_{v,1}+\frac{4C_6'\bar{\tau}_t^2m\bar{\alpha}_t\sigma^2T}{B}\\
&+\left(\frac{80mL_{\phi v}^2L_v^2}{\lambda^2}+2mC_6'\bar{\tau}_t^2+\frac{16C_6'\bar{\tau}_t^2m^3L_{\phi v}^2}{\bar{\alpha}_tI^2}\right)\sum_{t=1}^{T}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2\\
&+\left(\frac{80L_v^2L_{\phi v}^2}{\lambda^2}+2C_6'\bar{\tau}_t^2+\frac{16C_6'\bar{\tau}_t^2m^2L_{\phi v}^2}{\bar{\alpha}_tI^2}\right)\sum_{t=1}^{T}\|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2\Bigg]
\end{aligned}\tag{42}$$

Combining inequalities (37), (40), (38), (39), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla F(\mathbf{x}_t)\|^2+\sum_{t=1}^{T}\delta_{z,t}+\frac{1}{m}\sum_{t=1}^{T}\delta_{y,t}+\frac{1}{m}\sum_{t=1}^{T}\delta_{v,t}\right]\\
&\leq\mathbb{E}\Bigg[\frac{2}{\eta_t}(F(\mathbf{x}_1)-F(\mathbf{x}^*))+\frac{3\delta_{z,1}}{\beta_t}+6\beta_tC_5'T\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)+\frac{4C_1'}{m\tau_t\tau\lambda}\delta_{y,1}+\frac{4C_2'}{m\lambda\bar{\tau}_t}\delta_{v,1}\\
&\quad+\left(\frac{6C_3'}{\beta_t}+\frac{32C_1'L_y^2}{\tau_t^2\tau^2\lambda^2}+\frac{20C_2'L_v^2}{\lambda^2\bar{\tau}_t^2}\right)\sum_{t=1}^{T}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2-\frac{1}{2}\sum_{t=1}^{T}\|\mathbf{z}_{t+1}\|^2\\
&\quad+\left(\frac{6C_3'}{m\beta_t}-\frac{8C_1'}{m\tau_t^2\lambda}(\frac{1}{2\tau}-\frac{3L_g}{4})+\frac{20C_2'L_v^2}{m\lambda^2\bar{\tau}_t^2}\right)\sum_{t=1}^{T}\|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2\\
&\quad+\frac{6C_4'}{m\beta_t}\sum_{t=1}^{T}\|\mathbf{v}_t-\mathbf{v}_{t+1}\|^2+\frac{32C_1'}{m\lambda^2}\sum_{t=1}^{T}\tilde{\delta}_{s,t}+\frac{40C_2'}{m}\sum_{t=1}^{T}\tilde{\delta}_{u,t}\Bigg]\\
&\overset{(a)}{\leq}\mathbb{E}\Bigg[\frac{2}{\eta_t}(F(\mathbf{x}_1)-F(\mathbf{x}^*))+\frac{3\delta_{z,1}}{\beta_t}+6\beta_tC_5'T\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)+\frac{4C_1'}{m\tau_t\tau\lambda}\delta_{y,1}+\frac{4C_2'}{m\lambda\bar{\tau}_t}\delta_{v,1}\\
&\quad+\left(\frac{6C_3'}{\beta_t}+\frac{32C_1'L_y^2}{\tau_t^2\tau^2\lambda^2}+\frac{20C_2'L_v^2}{\lambda^2\bar{\tau}_t^2}+\frac{64C_1'L_{gy}^2}{\lambda^2}+80C_2'L_{\phi v}^2\right)\sum_{t=1}^{T}\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2-\frac{1}{2}\sum_{t=1}^{T}\|\mathbf{z}_{t+1}\|^2\\
&\quad+\left(\frac{6C_3'}{m\beta_t}-\frac{8C_1'}{m\tau_t^2\lambda}(\frac{1}{2\tau}-\frac{3L_g}{4})+\frac{20C_2'L_v^2}{m\lambda^2\bar{\tau}_t^2}+\frac{64C_1'L_{gy}^2}{m\lambda^2}+\frac{80C_2'L_{gyy}^2}{m}\right)\sum_{t=1}^{T}\|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2\\
&\quad+\left(\frac{6C_4'}{m\beta_t}+\frac{80C_2'L_{gyy}^2}{m}\right)\sum_{t=1}^{T}\|\mathbf{v}_t-\mathbf{v}_{t+1}\|^2+\frac{64C_1'}{m\lambda^2}\sum_{t=1}^{T}\delta_{s,t}+\frac{80C_2'}{m}\sum_{t=1}^{T}\delta_{u,t}
\end{aligned}\tag{43}$$

where (a) uses inequality (8) and (9).

Enlarge the value of constant C_2' so that $C_2' \geq \max\left\{\frac{64C_1'}{\lambda^2}+1, 80C_2'+1\right\}$.

Combining with inequalities (31), (32), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \|\nabla F(\mathbf{x}_t)\|^2 + \sum_{t=1}^T \delta_{z,t} + \frac{1}{m}\sum_{t=1}^T \delta_{y,t} + \frac{1}{m}\sum_{t=1}^T \delta_{v,t} + \frac{1}{m}\sum_{t=1}^T \delta_{s,t} + \frac{1}{m}\sum_{t=1}^T \delta_{u,t}\right] \\
&\leq \mathbb{E}\Bigg[\frac{2}{\eta_t}(F(\mathbf{x}_1)-F(\mathbf{x}^*)) + \frac{3\delta_{z,1}}{\beta_t} + 6\beta_t C_5' T\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right) + \frac{4C_1'}{m\tau_t\tau\lambda}\delta_{y,1} + \frac{4C_2'}{m\lambda\bar{\tau}_t}\delta_{v,1} + \frac{64C_1'}{\lambda^2 I\alpha_t}\delta_{s,1} + \frac{80C_2'}{I\bar{\alpha}_t}\delta_{u,1} \\
&\quad + \frac{128C_1'\alpha_t\sigma^2 T}{\lambda^2 B} + \frac{160C_2'\bar{\alpha}_t\sigma^2 T}{B} + \Bigg(\frac{6C_3'}{\beta_t} + \frac{32C_1'L_y^2}{\tau_t^2\tau^2\lambda^2} + \frac{20C_2'L_v^2}{\lambda^2\bar{\tau}_t^2} + \frac{64C_1'L_{gy}^2}{\lambda^2} + 80C_2'L_{\phi v}^2 \\
&\quad + \frac{512m^2L_{gy}^2}{\lambda^2\alpha_t I^2} + \frac{640C_2'm^2L_{\phi v}^2}{\bar{\alpha}_t I^2}\Bigg)\sum_{t=1}^T \|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2 - \frac{1}{2}\sum_{t=1}^T \|\mathbf{z}_{t+1}\|^2 + \Bigg(\frac{6C_3'}{m\beta_t} - \frac{8C_1'}{m\tau_t^2\lambda}\left(\frac{1}{2\tau}-\frac{3L_g}{4}\right) \\
&\quad + \frac{20C_2'L_v^2}{m\lambda^2\bar{\tau}_t^2} + \frac{64C_1'L_{gy}^2}{m\lambda^2} + \frac{80C_2'L_{gyy}^2}{m} + \frac{512mL_{gy}^2}{\lambda^2\alpha_t I^2} + \frac{640C_2'mL_{\phi v}^2}{\bar{\alpha}_t I^2}\Bigg)\sum_{t=1}^T \|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2 \\
&\quad + \left(\frac{6C_4'}{m\beta_t} + \frac{80C_2'L_{gyy}^2}{m} + \frac{640C_2'mL_{\phi v}^2}{\bar{\alpha}_t I^2}\right)\sum_{t=1}^T \|\mathbf{v}_t-\mathbf{v}_{t+1}\|^2\Bigg]
\end{aligned} \tag{44}$$

Combining with inequality (42), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \|\nabla F(\mathbf{x}_t)\|^2 + \sum_{t=1}^T \delta_{z,t} + \frac{1}{m}\sum_{t=1}^T \delta_{y,t} + \frac{1}{m}\sum_{t=1}^T \delta_{v,t} + \frac{1}{m}\sum_{t=1}^T \delta_{s,t} + \frac{1}{m}\sum_{t=1}^T \delta_{u,t}\right] \\
&\leq \mathbb{E}\Bigg[\frac{2}{\eta_t}(F(\mathbf{x}_1)-F(\mathbf{x}^*)) + \frac{3\delta_{z,1}}{\beta_t} + 6\beta_t C_5' T\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right) + \frac{4C_1'}{m\tau_t\tau\lambda}\delta_{y,1} + \frac{4C_2'}{m\lambda\bar{\tau}_t}\delta_{v,1} + \frac{64C_1'}{\lambda^2 I\alpha_t}\delta_{s,1} + \frac{80C_2'}{I\bar{\alpha}_t}\delta_{u,1} \\
&\quad + \frac{128C_1'\alpha_t\sigma^2 T}{\lambda^2 B} + \frac{160C_2'\bar{\alpha}_t\sigma^2 T}{B} + \Bigg(\frac{6C_3'}{\beta_t} + \frac{32C_1'L_y^2}{\tau_t^2\tau^2\lambda^2} + \frac{20C_2'L_v^2}{\lambda^2\bar{\tau}_t^2} + \frac{64C_1'L_{gy}^2}{\lambda^2} + 80C_2'L_{\phi v}^2 \\
&\quad + \frac{512m^2L_{gy}^2}{\lambda^2\alpha_t I^2} + \frac{640C_2'm^2L_{\phi v}^2}{\bar{\alpha}_t I^2}\Bigg)\sum_{t=1}^T \|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2 - \frac{1}{2}\sum_{t=1}^T \|\mathbf{z}_{t+1}\|^2 + \Bigg(\frac{6C_3'}{m\beta_t} - \frac{8C_1'}{m\tau_t^2\lambda}\left(\frac{1}{2\tau}-\frac{3L_g}{4}\right) \\
&\quad + \frac{20C_2'L_v^2}{m\lambda^2\bar{\tau}_t^2} + \frac{64C_1'L_{gy}^2}{m\lambda^2} + \frac{80C_2'L_{gyy}^2}{m} + \frac{512mL_{gy}^2}{\lambda^2\alpha_t I^2} + \frac{640C_2'mL_{\phi v}^2}{\bar{\alpha}_t I^2}\Bigg)\sum_{t=1}^T \|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2 \\
&\quad + \left(\frac{6C_4'}{m\beta_t} + \frac{80C_2'L_{gyy}^2}{m} + \frac{640C_2'mL_{\phi v}^2}{\bar{\alpha}_t I^2}\right)\Bigg[\frac{2mC_6'\bar{\tau}_t^2}{I\bar{\alpha}_t}\delta_{u,1} + \frac{16L_{\phi v}^2\bar{\tau}_t}{\lambda}\delta_{v,1} + \frac{4C_6'\bar{\tau}_t^2 m\bar{\alpha}_t\sigma^2 T}{B} \\
&\quad + \left(\frac{80mL_{\phi v}^2L_v^2}{\lambda^2} + 2mC_6'\bar{\tau}_t^2 + \frac{16C_6'\bar{\tau}_t^2 m^3L_{\phi v}^2}{\bar{\alpha}_t I^2}\right)\sum_{t=1}^T \|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2 \\
&\quad + \left(\frac{80L_v^2L_{\phi v}^2}{\lambda^2} + 2C_6'\bar{\tau}_t^2 + \frac{16C_6'\bar{\tau}_t^2 m^2L_{\phi v}^2}{\bar{\alpha}_t I^2}\right)\sum_{t=1}^T \|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2\Bigg]\Bigg]
\end{aligned} \tag{45}$$

Recall that $\bar{\tau}_t^2 \leq \min\{\frac{1}{4C_6'}, C_{\bar{\tau}}\frac{I^2}{m^2}\bar{\alpha}_t\}$, i.e. $2C_6'\bar{\tau}_t^2 + \frac{16C_6'\bar{\tau}_t^2 m^2L_{\phi v}^2}{\bar{\alpha}_t I^2} \leq 1$, and let

$$C_7' = \max\left\{18C_4', 240C_2'L_{gyy}^2, 1920C_2'L_{\phi v}^2\right\}\left(\frac{80L_v^2L_{\phi v}^2}{\lambda^2}+1\right)$$

then

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \|\nabla F(\mathbf{x}_t)\|^2 + \sum_{t=1}^T \delta_{z,t} + \frac{1}{m}\sum_{t=1}^T \delta_{y,t} + \frac{1}{m}\sum_{t=1}^T \delta_{v,t} + \frac{1}{m}\sum_{t=1}^T \delta_{s,t} + \frac{1}{m}\sum_{t=1}^T \delta_{u,t}\right] \\
&\le \mathbb{E}\Bigg[\frac{2}{\eta_t}(F(\mathbf{x}_1) - F(\mathbf{x}^*)) + \frac{3\delta_{z,1}}{\beta_t} + 6\beta_t C_5' T\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right) + \frac{4C_1'}{m\tau_t\tau\lambda}\delta_{y,1} + \frac{4C_2'}{m\lambda\bar{\tau}_t}\delta_{v,1} + \frac{64C_1'}{\lambda^2 I\alpha_t}\delta_{s,1} \\
&\quad + \frac{80C_2'}{I\bar{\alpha}_t}\delta_{u,1} + \frac{128C_1'\alpha_t\sigma^2 T}{\lambda^2 B} + \frac{160C_2'\bar{\alpha}_t\sigma^2 T}{B} + \Bigg(\frac{6C_3'}{\beta_t} + \frac{32C_1'L_y^2}{\tau_t^2\tau^2\lambda^2} + \frac{20C_2'L_v^2}{\lambda^2\bar{\tau}_t^2} + \frac{64C_1'L_{gy}^2}{\lambda^2} + 80C_2'L_{\phi v}^2 \\
&\quad + \frac{512m^2L_{gy}^2}{\lambda^2\alpha_t I^2} + \frac{640C_2'm^2L_{\phi v}^2}{\bar{\alpha}_t I^2}\Bigg)\sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 - \frac{1}{2}\sum_{t=1}^T \|\mathbf{z}_{t+1}\|^2 + \Bigg(\frac{6C_3'}{m\beta_t} - \frac{8C_1'}{m\tau_t^2\lambda}(\frac{1}{2\tau} - \frac{3L_g}{4}) \\
&\quad + \frac{20C_2'L_v^2}{m\lambda^2\bar{\tau}_t^2} + \frac{64C_1'L_{gy}^2}{m\lambda^2} + \frac{80C_2'L_{gyy}^2}{m} + \frac{512mL_{gy}^2}{\lambda^2\alpha_t I^2} + \frac{640C_2'mL_{\phi v}^2}{\bar{\alpha}_t I^2}\Bigg)\sum_{t=1}^T \|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2 \\
&\quad + C_7'\left(\frac{1}{m\beta_t} + \frac{1}{m} + \frac{m}{\bar{\alpha}_t I^2}\right)\left(\frac{2mC_6'\bar{\tau}_t^2}{I\bar{\alpha}_t}\delta_{u,1} + \frac{16L_{\phi v}^2\bar{\tau}_t}{\lambda}\delta_{v,1} + \frac{4C_6'\bar{\tau}_t^2 m\bar{\alpha}_t\sigma^2 T}{B}\right) \\
&\quad + C_7'\left(\frac{1}{\beta_t} + 1 + \frac{m^2}{\bar{\alpha}_t I^2}\right)\sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + C_7'\left(\frac{1}{m\beta_t} + \frac{1}{m} + \frac{m}{\bar{\alpha}_t I^2}\right)\sum_{t=1}^T \|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2\Bigg]\Bigg]
\end{aligned}
\tag{46}
$$

To ensure the coefficient of $\sum_{t=1}^T \|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2$, we set

$$
\begin{aligned}
\tau_t^2 &= C_8' \min\left\{\beta_t, \bar{\tau}_t^2, \frac{I^2\alpha_t}{m^2}, \frac{I^2\bar{\alpha}_t}{m^2}\right\} \\
&\le \frac{8C_1'}{9m\lambda}(\frac{1}{2\tau} - \frac{3L_g}{4}) \min\left\{\frac{m\beta_t}{6C_3'}, \frac{m\lambda^2\bar{\tau}_t^2}{20C_2'L_v^2}, \frac{m\lambda^2}{64C_1'L_{gy}^2}, \frac{m}{80C_2'L_{gyy}^2}, \frac{\lambda^2\alpha_t I^2}{512mL_{gy}^2}, \frac{\bar{\alpha}_t I^2}{640C_2'mL_{\phi v}^2}, \frac{m\beta_t}{C_7'}, \frac{m}{C_7'}, \frac{\bar{\alpha}_t I^2}{mC_7'}\right\}
\end{aligned}
$$

where $C_8' = \frac{8C_1'}{9\lambda}(\frac{1}{2\tau} - \frac{3L_g}{4}) \min\left\{\frac{1}{6C_3'}, \frac{\lambda^2}{20C_2'L_v^2}, \frac{\lambda^2}{64C_1'L_{gy}^2}, \frac{1}{80C_2'L_{gyy}^2}, \frac{\lambda^2}{512L_{gy}^2}, \frac{1}{640C_2'L_{\phi v}^2}, \frac{1}{C_7'}\right\}$. Let

$$
C_9' = 11\max\left\{3, 6C_5', \frac{4C_1'}{\tau\lambda}, \frac{4C_2'}{\lambda}, \frac{64C_1'}{\lambda^2}, 80C_2', \frac{128C_1'\sigma^2}{\lambda^2}, 160C_2'\sigma^2, 2C_7'C_6', \frac{16C_7'L_{\phi v}^2}{\lambda}, 4C_7'C_6'\sigma^2\right\}
$$

It follows

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \|\nabla F(\mathbf{x}_t)\|^2 + \sum_{t=1}^T \delta_{z,t} + \frac{1}{m}\sum_{t=1}^T \delta_{y,t} + \frac{1}{m}\sum_{t=1}^T \delta_{v,t} + \frac{1}{m}\sum_{t=1}^T \delta_{s,t} + \frac{1}{m}\sum_{t=1}^T \delta_{u,t}\right] \\
&\le \mathbb{E}\Bigg[\frac{2}{\eta_t}(F(\mathbf{x}_1) - F(\mathbf{x}^*)) + C_9'\Bigg(\frac{\delta_{z,1}}{\beta_t} + \frac{\delta_{y,1}}{m\tau_t} + \left(\frac{1}{m\bar{\tau}_t} + \frac{\bar{\tau}_t}{m\beta_t} + \frac{\bar{\tau}_t}{m} + \frac{m\bar{\tau}_t}{\bar{\alpha}_t I^2}\right)\delta_{v,1} \\
&\quad + \frac{\delta_{s,1}}{I\alpha_t} + \left(\frac{1}{I\bar{\alpha}_t} + \frac{\bar{\tau}_t^2}{I\bar{\alpha}_t\beta_t} + \frac{\bar{\tau}_t^2}{I\bar{\alpha}_t} + \frac{m^2\bar{\tau}_t^2}{I\bar{\alpha}_t^2 I^2}\right)\delta_{u,1} \\
&\quad + \beta_t T\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right) + \frac{\alpha_t T}{B} + \frac{\bar{\alpha}_t T}{B} + \frac{\bar{\tau}_t^2\bar{\alpha}_t T}{\beta_t B} + \frac{\bar{\tau}_t^2\bar{\alpha}_t T}{B} + \frac{m^2\bar{\tau}_t^2 T}{BI^2}\Bigg) \\
&\quad + \Bigg(\frac{6C_3'}{\beta_t} + \frac{32C_1'L_y^2}{\tau_t^2\tau^2\lambda^2} + \frac{20C_2'L_v^2}{\lambda^2\bar{\tau}_t^2} + \frac{64C_1'L_{gy}^2}{\lambda^2} + 80C_2'L_{\phi v}^2 \\
&\quad + \frac{512m^2L_{gy}^2}{\lambda^2\alpha_t I^2} + \frac{640C_2'm^2L_{\phi v}^2}{\bar{\alpha}_t I^2}\Bigg)\sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 - \frac{1}{2}\sum_{t=1}^T \|\mathbf{z}_{t+1}\|^2 \\
&\quad + C_7'\left(\frac{1}{\beta_t} + 1 + \frac{m^2}{\bar{\alpha}_t I^2}\right)\sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2\Bigg]\Bigg]
\end{aligned}
\tag{47}
$$

Setting $\beta_t, \bar{\alpha}_t \leq \frac{\epsilon^2}{18C_9'\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)}$, $\alpha_t \leq \frac{B}{18C_9'}\epsilon^2$, $\bar{\tau}_t^2 \leq \frac{C_{\bar{\tau}}I^2}{m^2}\bar{\alpha}_t$, we have

$$C_9'\left(\beta_t\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)+\frac{\alpha_t}{B}+\frac{\bar{\alpha}_t}{B}+\frac{\bar{\tau}_t^2\bar{\alpha}_t}{\beta_t B}+\frac{\bar{\tau}_t^2\bar{\alpha}_t}{B}+\frac{m^2\bar{\tau}_t^2}{BI^2}\right) \leq \frac{\epsilon^2}{3} \tag{48}$$

and

$$\begin{aligned}
\tau_t^2 &\leq C_8' \min\left\{\beta_t, \bar{\tau}_t^2, \frac{I^2\alpha_t}{m^2}, \frac{I^2\bar{\alpha}_t}{m^2},\right\} \\
&= \frac{C_8'}{18C_9'}\min\left\{\frac{\epsilon^2}{\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}}, \frac{C_{\bar{\tau}}I^2\epsilon^2}{m^2\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)}\epsilon^2, \frac{I^2B}{m^2}\epsilon^2, \frac{I^2\epsilon^2}{m^2\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)}\epsilon^2,\right\} \\
&= \frac{C_8'}{18C_9'}\frac{I^2\min\{1, C_{\bar{\tau}}\}}{m^2\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)}\epsilon^2
\end{aligned} \tag{49}$$

To ensure the coefficient of $\sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2$ is non-positive, we set

$$\begin{aligned}
\eta_t^2 &\leq C_{11}'\frac{I^2\epsilon^2}{m^2}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1} \\
&\leq C_{10}'\min\left\{\beta_t, \tau_t^2, \bar{\tau}_t^2, \frac{\alpha_t I^2}{m^2}, \frac{\bar{\alpha}_t I^2}{m^2}\right\} \\
&\leq \frac{1}{20}\min\left\{\frac{\beta_t}{6C_3'}, \frac{\tau_t^2\tau^2\lambda^2}{32C_1'L_y^2}, \frac{\lambda^2\bar{\tau}_t^2}{20C_2'L_v^2}, \frac{\lambda^2}{64C_1'L_{gy}^2}, \frac{1}{80C_2'L_{\phi v}^2}, \frac{\lambda^2\alpha_t I^2}{512m^2L_{gy}^2}, \frac{\bar{\alpha}_t I^2}{640C_2'm^2L_{\phi v}^2}, \frac{\beta_t}{C_7'}, \frac{1}{C_7'}, \frac{\bar{\alpha}_t I^2}{m^2C_7'}\right\}
\end{aligned} \tag{50}$$

where $C_{10}' = \frac{1}{20}\min\left\{\frac{1}{6C_3'}, \frac{\tau^2\lambda^2}{32C_1'L_y^2}, \frac{\lambda^2}{20C_2'L_v^2}, \frac{\lambda^2}{64C_1'L_{gy}^2}, \frac{1}{80C_2'L_{\phi v}^2}, \frac{\lambda^2}{512L_{gy}^2}, \frac{1}{640C_2'L_{\phi v}^2}, \frac{1}{C_7'}, \frac{1}{C_7'}, \frac{1}{C_7'}\right\}$, and $C_{11}' = C_{10}'\min\left\{\frac{1}{18C_9'}, \frac{C_8'\min\{1,C_{\bar{\tau}}\}}{18C_9'}, \frac{C_{\bar{\tau}}}{18C_9'}\right\}$

Thus, with $T = c_T\epsilon^{-3} := 6\Delta\frac{m}{\sqrt{C_{11}'}I}\left(\frac{\mathbb{I}(I<m)}{\sqrt{I}}+\frac{1}{\sqrt{B}}\right)\epsilon^{-3} \geq 6\Delta\frac{m}{\sqrt{C_{11}'}I}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{1/2}\epsilon^{-3}$, we have

$$\frac{2\Delta}{\eta_t T} = 2\Delta\frac{m}{\sqrt{C_{11}'}I\sqrt{\min\{I,B\}}}\epsilon^{-1}\frac{1}{T} = \frac{\epsilon^2}{3} \tag{51}$$

$\frac{C_9'}{T}\left[\frac{\delta_{z,1}}{\beta_t}+\frac{\delta_{y,1}}{m\tau_t}+\left(\frac{1}{m\bar{\tau}_t}+\frac{\bar{\tau}_t}{m\beta_t}+\frac{\bar{\tau}_t}{m}+\frac{m\bar{\tau}_t}{\bar{\alpha}_t I^2}\right)\delta_{v,1}+\frac{\delta_{s,1}}{I\alpha_t}+\left(\frac{1}{I\bar{\alpha}_t}+\frac{\bar{\tau}_t^2}{I\bar{\alpha}_t\beta_t}+\frac{\bar{\tau}_t^2}{I\bar{\alpha}_t}+\frac{m^2\bar{\tau}_t^2}{\bar{\alpha}_t^2 I^3}\right)\delta_{u,1}\right] \leq \frac{\epsilon^2}{3}$ can be achieved by processing all lower problems at the beginning and finding good initial solutions $\delta_{z,1}, \delta_{s,1}, \delta_{u,1}$ with accuracy $\mathcal{O}(\epsilon)$ with complexity $\mathcal{O}(\epsilon^{-1})$, and $\delta_{y,1}, \delta_{v,1}$ with accuracy $\mathcal{O}(1)$ with complexity $\mathcal{O}(1)$. Denote the iteration number for initialization as T_0. Then the total iteration complexity is $\mathcal{O}\left(\frac{m\epsilon^{-3}\mathbb{I}(I<m)}{I\sqrt{I}}+\frac{m\epsilon^{-3}}{I\sqrt{B}}\right)$.

□

C. Convergence Analysis of RE-BSVRB

C.1. Convergence Analysis of RE-BSVRB$^{\text{v1}}$

We present the formal statement of Theorem 6.1 for RE-BSVRB$^{\text{v1}}$.

Theorem C.1. *Suppose Assumptions 5.1 and 5.2 hold and the PL condition holds. Set* $\alpha_1 = \bar{\alpha}_1 = \beta_1 \leq \frac{1}{2}$, $\tau_1 = \sqrt{\frac{C_8 I\alpha_1}{m}}$, $\eta_1 = \min\left\{\frac{1}{2L_F}, \sqrt{\frac{C_9 I^2\alpha_1}{m^2}}\right\}$, $T_1 = O\left(\max\left\{\frac{m}{\mu\eta_1}, \frac{m}{\mu\beta_1}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1}, \frac{m}{\mu\tau_1}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1}\right\}\right)$. *Define a constant* $\epsilon_1 = \frac{7C_{10}(\beta_1+\alpha_1+\bar{\alpha}_1)}{\mu}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)$ *and* $\epsilon_k = \epsilon_1/2^{k-1}$. *For* $k \geq 2$, *setting* $\beta_k = \alpha_k = \bar{\alpha}_k \leq \frac{\mu\epsilon_k}{21C_{10}}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1}$, $\tau_k = \frac{\sqrt{C_8\alpha_k}I}{m}$, $\eta_k = \sqrt{C_9}\sqrt{\min\left\{\tau_k^2, \frac{\alpha_k I^2}{m^2}\right\}}$, $T_k = O\left(\max\left\{\frac{1}{\mu\eta_k}, \frac{1}{\beta_k}, \frac{1}{\tau_k}\right\}\right)$, *where and* $C_1 \sim C_{11}$ *are as used in Theorem B.5, then after* $K = O(\log(\epsilon_1/\epsilon))$ *stages, the output of RE-BSVRB$^{\text{v1}}$ satisfies* $\mathbb{E}[F(\mathbf{x}_K) - F(\mathbf{x}^*)] \leq \epsilon$.

Proof. Following from the proof of Theorem B.5, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla F(\mathbf{x}_t)\|^2+\sum_{t=1}^{T}\delta_{z,t}+\frac{1}{m}\sum_{t=1}^{T}\delta_{y,t}+\frac{1}{m}\sum_{t=1}^{T}\delta_{H,t}+\frac{1}{m}\sum_{t=1}^{T}\delta_{s,t}\right]\\
&\leq \mathbb{E}\left[\frac{2\Delta}{\eta_t}+C_{10}\left(\frac{\delta_{z,1}}{\beta_{t+1}}+\frac{\delta_{s,1}}{I\alpha_t}+\frac{\delta_{y,1}}{\tau_t m}+\frac{\delta_{H,1}}{I\bar{\alpha}_t}+\frac{\bar{\alpha}_t T}{B}+\frac{\alpha_t T}{B}+\beta_{t+1}T\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)+\frac{I\bar{\alpha}_{t+1}^2 T}{m\beta_{t+1}B}\right)\right] \qquad (52)\\
&\overset{(a)}{\leq} \mathbb{E}\left[\frac{2\Delta}{\eta_t}+C_{10}\left(\frac{\delta_{z,1}}{\beta_{t+1}}+\frac{\delta_{s,1}}{m\alpha_t}+\frac{\delta_{y,1}}{\tau_t m}+\frac{\delta_{H,1}}{m\bar{\alpha}_t}+(\bar{\alpha}_t+\alpha_t+\beta_{t+1})T\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)\right)\right]
\end{aligned}$$

where in (a) we redefine the constant $C_{10}=\frac{m}{I}C_{10}$ and use the setting $\bar{\alpha}_t=\beta_{t+1}$.

From Theorem B.5, we know that it is required that $\alpha_0,\bar{\alpha}_0,\beta_0\leq\frac{1}{2}$, $\tau_0=\sqrt{C_8}\sqrt{\min\left\{\frac{I\beta_0}{m},\frac{I^2\bar{\alpha}_0}{m^2},\frac{I^2\alpha_0}{m^2}\right\}}$, $\eta_0=\min\left\{\frac{1}{2L_F},\sqrt{C_9}\sqrt{\min\left\{\frac{I\beta_0}{m},\frac{I^2\bar{\alpha}_0}{m^2},\frac{I^2\alpha_0}{m^2}\right\}}\right\}$.

Without loss of generality, let us set $\alpha_0=\bar{\alpha}_0=\beta_0$ and assume that $\epsilon_0=2\Delta>\frac{7C_{10}(\beta_0+\alpha_0+\bar{\alpha}_0)}{\mu}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)$. The case that $2\Delta\leq\frac{7C_{10}(\beta_0+\alpha_0+\bar{\alpha}_0)}{\mu}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)$ can be simply covered by our proof. Then denotes $\epsilon_1=\frac{7C_{10}(\beta_0+\alpha_0+\bar{\alpha}_0)}{\mu}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)$, and $\epsilon_k=\epsilon_1/2^{k-1}$.

In the first epoch ($k=1$), we have initialization such that $F(\mathbf{x}_1)-F(\mathbf{x}^*)\leq\Delta$. In the following, we let the last subscript denote the epoch index. Setting $\eta_1=\eta_0$, $\beta_1=\beta_0$, $\alpha_1=\alpha_0$, $\bar{\alpha}_1=\bar{\alpha}_0$, $\tau_1=\tau_0$, and $T_1=\max\left\{\frac{7\Delta}{\mu\eta_1},\frac{7mC_{10}}{\mu\beta_1}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1}(\delta_{z,0}+\delta_{s,0}+\delta_{w,0}),\frac{7C_{10}m}{\mu\tau_1}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1}\delta_{y,0}\right\}$. We bound the error of the first stage's output as follows,

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla F(\mathbf{x}_1)\|^2+\delta_{z,1}+\frac{1}{m}\delta_{y,1}+\frac{1}{m}\delta_{s,1}+\frac{1}{m}\delta_{H,1}\right]\\
&\leq\frac{2\Delta}{\eta_1T_1}+\frac{C_{10}}{T_1}\left(\frac{1}{\beta_1}\delta_{z,0}+\frac{1}{\tau_1 m}\delta_{y,0}+\frac{1}{\alpha_1 m}\delta_{s,0}+\frac{1}{\bar{\alpha}_1 m}\delta_{w,0}\right)+C_{10}(\beta_1+\alpha_1+\bar{\alpha}_1)\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right) \qquad (53)\\
&\leq\mu\epsilon_1
\end{aligned}$$

where the first inequality uses (52) and the fact that the output of each epoch is randomly sampled from all iterations, and the last line uses the choice of $\eta_1,\beta_1,\alpha_1,\bar{\alpha}_1,\tau_1,T_1,\epsilon_1$. If follows that

$$\mathbb{E}[F(\mathbf{x}_1)-\mathcal{F}(\mathbf{x}^*)]\leq\frac{1}{2\mu}\mathbb{E}[\|\nabla F(\mathbf{x}_1)\|^2]\leq\frac{\epsilon_1}{2}. \qquad (54)$$

Starting from the second stage, we will prove by induction. Suppose we are at k-th stage. Assuming that the output of $(k-1)$-the stage satisfies that $\mathbb{E}[F(\mathbf{x}_{k-1})-F(\mathbf{x}^*)]\leq\epsilon_{k-1}$ and $\mathbb{E}\left[\delta_{z,k-1}+\frac{\delta_{y,k-1}}{m}+\frac{\delta_{s,k-1}}{m}+\frac{\delta_{w,k-1}}{m}\right]\leq\mu\epsilon_{k-1}$, and setting $\beta_k=\alpha_k=\bar{\alpha}_k\leq\frac{\mu\epsilon_k}{21C_{10}}\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)^{-1}$, $\tau_k^2=C_8\frac{\alpha_k I^2}{m^2}$, $\eta_k^2=C_9\min\left\{\tau_k^2,\frac{\alpha_k I^2}{m^2}\right\}$, $T_k=\max\left\{\frac{28}{\mu\eta_k},\frac{7C_{10}}{\beta_k},\frac{7C_{10}}{\tau_k}\right\}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla F(\mathbf{x}_k)\|^2+\delta_{z,k}+\frac{1}{m}\delta_{y,k}+\frac{1}{m}\delta_{s,k}+\frac{1}{m}\delta_{w,k}\right]\\
&\leq\mathbb{E}\left[\frac{2(F(\mathbf{x}_{k-1})-F(\mathbf{x}^*))}{\eta_kT_k}+\frac{C_{10}}{T_k}\left(\frac{1}{\beta_k}\delta_{z,k-1}+\frac{1}{\tau_k m}\delta_{y,k-1}+\frac{1}{\alpha_k m}\delta_{s,k-1}+\frac{1}{\bar{\alpha}_k m}\delta_{w,k-1}\right)\right.\\
&\qquad\left.+C_{10}(\beta_k+\alpha_k+\bar{\alpha}_k)\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)\right] \qquad (55)\\
&\leq\mathbb{E}\left[\frac{2\epsilon_{k-1}}{\eta_kT_k}+\frac{C_{10}\mu\epsilon_{k-1}}{T_k}\left(\frac{1}{\beta_k}+\frac{1}{\tau_k}+\frac{1}{\alpha_k}+\frac{1}{\bar{\alpha}_k}\right)+C_{10}(\beta_k+\alpha_k+\bar{\alpha}_k)\left(\frac{\mathbb{I}(I<m)}{I}+\frac{1}{B}\right)\right]\\
&\leq\mu\epsilon_k
\end{aligned}$$

It follows that

$$\mathbb{E}[F(\mathbf{x}_k) - F(\mathbf{x}^*)] \leq \frac{1}{2\mu}\mathbb{E}[\|\nabla F(\mathbf{x}_k)\|^2] \leq \frac{\epsilon_k}{2}. \tag{56}$$

Thus, after $K = 1 + \log_2(\epsilon_1/\epsilon) \leq \log_2(\epsilon_0/\epsilon)$ stages, $\mathbb{E}[F(\mathbf{x}_k) - F(\mathbf{x}^*)] \leq \epsilon$.

□

C.2. Convergence Analysis of RE-BSVRB^{v2}

We present the formal statement of Thoerem 6.1 for RE-BSVRB^{v2}.

Theorem C.2. *Suppose Assumptions 5.1 and 5.2 hold and the PL condition holds. Set* $\beta_1 = \alpha_1 = \bar{\alpha}_1 \leq \frac{1}{2}$, $\bar{\tau}_1 = \min\left\{\frac{\lambda}{8L_{\phi v}^2}, \frac{\lambda}{2}, \frac{1}{\lambda}, \frac{\sqrt{C_{\bar{\tau}}}I\bar{\alpha}_1^{1/2}}{m}\right\}$, $\tau_1 = \sqrt{C_8'}\sqrt{\min\left\{\bar{\tau}_1^2, \frac{I^2\alpha_1}{m^2}\right\}}$, $\eta_1 = \min\left\{\frac{1}{2L_f}, \sqrt{C_{10}'}\min\left\{\sqrt{m}\tau_1, \sqrt{m}\bar{\tau}_1, \frac{\sqrt{\alpha_1}I}{m}\right\}\right\}$, $T_1 = O\left(\max\left\{\frac{1}{\mu\eta_1}, \frac{\min\{I,B\}}{\mu\beta_1^2}, \frac{I^2\min\{I,B\}}{\mu\tau_1^3m^3}\right\}\right)$. *Define a constant* $\epsilon_1 = \frac{7C_9'(\beta_1+\alpha_1+\bar{\alpha}_1)}{\mu\min\{I,B\}}$ *and* $\epsilon_k = \epsilon_1/2^{k-1}$. *For* $k \geq 2$, *setting* $\beta_k = \alpha_k = \bar{\alpha}_k \leq \frac{\mu\epsilon_k\min\{I,B\}}{21C_9'}$, $\bar{\tau}_k^2 = \frac{C_{\bar{\tau}}I^2}{m^2}\bar{\alpha}_k$, $\tau_k = \sqrt{C_8'}\sqrt{\min\left\{\bar{\tau}_k^2, \frac{I^2\alpha_k}{m^2}\right\}}$, $\eta_k = \sqrt{C_{10}'}\sqrt{\min\left\{m\tau_k^2, m\bar{\tau}_k^2, \frac{\alpha_kI^2}{m^2}\right\}}$, $T_k = O\left(\max\left\{\frac{1}{\mu\eta_k}, \frac{1}{\beta_k}, \frac{1}{\tau_km}, \frac{1}{\bar{\tau}_km}\right\}\right)$, *where and* $C_1' \sim C_{11}'$ *are as used in Theorem B.9, then after* $K = O(\log(\epsilon_1/\epsilon))$ *stages, the output of RE-BSVRB^{v2} satisfies* $\mathbb{E}[F(\mathbf{x}_K) - F(\mathbf{x}^*)] \leq \epsilon$.

Proof. Following from the proof of Theorem B.9, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T\|\nabla F(\mathbf{x}_t)\|^2 + \sum_{t=1}^T\delta_{z,t} + \frac{1}{m}\sum_{t=1}^T\delta_{y,t} + \frac{1}{m}\sum_{t=1}^T\delta_{v,t} + \frac{1}{m}\sum_{t=1}^T\delta_{s,t} + \frac{1}{m}\sum_{t=1}^T\delta_{u,t}\right] \\
&\leq \mathbb{E}\left[\frac{2\Delta}{\eta_0} + C_9'\left[\frac{\delta_{z,1}}{\beta_0} + \frac{\delta_{y,1}}{m\tau_0} + \left(\frac{1}{m\bar{\tau}_0} + \frac{\bar{\tau}_0}{m\beta_0} + \frac{\bar{\tau}_0}{m} + \frac{m\bar{\tau}_0}{\bar{\alpha}_0I^2}\right)\delta_{v,1} + \frac{\delta_{s,1}}{I\alpha_0} + \left(\frac{1}{I\bar{\alpha}_0} + \frac{\bar{\tau}_0^2}{I\bar{\alpha}_0\beta_0} + \frac{\bar{\tau}_0^2}{I\bar{\alpha}_0} + \frac{m^2\bar{\tau}_0^2}{I\bar{\alpha}_0^2I^2}\right)\delta_{u,1}\right.\right. \\
&\quad\left.\left. + T\left(\frac{\beta_t}{\min\{I,B\}} + \frac{\alpha_0}{B} + \frac{\bar{\alpha}_0}{B} + \frac{\bar{\tau}_0^2\bar{\alpha}_0}{\beta_0B} + \frac{\bar{\tau}_0^2\bar{\alpha}_0}{B} + \frac{m^2\bar{\tau}_0^2}{BI^2}\right)\right]\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\frac{2\Delta}{\eta_0} + C_9'\left[\frac{1}{\beta_0}\delta_{z,1} + \frac{1}{m\tau_0}\delta_{y,1} + \frac{1}{m\alpha_0}\delta_{s,1} + \frac{1}{m\bar{\alpha}_0}\delta_{u,1} + \frac{1}{m\bar{\tau}_0}\delta_{v,1} + \frac{(\beta_0+\alpha_0+\bar{\alpha}_0)T}{\min\{I,B\}}\right]\right]
\end{aligned} \tag{57}$$

where in (a) we enlarge the constant C_9' and use the setting $\bar{\alpha}_0 = \beta_0$ and $\bar{\tau}_0^2 = \frac{C_{\bar{\tau}}I^2}{m^2}\bar{\alpha}_0$.

From Theorem B.9, we know that it is required that $\beta_0, \alpha_0, \bar{\alpha}_0 \leq \frac{1}{2}$, $\bar{\tau}_0 = \min\left\{\frac{\lambda}{8L_{\phi v}^2}, \frac{\lambda}{2}, \frac{1}{\lambda}, \frac{1}{2\sqrt{C_6'}}, \frac{\sqrt{C_{\bar{\tau}}}I\bar{\alpha}_0^{1/2}}{m}\right\}$, $\tau_0 = \sqrt{C_8'}\sqrt{\min\left\{\beta_0, \bar{\tau}_0^2, \frac{I^2\alpha_0}{m^2}, \frac{I^2\bar{\alpha}_0}{m^2},\right\}}$, $\eta_0 = \min\left\{\frac{1}{2L_f}, \sqrt{C_{10}'}\sqrt{\min\left\{\beta_0, \tau_0^2, \bar{\tau}_0^2, \frac{\alpha_0I^2}{m^2}, \frac{\bar{\alpha}_0I^2}{m^2}\right\}}\right\}$.

Without loss of generality, set $\beta_0 = \alpha_0 = \bar{\alpha}_0$ and let us assume that $\epsilon_0 = 2\Delta > \frac{7C_9'(\beta_0+\alpha_0+\bar{\alpha}_0)}{\mu}\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right)$. The case that $2\Delta \leq \frac{7C_9'(\beta_0+\alpha_0+\bar{\alpha}_0)}{\mu}\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right)$ can be simply covered by our proof. Then denotes $\epsilon_1 = \frac{7C_9'(\beta_0+\alpha_0+\bar{\alpha}_0)}{\mu}\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right)$, and $\epsilon_k = \epsilon_1/2^{k-1}$.

In the first epoch ($k = 1$), we have initialization such that $F(\mathbf{x}_1) - F(\mathbf{x}^*) \leq \Delta$. In the following, we let the last subscript denote the epoch index. Setting $\eta_1 = \eta_0$, $\beta_1 = \beta_0$, $\alpha_1 = \alpha_0$, $\bar{\alpha}_1 = \bar{\alpha}_0$, $\tau_1 = \tau_0$, $\bar{\tau}_1 = \bar{\tau}_0$, and

$$T_1 = \max\left\{\frac{7\Delta}{\mu\eta_1}, \max\left\{\frac{7}{\mu\beta_1}(\delta_{z,0} + \delta_{s,0} + \delta_{u,0}), \frac{7I^2C_8'}{\mu\tau_1m}\delta_{y,0}, \frac{7I^2C_{\bar{\tau}}}{\mu\bar{\tau}_1m}\delta_{v,0}\right\}\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right)^{-1}\right\}$$

We bound the error of the first stage's output as follows,

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla F(\mathbf{x}_1)\|^2 + \delta_{z,1} + \frac{1}{m}\delta_{y,1} + \frac{1}{m}\delta_{v,1} + \frac{1}{m}\delta_{s,1} + \frac{1}{m}\delta_{u,1}\right] \\
&\leq \frac{2\Delta}{\eta_1 T_1} + \frac{C_9'}{T_1}\left(\frac{1}{\beta_1}\delta_{z,0} + \frac{1}{\tau_1 m}\delta_{y,0} + \frac{1}{\tau_1 m}\delta_{v,0} + \frac{1}{\alpha_1 m}\delta_{s,0} + \frac{1}{\bar{\alpha}_1 m}\delta_{u,0}\right) + \frac{C_9'(\beta_1 + \alpha_1 + \bar{\alpha}_1)}{\min\{I, B\}} \\
&\leq \mu\epsilon_1
\end{aligned} \tag{58}$$

where the first inequality uses (57) and the fact that the output of each epoch is randomly sampled from all iterations, and the last line uses the choice of $\eta_1, \beta_1, \alpha_1, \bar{\alpha}_1, \tau_1, \bar{\tau}_1, T_1, \epsilon_1$. If follows that

$$\mathbb{E}[F(\mathbf{x}_1) - \mathcal{F}(\mathbf{x}^*)] \leq \frac{1}{2\mu}\mathbb{E}[\|\nabla F(\mathbf{x}_1)\|^2] \leq \frac{\epsilon_1}{2}. \tag{59}$$

Starting from the second stage, we will prove by induction. Suppose we are at k-th stage. Assuming that the output of $(k-1)$-the stage satisfies that $\mathbb{E}[F(\mathbf{x}_{k-1}) - F(\mathbf{x}^*)] \leq \epsilon_{k-1}$ and $\mathbb{E}\left[\delta_{z,k-1} + \frac{\delta_{y,k-1}}{m} + \frac{\delta_{v,k-1}}{m} + \frac{\delta_{s,k-1}}{m} + \frac{\delta_{u,k-1}}{m}\right] \leq \mu\epsilon_{k-1}$, and setting $\beta_k = \alpha_k = \bar{\alpha}_k \leq \frac{\mu\epsilon_k}{21C_9'}\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right)^{-1}$, $\bar{\tau}_k^2 = \frac{C_{\bar{\tau}} I^2}{m^2}\bar{\alpha}_k$, $\tau_k^2 = C_8' \min\left\{\beta_k, \bar{\tau}_k^2, \frac{I^2\alpha_k}{m^2}, \frac{I^2\bar{\alpha}_k}{m^2}, \right\}$, $\eta_k = C_{10}' \min\left\{\beta_k, \tau_k^2, \bar{\tau}_k^2, \frac{\alpha_k I^2}{m^2}, \frac{\bar{\alpha}_k I^2}{m^2}\right\}$, $T_k = \max\left\{\frac{28}{\mu\eta_k}, \frac{7C_9'}{\beta_k}, \frac{7C_9'}{\tau_k}, \frac{7C_9'}{\bar{\tau}_k}\right\}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla F(\mathbf{x}_k)\|^2 + \delta_{z,k} + \frac{1}{m}\delta_{y,k} + \frac{1}{m}\delta_{v,k} + \frac{1}{m}\delta_{s,k} + \frac{1}{m}\delta_{u,k}\right] \\
&\leq \mathbb{E}\left[\frac{2(F(\mathbf{x}_{k-1}) - F(\mathbf{x}^*))}{\eta_k T_k} + \frac{C_9'}{T_k}\left(\frac{1}{\beta_k}\delta_{z,k-1} + \frac{1}{\tau_k m}\delta_{y,k-1} + \frac{1}{\tau_k m}\delta_{v,k-1} + \frac{1}{\alpha_k m}\delta_{s,k-1} + \frac{1}{\bar{\alpha}_k m}\delta_{u,k-1}\right)\right. \\
&\quad \left. + \frac{C_9'(\beta_k + \alpha_k + \bar{\alpha}_k)}{\min\{I, B\}}\right] \\
&\leq \mathbb{E}\left[\frac{2\epsilon_{k-1}}{\eta_k T_k} + \frac{C_9'\mu\epsilon_{k-1}}{T_k}\left(\frac{1}{\beta_k} + \frac{1}{\tau_k} + \frac{1}{\bar{\tau}_k} + \frac{1}{\alpha_k} + \frac{1}{\bar{\alpha}_k}\right) + \frac{C_9'(\beta_k + \alpha_k + \bar{\alpha}_k)}{\min\{I, B\}}\right] \\
&\leq \mu\epsilon_k
\end{aligned} \tag{60}$$

It follows that

$$\mathbb{E}[F(\mathbf{x}_k) - F(\mathbf{x}^*)] \leq \frac{1}{2\mu}\mathbb{E}[\|\nabla F(\mathbf{x}_k)\|^2] \leq \frac{\epsilon_k}{2}. \tag{61}$$

Thus, after $K = 1 + \log_2(\epsilon_1/\epsilon) \leq \log_2(\epsilon_0/\epsilon)$ stages, $\mathbb{E}[F(\mathbf{x}_k) - F(\mathbf{x}^*)] \leq \epsilon$.

□

D. Proof of Lemmas

D.1. Proof of Lemma B.2

Proof. Due the smoothness of F, we can prove that under $\eta_t L_F \leq 1/2$

$$\begin{aligned}
F(\mathbf{x}_{t+1}) &\leq F(\mathbf{x}_t) + \nabla F(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L_F}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&= F(\mathbf{x}_t) - \eta_t \nabla F(\mathbf{x}_t)^\top \mathbf{z}_{t+1} + \frac{L_F \eta_t^2}{2}\|\mathbf{z}_{t+1}\|^2 \\
&= F(\mathbf{x}_t) + \frac{\eta_t}{2}\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2 - \frac{\eta_t}{2}\|\nabla F(\mathbf{x}_t)\|^2 + (\frac{L_F\eta_t^2}{2} - \frac{\eta_t}{2})\|\mathbf{z}_{t+1}\|^2 \\
&\leq F(\mathbf{x}_t) + \frac{\eta_t}{2}\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2 - \frac{\eta_t}{2}\|\nabla F(\mathbf{x}_t)\|^2 - \frac{\eta_t}{4}\|\mathbf{z}_{t+1}\|^2
\end{aligned}$$

□

D.2. Proof of Lemma B.4

Proof. Consider the updates

$$\mathbf{h}_{i,t+1} = \begin{cases} \Pi_\Omega \left[(1-\alpha)\mathbf{h}_{i,t} + \alpha h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) + \gamma[h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t - h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t)]\right] & \text{if } i \in I_t \\ \mathbf{h}_{i,t} & \text{o.w.} \end{cases}$$

Define

$$\begin{aligned} \tilde{h}_{i,t} &= \Pi_\Omega \left[(1-\alpha)\mathbf{h}_{i,t} + \alpha h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) + \gamma[h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t - h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t)]\right] \\ \bar{h}_{i,t} &= (1-\alpha)\mathbf{h}_{i,t} + \alpha h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) + \gamma[h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t - h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t)] \end{aligned}$$

We have

$$\begin{aligned} &\mathbb{E}_t\left[\|\mathbf{h}_{i,t+1} - h_i(\mathbf{e}_{i,t})\|^2\right] \\ &= \mathbb{E}_{I_t}\mathbb{E}_{\mathcal{B}_i^t}\left[\|\mathbf{h}_{i,t+1} - h_i(\mathbf{e}_{i,t})\|^2\right] \\ &= \frac{I}{m}\mathbb{E}_{\mathcal{B}_i^t}\left[\|\tilde{h}_{i,t} - h_i(\mathbf{e}_{i,t})\|^2\right] + \left(1-\frac{I}{m}\right)\|\mathbf{h}_{i,t} - h_i(\mathbf{e}_{i,t})\|^2 \\ &= \frac{I}{m}\mathbb{E}_{\mathcal{B}_i^t}\left[\|\tilde{h}_{i,t} - h_i(\mathbf{e}_{i,t})\|^2\right] + \left(1-\frac{I}{m}\right)\|\mathbf{h}_{i,t} - h_i(\mathbf{e}_{i,t-1}) + h_i(\mathbf{e}_{i,t-1}) - h_i(\mathbf{e}_{i,t})\|^2 \\ &= \frac{I}{m}\mathbb{E}_{\mathcal{B}_i^t}\left[\|\tilde{h}_{i,t} - h_i(\mathbf{e}_{i,t})\|^2\right] + \left(1-\frac{I}{m}\right)\|\mathbf{h}_{i,t} - h_i(\mathbf{e}_{i,t-1})\|^2 + \left(1-\frac{I}{m}\right)\|h_i(\mathbf{e}_{i,t-1}) - h_i(\mathbf{e}_{i,t})\|^2 \\ &\quad + \underbrace{2\left(1-\frac{I}{m}\right)\left\langle \mathbf{h}_{i,t} - h_i(\mathbf{e}_{i,t-1}), h_i(\mathbf{e}_{i,t-1}) - h_i(\mathbf{e}_{i,t})\right\rangle}_{ⓐ} \end{aligned} \tag{62}$$

It follows from the non-expansive property of projection that

$$\begin{aligned} &\mathbb{E}_{\mathcal{B}_i^t}\left[\|\tilde{h}_{i,t} - h_i(\mathbf{e}_{i,t})\|^2\right] \le \mathbb{E}_{\mathcal{B}_i^t}\left[\|\bar{h}_{i,t} - h_i(\mathbf{e}_{i,t})\|^2\right] \\ &= \mathbb{E}_{\mathcal{B}_i^t}\left[\left\|(1-\alpha)\mathbf{h}_{i,t} + \alpha h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) + \gamma[h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t - h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t)] - h_i(\mathbf{e}_{i,t})\right\|^2\right] \\ &= \mathbb{E}_{\mathcal{B}_i^t}\Bigg[\Big\|(1-\alpha)[\mathbf{h}_{i,t} - h_i(\mathbf{e}_{i,t-1})] + (1-\alpha)[h_i(\mathbf{e}_{i,t-1}) - h_i(\mathbf{e}_{i,t})] \\ &\quad + \alpha[h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) - h_i(\mathbf{e}_{i,t})] + \gamma[h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t - h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t)]\Big\|^2\Bigg] \\ &\overset{(a)}{=} \mathbb{E}_{\mathcal{B}_i^t}\Bigg[\Big\|(1-\alpha)[\mathbf{h}_{i,t} - h_i(\mathbf{e}_{i,t-1})] + (1-\alpha)[h_i(\mathbf{e}_{i,t-1}) - h_i(\mathbf{e}_{i,t})] \\ &\quad + \gamma[h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) - h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t)]\Big\|^2\Bigg] + \alpha^2\mathbb{E}_{\mathcal{B}_i^t}\left[\left\|h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) - h_i(\mathbf{e}_{i,t})\right\|^2\right] \\ &\quad + 2\gamma\alpha\mathbb{E}_{\mathcal{B}_i^t}\left[\langle h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) - h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t), h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) - h_i(\mathbf{e}_{i,t})\rangle\right] \\ &\overset{(b)}{=} (1-\alpha)^2\left\|\mathbf{h}_{i,t} - h_i(\mathbf{e}_{i,t-1})\right\|^2 \\ &\quad + \mathbb{E}_{\mathcal{B}_i^t}\left[\left\|(1-\alpha)[h_i(\mathbf{e}_{i,t-1}) - h_i(\mathbf{e}_{i,t})] + \gamma[h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) - h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t)]\right\|^2\right] \\ &\quad + \underbrace{2(1-\alpha)(1-\alpha-\gamma)\left\langle \mathbf{h}_{i,t} - h_i(\mathbf{e}_{i,t-1}), h_i(\mathbf{e}_{i,t-1}) - h_i(\mathbf{e}_{i,t})\right\rangle}_{ⓑ} \\ &\quad + \frac{\alpha^2\sigma^2}{B} + 2\gamma\alpha\mathbb{E}_{\mathcal{B}_i^t}\left[\langle h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) - h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t), h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) - h_i(\mathbf{e}_{i,t})\rangle\right] \end{aligned} \tag{63}$$

where (a) follows from $\mathbb{E}_{\mathcal{B}_i^t}[h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) - h_i(\mathbf{e}_{i,t})] = 0$, (b) follows from $\mathbb{E}_{\mathcal{B}_i^t}[h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t) - h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t)] = h_i(\mathbf{e}_{i,t}) - h_i(\mathbf{e}_{i,t-1})$.

Combining inequalities (62) and (63) gives

$$
\begin{aligned}
&\mathbb{E}_t\left[\|\mathbf{h}_{i,t+1}-h_i(\mathbf{e}_{i,t})\|^2\right]\\
&=\left(1-\frac{I}{m}+\frac{(1-\alpha)^2 I}{m}\right)\left\|\mathbf{h}_{i,t}-h_i(\mathbf{e}_{i,t-1})\right\|^2\\
&\quad+\frac{I}{m}\mathbb{E}_{\mathcal{B}_i^t}\left[\left\|(1-\alpha)[h_i(\mathbf{e}_{i,t-1})-h_i(\mathbf{e}_{i,t})]+\gamma[h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t)]\right\|^2\right]\\
&\quad+\frac{I}{m}ⓑ+\frac{\alpha^2 I\sigma^2}{Bm}+\left(1-\frac{I}{m}\right)\left\|h_i(\mathbf{e}_{i,t-1})-h_i(\mathbf{e}_{i,t})\right\|^2+ⓐ\\
&\quad+\frac{2\gamma\alpha I}{m}\mathbb{E}_{\mathcal{B}_i^t}\left[\langle h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t),h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t})\rangle\right]\\
&\overset{(a)}{=}\left(1-\frac{I}{m}+\frac{(1-\alpha)^2 I}{m}\right)\left\|\mathbf{h}_{i,t}-h_i(\mathbf{e}_{i,t-1})\right\|^2\\
&\quad+\frac{I}{m}\mathbb{E}_{\mathcal{B}_i^t}\left[\left\|(1-\alpha)[h_i(\mathbf{e}_{i,t-1})-h_i(\mathbf{e}_{i,t})]+\gamma[h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t)]\right\|^2\right]\\
&\quad+\frac{\alpha^2 I\sigma^2}{Bm}+\left(1-\frac{I}{m}\right)\left\|h_i(\mathbf{e}_{i,t-1})-h_i(\mathbf{e}_{i,t})\right\|^2\\
&\quad+\frac{2\gamma\alpha I}{m}\mathbb{E}_{\mathcal{B}_i^t}\left[\langle h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t),h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t})\rangle\right]\\
&=\left(1-\frac{I}{m}+\frac{(1-\alpha)^2 I}{m}\right)\left\|\mathbf{h}_{i,t}-h_i(\mathbf{e}_{i,t-1})\right\|^2+\frac{(1-\alpha)^2 I}{m}\left\|h_i(\mathbf{e}_{i,t-1})-h_i(\mathbf{e}_{i,t})\right\|^2\\
&\quad+\frac{\gamma^2 I}{m}\mathbb{E}_{\mathcal{B}_i^t}\left[\left\|h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t)\right\|^2\right]-\frac{2(1-\alpha)\gamma I}{m}\left\|h_i(\mathbf{e}_{i,t-1})-h_i(\mathbf{e}_{i,t})\right\|^2\\
&\quad+\frac{\alpha^2 I\sigma^2}{Bm}+\left(1-\frac{I}{m}\right)\left\|h_i(\mathbf{e}_{i,t-1})-h_i(\mathbf{e}_{i,t})\right\|^2\\
&\quad+\frac{2\gamma\alpha I}{m}\mathbb{E}_{\mathcal{B}_i^t}\left[\langle h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t),h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t})\rangle\right]\\
&\overset{(b)}{=}\left(1-\frac{\alpha I}{m}\right)\left\|\mathbf{h}_{i,t}-h_i(\mathbf{e}_{i,t-1})\right\|^2+\frac{4mL^2}{I}\|\mathbf{e}_{i,t-1}-\mathbf{e}_{i,t}\|^2+\frac{\alpha^2 I\sigma^2}{Bm}\\
&\quad+\frac{2\gamma\alpha I}{m}\mathbb{E}_{\mathcal{B}_i^t}\left[\langle h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t),h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t})\rangle\right]\\
&\overset{(c)}{=}\left(1-\frac{\alpha I}{m}\right)\left\|\mathbf{h}_{i,t}-h_i(\mathbf{e}_{i,t-1})\right\|^2+\frac{8mL^2}{I}\|\mathbf{e}_{i,t-1}-\mathbf{e}_{i,t}\|^2+\frac{2\alpha^2 I\sigma^2}{Bm}
\end{aligned}
\tag{64}
$$

where (a) is due to ⓐ $+\frac{I}{m}$ⓑ $=0$, which follows from the setting $\gamma=\frac{m-\alpha I}{(1-\alpha)I}$, (b) is due to $1-\frac{I}{m}+\frac{(1-\alpha)^2 I}{m}\le\frac{2(1-\alpha)\gamma I}{m}$ and $\gamma\le\frac{2m}{I}$, which follows from $\alpha\le\frac{1}{2}$, (c) is due to

$$
\begin{aligned}
&\frac{2\gamma\alpha I}{m}\mathbb{E}_{\mathcal{B}_i^t}\left[\langle h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t),h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t})\rangle\right]\\
&\le\frac{I}{m}\mathbb{E}_{\mathcal{B}_i^t}\left[\gamma^2\|h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t-1};\mathcal{B}_i^t)\|^2+\alpha^2\|h_i(\mathbf{e}_{i,t};\mathcal{B}_i^t)-h_i(\mathbf{e}_{i,t})\|^2\right]\\
&\le\frac{4mL^2}{I}\|\mathbf{e}_{i,t-1}-\mathbf{e}_{i,t}\|^2+\frac{\alpha^2 I\sigma^2}{mB}
\end{aligned}
$$

Then by taking expectation over all randomness and summing over $i=1,\ldots,m$, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^m\|\mathbf{h}_{i,t+1}-h_i(\mathbf{e}_{i,t})\|^2\right]\\
&\le\left(1-\frac{\alpha I}{m}\right)\mathbb{E}\Big[\sum_{i=1}^m\|\mathbf{h}_{i,t}-h_i(\mathbf{e}_{i,t-1})\|^2\Big]+\frac{8mL^2}{I}\mathbb{E}\Big[\sum_{i=1}^m\|\mathbf{e}_{i,t-1}-\mathbf{e}_{i,t}\|^2\Big]+\frac{2\alpha^2 I\sigma^2}{B}
\end{aligned}
\tag{65}
$$

□

D.3. Proof of Lemma B.6

Proof.

$$
\begin{aligned}
&\|\Delta_t - \nabla F(\mathbf{x}_t)\|^2 \\
&= \Bigg\| \frac{1}{m}\sum_{i=1}^m \nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t})\mathbb{E}_t[[H_{i,t}]^{-1}]\nabla_y f_i(\mathbf{x}_t, \mathbf{y}_{i,t}) \\
&\quad - \frac{1}{m}\sum_{i=1}^m \nabla_x f_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t)) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))[\nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\nabla_y f_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t)) \Bigg\|^2 \\
&\le \frac{1}{m}\sum_{i=1}^m 2\left\|\nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \nabla_x f_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))\right\|^2 \\
&\quad + 6\left\|\nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t})[\mathbb{E}_t[[H_{i,t}]^{-1}] - [\nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}]\nabla_y f_i(\mathbf{x}_t, \mathbf{y}_{i,t})\right\|^2 \\
&\quad + 6\left\|[\nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))][\nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\nabla_y f_i(\mathbf{x}_t, \mathbf{y}_{i,t})\right\|^2 \\
&\quad + 6\left\|\nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))[\nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}[\nabla_y f_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \nabla_y f_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]\right\|^2 \\
&\le \frac{1}{m}\sum_{i=1}^m \left(2L_{fx}^2 + \frac{L_{gxy}^2 C_{fy}^2}{\lambda^2} + \frac{6C_{gxy}^2 L_{fy}^2}{\lambda^2}\right)\|\mathbf{y}_{i,t} - \mathbf{y}_i(\mathbf{x}_t)\|^2 + 6C_{gxy}^2 C_{fy}^2 \|\mathbb{E}_t[[H_{i,t}]^{-1}] - [\nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))]^{-1}\|^2 \\
&\overset{(a)}{\le} \frac{1}{m}\sum_{i=1}^m C_1\|\mathbf{y}_{i,t} - \mathbf{y}_i(\mathbf{x}_t)\|^2 + C_2\|H_{i,t} - \nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t}))\|^2
\end{aligned} \tag{66}
$$

where $C_1 := \left(2L_{fx}^2 + \frac{L_{gxy}^2 C_{fy}^2}{\lambda^2} + \frac{6C_{gxy}^2 L_{fy}^2}{\lambda^2} + \frac{12C_{gxy}^2 C_{fy}^2 L_{gyy}^2}{\lambda^4}\right)$, $C_2 := \frac{12C_{gxy}^2 C_{fy}^2}{\lambda^4}$, and (a) uses the fact that $[H_{i,t}]^{-1}$ is irrelevant to the randomness at iteration t, which means $[H_{i,t}]^{-1} = E_t[[H_{i,t}]^{-1}]$, and the Lipschitz continuity of $\nabla_{yy}^2 g_i(\mathbf{x}, \mathbf{y}_i)$. □

D.4. Proof of Lemma B.7

Proof.

$$
\begin{aligned}
&\mathbb{E}_t[\|\mathbf{z}_{t+1} - \Delta_t\|^2] \\
&= \mathbb{E}_t\left[\left\|(1-\beta_t)(\mathbf{z}_t - \Delta_{t-1}) + (1-\beta_t)(\Delta_{t-1} - \widetilde{G}_t) + G_t - \Delta_t\right\|^2\right] \\
&= (1-\beta_t)^2\|\mathbf{z}_t - \Delta_{t-1}\|^2 + 2(1-\beta_t)^2\mathbb{E}_t\left[\|\Delta_{t-1} - \widetilde{G}_t + G_t - \Delta_t\|^2\right] + 2\beta_t^2\mathbb{E}_t\left[\|G_t - \Delta_t\|^2\right] \\
&\overset{(a)}{\le} (1-\beta_t)^2\|\mathbf{z}_t - \Delta_{t-1}\|^2 + 2(1-\beta_t)^2\mathbb{E}_t\left[\|G_t - \widetilde{G}_t\|^2\right] + 2\beta_t^2\mathbb{E}_t\left[\|G_t - \Delta_t\|^2\right]
\end{aligned} \tag{67}
$$

where (a) use the standard inequality $\mathbb{E}[\|a-\mathbb{E}[a]\|^2] \leq \mathbb{E}[\|a\|^2]$, and $\mathbb{E}_t[G_t] = \Delta_t$, $\mathbb{E}_t[\widetilde{G}_t] = \Delta_{t-1}$. We further bound the last two terms as following

$$
\begin{aligned}
&\mathbb{E}_t\left[\|G_t - \Delta_t\|^2\right] \\
&\leq \mathbb{E}_t\Bigg[\Bigg\|\frac{1}{I}\sum_{i\in\mathcal{I}_t}\Big(\nabla_x f_i(\mathbf{x}_t,\mathbf{y}_{i,t};\mathcal{B}_i^t) - \nabla_{xy}^2 g_i(\mathbf{x}_t,\mathbf{y}_{i,t};\widetilde{\mathcal{B}}_i^t)[H_{i,t}]^{-1}\nabla_y f_i(\mathbf{x}_t,\mathbf{y}_{i,t};\mathcal{B}_i^t)\Big) \\
&\quad - \frac{1}{m}\sum_{i=1}^m\left(\nabla_x f_i(\mathbf{x}_t,\mathbf{y}_{i,t}) - \nabla_{xy}^2 g_i(\mathbf{x}_t,\mathbf{y}_{i,t})[H_{i,t}]^{-1}\nabla_y f_i(\mathbf{x}_t,\mathbf{y}_{i,t})\right)\Bigg\|^2\Bigg] \\
&\leq \mathbb{E}_t\Bigg[2\Bigg\|\frac{1}{I}\sum_{i\in\mathcal{I}_t}\Big(\nabla_x f_i(\mathbf{x}_t,\mathbf{y}_{i,t};\mathcal{B}_i^t) - \nabla_{xy}^2 g_i(\mathbf{x}_t,\mathbf{y}_{i,t};\widetilde{\mathcal{B}}_i^t)[H_{i,t}]^{-1}\nabla_y f_i(\mathbf{x}_t,\mathbf{y}_{i,t};\mathcal{B}_i^t)\Big) \\
&\quad - \frac{1}{m}\sum_{i=1}^m\Big(\nabla_x f_i(\mathbf{x}_t,\mathbf{y}_{i,t};\mathcal{B}_i^t) - \nabla_{xy}^2 g_i(\mathbf{x}_t,\mathbf{y}_{i,t};\widetilde{\mathcal{B}}_i^t)[H_{i,t}]^{-1}\nabla_y f_i(\mathbf{x}_t,\mathbf{y}_{i,t};\mathcal{B}_i^t)\Big)\Bigg\|^2 \\
&\quad + 2\Bigg\|\frac{1}{m}\sum_{i=1}^m\Big(\nabla_x f_i(\mathbf{x}_t,\mathbf{y}_{i,t};\mathcal{B}_i^t) - \nabla_{xy}^2 g_i(\mathbf{x}_t,\mathbf{y}_{i,t};\widetilde{\mathcal{B}}_i^t)[H_{i,t}]^{-1}\nabla_y f_i(\mathbf{x}_t,\mathbf{y}_{i,t};\mathcal{B}_i^t)\Big) \\
&\quad - \frac{1}{m}\sum_{i=1}^m\left(\nabla_x f_i(\mathbf{x}_t,\mathbf{y}_{i,t}) - \nabla_{xy}^2 g_i(\mathbf{x}_t,\mathbf{y}_{i,t})[H_{i,t}]^{-1}\nabla_y f_i(\mathbf{x}_t,\mathbf{y}_{i,t})\right)\Bigg\|^2\Bigg] \\
&\leq \frac{8(2C_{fx}^2 + \frac{2C_{gxy}^2C_{fy}^2}{\lambda^2})}{I} + \frac{4\sigma^2}{B} + \frac{8(\frac{C_{gxy}^2+C_{fy}^2}{\lambda^2})\sigma^2}{B} =: C_5\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right),
\end{aligned} \tag{68}
$$

where $C_5 = \max\{8(2C_{fx}^2 + \frac{2C_{gxy}^2C_{fy}^2}{\lambda^2}), 4\sigma^2 + 8(\frac{C_{gxy}^2+C_{fy}^2}{\lambda^2})\sigma^2\}$, and

$$
\begin{aligned}
&\mathbb{E}_t\left[\|G_t - \widetilde{G}_t\|^2\right] \\
&= \mathbb{E}_t\Bigg[\Bigg\|\frac{1}{I}\sum_{i\in\mathcal{I}_t}\Big(\nabla_x f_i(\mathbf{x}_t,\mathbf{y}_{i,t};\mathcal{B}_i^t) - \nabla_{xy}^2 g_i(\mathbf{x}_t,\mathbf{y}_{i,t};\widetilde{\mathcal{B}}_i^t)[H_{i,t}]^{-1}\nabla_y f_i(\mathbf{x}_t,\mathbf{y}_{i,t};\mathcal{B}_i^t)\Big) \\
&\quad - \Big(\nabla_x f_i(\mathbf{x}_{t-1},\mathbf{y}_{i,t-1};\mathcal{B}_i^t) - \nabla_{xy}^2 g_i(\mathbf{x}_{t-1},\mathbf{y}_{i,t-1};\widetilde{\mathcal{B}}_i^t)[H_{i,t-1}]^{-1}\nabla_y f_i(\mathbf{x}_{t-1},\mathbf{y}_{i,t-1};\mathcal{B}_i^t)\Big)\Bigg\|^2\Bigg] \\
&\leq \frac{1}{m}\sum_{i=1}^m(2L_{fx}^2 + \frac{6L_{gxy}^2C_{fy}^2}{\lambda^2} + \frac{6C_{gxy}^2L_{fy}^2}{\lambda^2})(\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2 + \|\mathbf{y}_{i,t}-\mathbf{y}_{i,t-1}\|^2) + \frac{6C_{gxy}^2C_{fy}^2}{\lambda^4}\|H_{i,t}-H_{i,t-1}\|^2 \\
&=: C_3\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2 + \frac{C_3}{m}\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2 + \frac{C_4}{m}\|H_t-H_{t-1}\|^2.
\end{aligned} \tag{69}
$$

Then we have

$$
\begin{aligned}
&\mathbb{E}_t[\|\mathbf{z}_{t+1}-\Delta_t\|^2] \\
&\leq (1-\beta_t)^2\|\mathbf{z}_t-\Delta_{t-1}\|^2 + 2(1-\beta_t)^2\left(C_3\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2 + \frac{C_3}{m}\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2 + \frac{C_4}{m}\|\mathbf{v}_t-\mathbf{v}_{t-1}\|^2\right) \\
&\quad + 2\beta_t^2C_5\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right) \\
&\leq (1-\beta_t)\|\mathbf{z}_t-\Delta_{t-1}\|^2 + 2C_3\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2 + \frac{2C_3}{m}\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2 + \frac{2C_4}{m}\|H_t-H_{t-1}\|^2 + 2\beta_t^2C_5\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right)
\end{aligned} \tag{70}
$$

□

D.5. Proof of Lemma B.8

Proof. By Lemma6 in (Jiang et al., 2022), we have, for $\bar{\alpha}_{t+1} \leq 1/2$,

$$\begin{aligned}\|H_{t+1} - H_t\|^2 \leq \frac{2I\bar{\alpha}_{t+1}^2\sigma^2}{B} + \frac{4I\bar{\alpha}_{t+1}^2}{m}\mathbb{E}\left[\sum_{i=1}^m \|H_{i,t} - \nabla_{yy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2\right] \\ + \frac{9m^2L_{gyy}^2}{I}\mathbb{E}[m\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2]\end{aligned} \tag{71}$$

□

D.6. Proof of Lemma B.11

$$\begin{aligned}&\|\Delta_t - \nabla F(\mathbf{x}_t)\|^2 \\ &= \left\|\frac{1}{m}\sum_{i=1}^m \left(\nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t})\mathbf{v}_{i,t}\right) - \frac{1}{m}\sum_{i=1}^m \left(\nabla_x f_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t)) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_i(\mathbf{x}_t))\mathbf{v}_i(\mathbf{x}_t)\right)\right\|^2 \\ &\leq \frac{1}{m}\sum_{i=1}^m (2L_{fx}^2 + \frac{4L_{gxy}^2C_{fy}^2}{\lambda^2})\|\mathbf{y}_{i,t} - \mathbf{y}_i(\mathbf{x}_t)\|^2 + 4C_{gxy}^2\|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t)\|^2 \\ &\leq \frac{1}{m}\sum_{i=1}^m (2L_{fx}^2 + \frac{4L_{gxy}^2C_{fy}^2}{\lambda^2})\|\mathbf{y}_{i,t} - \mathbf{y}_i(\mathbf{x}_t)\|^2 + 8C_{gxy}^2\|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2 + 8C_{gxy}^2L_v^2\|\mathbf{y}_{i,t} - \mathbf{y}_i(\mathbf{x}_t)\|^2 \\ &=: \frac{C_1'}{m}\|\mathbf{y}_t - \mathbf{y}(\mathbf{x}_t)\|^2 + \frac{C_2'}{m}\|\mathbf{v}_t - \mathbf{v}(\mathbf{x}_t, \mathbf{y}_t)\|^2\end{aligned} \tag{72}$$

D.7. Proof of Lemma B.12

Proof.

$$\begin{aligned}&\mathbb{E}_t[\|\mathbf{z}_{t+1} - \Delta_t\|^2] \\ &= \mathbb{E}_t\left[\left\|(1-\beta_t)(\mathbf{z}_t - \Delta_{t-1}) + (1-\beta_t)(\Delta_{t-1} - \widetilde{G}_t) + G_t - \Delta_t\right\|^2\right] \\ &= (1-\beta_t)^2\|\mathbf{z}_t - \Delta_{t-1}\|^2 + 2(1-\beta_t)^2\mathbb{E}_t\left[\|\Delta_{t-1} - \widetilde{G}_t + G_t - \Delta_t\|^2\right] + 2\beta_t^2\mathbb{E}_t\left[\|G_t - \Delta_t\|^2\right] \\ &\overset{(a)}{\leq} (1-\beta_t)^2\|\mathbf{z}_t - \Delta_{t-1}\|^2 + 2(1-\beta_t)^2\mathbb{E}_t\left[\|G_t - \widetilde{G}_t\|^2\right] + 2\beta_t^2\mathbb{E}_t\left[\|G_t - \Delta_t\|^2\right]\end{aligned} \tag{73}$$

where (a) use the standard inequality $\mathbb{E}[\|a - \mathbb{E}[a]\|^2] \leq \mathbb{E}[\|a\|^2]$, and $\mathbb{E}_t[G_t] = \Delta_t$, $\mathbb{E}_t[\widetilde{G}_t] = \Delta_{t-1}$. We further bound the last two terms as following

$$\begin{aligned}&\mathbb{E}_t\left[\|G_t - \Delta_t\|^2\right] \\ &\leq \mathbb{E}_t\left[\left\|\frac{1}{I}\sum_{i\in\mathcal{I}_t}\left(\nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t)\mathbf{v}_{i,t}\right) - \frac{1}{m}\sum_{i=1}^m \left(\nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t})\mathbf{v}_{i,t}\right)\right\|^2\right] \\ &\leq \mathbb{E}_t\left[2\left\|\frac{1}{I}\sum_{i\in\mathcal{I}_t}\left(\nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t)\mathbf{v}_{i,t}\right)\right.\right. \\ &\quad \left.\left. - \frac{1}{m}\sum_{i=1}^m \left(\nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t)\mathbf{v}_{i,t}\right)\right\|^2\right. \\ &\quad \left. + 2\left\|\frac{1}{m}\sum_{i=1}^m \left(\nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t)\mathbf{v}_{i,t}\right) - \frac{1}{m}\sum_{i=1}^m \left(\nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t})\mathbf{v}_{i,t}\right)\right\|^2\right] \\ &\leq \frac{8(2C_{fx}^2 + 2C_{gxy}^2\mathcal{V}^2)}{I} + \frac{4\sigma^2}{B} + \frac{4\sigma^2\mathcal{V}^2}{B} \leq C_5'\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right),\end{aligned} \tag{74}$$

where $C_5' = \max\{8(2C_{fx}^2 + 2C_{gxy}^2\mathcal{V}^2, 4\sigma^2 + 4\sigma^2\mathcal{V}^2)\}$, and

$$\begin{aligned}
&\mathbb{E}_t\left[\|G_t - \widetilde{G}_t\|^2\right] \\
&= \mathbb{E}_t\left[\left\|\frac{1}{I}\sum_{i\in\mathcal{I}_t}\left(\nabla_x f_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \mathcal{B}_i^t) - \nabla_{xy}^2 g_i(\mathbf{x}_t, \mathbf{y}_{i,t}; \widetilde{\mathcal{B}}_i^t)\mathbf{v}_{i,t}\right)\right.\right. \\
&\quad \left.\left. - \left(\nabla_x f_i(\mathbf{x}_{t-1}, \mathbf{y}_{i,t-1}; \mathcal{B}_i^t) - \nabla_{xy}^2 g_i(\mathbf{x}_{t-1}, \mathbf{y}_{i,t-1}; \widetilde{\mathcal{B}}_i^t)\mathbf{v}_{i,t-1}\right)\right\|^2\right] \qquad (75) \\
&\le \frac{1}{m}\sum_{i=1}^m (2L_{fx}^2 + \frac{4L_{gxy}^2 C_{fy}^2}{\lambda^2})(\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \|\mathbf{y}_{i,t} - \mathbf{y}_{i,t-1}\|^2) + 4C_{gxy}^2\|\mathbf{v}_{i,t} - \mathbf{v}_{i,t-1}\|^2 \\
&=: C_3'\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \frac{C_3'}{m}\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2 + \frac{C_4'}{m}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathbb{E}_t[\|\mathbf{z}_{t+1} - \Delta_t\|^2] \\
&\le (1-\beta_t)^2\|\mathbf{z}_t - \Delta_{t-1}\|^2 + 2(1-\beta_t)^2\left(C_3'\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \frac{C_3'}{m}\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2 + \frac{C_4'}{m}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2\right) \\
&\quad + 2\beta_t^2 C_5'\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right) \\
&\le (1-\beta_t)\|\mathbf{z}_t - \Delta_{t-1}\|^2 + 2C_3'\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \frac{2C_3'}{m}\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2 + \frac{2C_4'}{m}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 + 2\beta_t^2 C_5'\left(\frac{\mathbb{I}(I<m)}{I} + \frac{1}{B}\right)
\end{aligned} \qquad (76)$$

□

D.8. Proof of Lemma B.13

Proof. Consider updates $\mathbf{v}_{i,t+1} = \Pi_{\mathcal{V}}[\mathbf{v}_{i,t} - \bar{\tau}_t\mathbf{u}_{i,t}]$. Note that $\mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}) = \Pi_{\mathcal{V}}[\mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}), \mathbf{x}_t, \mathbf{y}_{i,t})]$

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{v}_{i,t+1} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2] \\
&= \mathbb{E}\left[\|\Pi_{\mathcal{V}}[\mathbf{v}_{i,t} - \bar{\tau}_t\mathbf{u}_{i,t}] - \Pi_{\mathcal{V}}[\mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}), \mathbf{x}_t, \mathbf{y}_{i,t})]\|^2\right] \\
&\le \mathbb{E}\left[\|\mathbf{v}_{i,t} - \bar{\tau}_t\mathbf{u}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}) + \bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}), \mathbf{x}_t, \mathbf{y}_{i,t})\|^2\right] \\
&\le \mathbb{E}\left[\|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \bar{\tau}_t\mathbf{u}_{i,t} + \bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t}) - \bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t}) + \bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}), \mathbf{x}_t, \mathbf{y}_{i,t})\|^2\right] \\
&\le \mathbb{E}\big[\|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2 + \| - \bar{\tau}_t\mathbf{u}_{i,t} + \bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t}) - \bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t}) + \bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}), \mathbf{x}_t, \mathbf{y}_{i,t})\|^2 \\
&\quad + \langle\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}), -\bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t}) + \bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}), \mathbf{x}_t, \mathbf{y}_{i,t})\rangle \\
&\quad + \langle\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}), -\bar{\tau}_t\mathbf{u}_{i,t} + \bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t})\rangle\big] \\
&\overset{(a)}{\le} \mathbb{E}\big[\|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2 + 2\bar{\tau}_t^2 L_{\phi v}^2\|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2 - \lambda\bar{\tau}_t\|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2 \\
&\quad + 2\bar{\tau}_t^2\|\mathbf{u}_{i,t} - \nabla_v\phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t})\|^2 + \langle\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}), -\bar{\tau}_t\mathbf{u}_{i,t} + \bar{\tau}_t\nabla_v\phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t})\rangle\big] \\
&\overset{(b)}{\le} \mathbb{E}\big[\|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2 + 2\bar{\tau}_t^2 L_{\phi v}^2\|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2 - \frac{3\lambda\bar{\tau}_t}{4}\|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2 \\
&\quad + 2\bar{\tau}_t^2\|\mathbf{u}_{i,t} - \nabla_v\phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t})\|^2 + 4\lambda\bar{\tau}_t\|\mathbf{u}_{i,t} - \nabla_v\phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t})\|^2\big] \\
&\overset{(c)}{\le} \mathbb{E}\big[(1 - \frac{\lambda\bar{\tau}_t}{2})\|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2 + 5\lambda\bar{\tau}_t\|\mathbf{u}_{i,t} - \nabla_v\phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t})\|^2\big]
\end{aligned} \qquad (77)$$

where (a) uses the λ-strong convexity of ϕ_i, (b) uses

$$\begin{aligned}
&\langle \mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}), \bar{\tau}_t \mathbf{u}_{i,t} - \bar{\tau}_t \nabla_v \phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t}) \rangle \\
&= \langle \frac{\sqrt{\lambda \bar{\tau}_t}}{2} (\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})), 2\sqrt{\lambda \bar{\tau}_t} (\mathbf{u}_{i,t} - \nabla_v \phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t})) \rangle \\
&\leq \frac{\lambda \bar{\tau}_t}{4} \|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2 + 4\lambda \bar{\tau}_t \|\mathbf{u}_{i,t} - \nabla_v \phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t})\|^2
\end{aligned} \tag{78}$$

and (c) uses the assumption $\bar{\tau}_t \leq \min\left\{\frac{\lambda}{8L_{\phi v}^2}, \frac{\lambda}{2}\right\}$,

Then

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{v}_{i,t+1} - \mathbf{v}_i(\mathbf{x}_{t+1}, \mathbf{y}_{i,t+1})\|^2] \\
&\leq (1 + \frac{\lambda \bar{\tau}_t}{4}) \mathbb{E}[\|\mathbf{v}_{i,t+1} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2] + (1 + \frac{4}{\lambda \bar{\tau}_t}) \mathbb{E}[\|\mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \mathbf{v}_i(\mathbf{x}_{t+1}, \mathbf{y}_{i,t+1})\|^2] \\
&\leq (1 - \frac{\lambda \bar{\tau}_t}{4}) \mathbb{E}\|\mathbf{v}_{i,t} - \mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t})\|^2] + 10\lambda \bar{\tau}_t \mathbb{E}[\|\mathbf{u}_{i,t} - \nabla_v \phi_i(\mathbf{v}_{i,t}, \mathbf{x}_t, \mathbf{y}_{i,t})\|^2] \\
&\quad + \frac{5}{\lambda \bar{\tau}_t} E[\|\mathbf{v}_i(\mathbf{x}_t, \mathbf{y}_{i,t}) - \mathbf{v}_i(\mathbf{x}_{t+1}, \mathbf{y}_{i,t+1})\|^2]
\end{aligned} \tag{79}$$

where we use the assumption $\bar{\tau}_t \leq \frac{1}{\lambda}$. Take summation over all blocks $i = 1, \ldots, m$, we have

$$\mathbb{E}[\delta_{v,t+1}] \leq (1 - \frac{\lambda \tau_t}{4}) \mathbb{E}[\delta_{v,t}] + 10\lambda \bar{\tau}_t \mathbb{E}[\tilde{\delta}_{u,t}] + \frac{5L_v^2}{\lambda \tau_t} E[\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2] + \frac{5L_v^2}{\lambda \tau_t} \mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2] \tag{80}$$

□

E. Numeric Results of Hyper-parameter Optimization Experiment

Table 4. Testing accuracies and standard deviation over 3 runs with different random seeds from logistic regression, BSVRB$^{\text{v1}}$ with $m = 1$ lower-level problem, and BSVRB$^{\text{v1}}$ with $m = 100$ lower-level problems on various noise level of dataset *a8a*. Noise level represents the proportion of training sample labels that are flipped. 70% of the positive samples are removed from training data except for noise level $0*$, which means no label noise and no data imbalance.

Noise Level	Logistic Regression	BSVRB$^{\text{v1}}(m = 1)$	BSVRB$^{\text{v1}}(m = 100)$
0*	0.8528 ± 0.0005	0.8526± 0.0002	**0.8509± 0.0011**
0	0.8442 ± 0.0009	0.8426 ± 0.0016	**0.8477 ± 0.0013**
0.1	0.8285± 0.0034	0.8303 ± 0.0100	**0.8400± 0.0025**
0.2	0.8250±0.0066	0.8185±0.0090	**0.8388 ± 0.0024**
0.3	0.7929±0.0081	0.8118 ± 0.0047	**0.8239 ± 0.0015**
0.4	0.7715±0.0025	0.7749 ± 0.0079	**0.8051± 0.0013**