SPABA: A Single-Loop and Probabilistic Stochastic Bilevel Algorithm Achieving Optimal Sample Complexity

Anonymous Authors[1]

Abstract

While stochastic bilevel optimization methods have been extensively studied for addressing large-scale nested optimization problems in machine learning, it remains an open question whether the optimal complexity bounds for solving bilevel optimization are the same as those in single-level optimization. Our main result resolves this question: SPABA, an adaptation of the PAGE method for nonconvex optimization in (Li et al., 2021) to the bilevel setting, can achieve optimal sample complexity in both the finite-sum and expectation settings. We show the optimality of SPABA by proving that there is no gap in complexity analysis between stochastic bilevel and single-level optimization when implementing PAGE. Notably, as indicated by the results of (Dagréou et al., 2022), there might exist a gap in complexity analysis when implementing other stochastic gradient estimators, like SGD and SAGA. In addition to SPABA, we propose several other single-loop stochastic bilevel algorithms, that either match or improve the state-of-the-art sample complexity results, leveraging our convergence rate and complexity analysis. Numerical experiments demonstrate the superior practical performance of the proposed methods.

1. Introduction

Bilevel optimization, where one optimization problem is nested within the constraints of another, has extensive applications in fields such as transportation (Yang & Bell, 2001) and game theory (Von Stackelberg, 1952). In recent years, bilevel optimization has gained popularity in the machine learning community due to its broad range of applications, including hyperparameter optimization (Pedregosa, 2016; MacKay et al., 2019; Lorraine et al., 2020), meta-learning (Franceschi et al., 2018; Ji et al., 2020), and neural architecture search (Liu et al., 2018; Liang et al., 2019). Refer to recent survey papers (Liu et al., 2021; Zhang et al., 2023) for more applications of bilevel optimization in machine learning, computer vision and signal processing.

Bilevel optimization tackles challenges arising from hierarchical optimization, where decision variables in the upper level are also involved in the lower level. Typically, the bilevel optimization problems are formulated as

$$\min_{x\in\mathbb{R}^{d_x}} H(x) := f(x, y^*(x)) \tag{1}$$

$$\text{s.t. } y^*(x) := \arg\min_{y\in\mathbb{R}^{d_y}} g(x, y), \tag{2}$$

where the upper-level (UL) objective $f(x, y)$ and the lower-level (LL) objective $g(x, y)$ are two smooth real valued functions defined on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. In this work, we focus on the setting where the LL objective $g(x, y)$ is strongly convex with respect to (w.r.t.) y for any x, and the UL objective $f(x, y)$ is possibly nonconvex.

A commonly employed strategy for solving bilevel problems involves utilizing implicit differentiation, which yields the following expression for the hypergradient:

$$\nabla H(x) = \nabla_1 f(x, y^*(x)) - \nabla^2_{12} g(x, y^*(x))z^*(x), \tag{3}$$

where $z^*(x) = \left[\nabla^2_{22} g(x, y^*(x))\right]^{-1} \nabla_2 f(x, y^*(x))$. The practical implementation of the gradient descent method encounters several challenges, including: the computation of the exact solution $y^*(x)$ at the lower level, and the inversion of the Hessian $\nabla^2_{22} g$ at the point $(x, y^*(x))$. Utilizing a warm start strategy in the LL updates, results by (Ji et al., 2021; Liu et al., 2023) demonstrate that deterministic bilevel algorithms based on approximate implicit differentiation (AID) can achieve a convergence rate of $\mathcal{O}(\epsilon^{-1})$. The convergence rate matches that of the gradient descent method for nonconvex single-level optimization.

However, deterministic approaches necessitate the evaluation of the full gradient at every iteration, demanding substantial computational resources. This drawback renders these methods unsuitable for large-scale machine learning

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

tasks. In many applications of interest, the objective functions f and g have the **finite-sum form**:

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

which captures the standard empirical risk minimization problems in machine learning. Additionally, when dealing with a substantial or potentially infinite number of data samples, such as in online or streaming scenarios, f and g are commonly represented using the **expectation form**:

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

To improve sample efficiency compared to full-batch methods, it makes sense to apply stochastic techniques from single-level optimization to the bilevel optimization context. Unfortunately, the practical implementation of stochastic algorithms faces various challenges, such as: computing exactly the solution $y^*(x)$ at the lower level; inverting the inversion of the Hessian $\nabla^2_{22}g$; and addressing the nonlinear characteristics of ∇H within functions f and g. Therefore, a natural question follows: *Whether the optimal complexity bounds for solving bilevel optimization are the same as those in single-level optimization?* In fact, even more basic questions are open:

Question 1. *Is there a gap in complexity analysis between stochastic bilevel and single-level optimization when implementing the same stochastic gradient estimator?*

In the literature, various strategies have been proposed to tackle these challenges. For instance, the existing methods (Ghadimi & Wang, 2018; Ji et al., 2021; Yang et al., 2021; Chen et al., 2021; Guo et al., 2021; Khanduri et al., 2021b; Hong et al., 2023) employ one or multiple iterations of stochastic gradient descent (SGD) for the LL problem while incorporating truncated stochastic Neumann series to approximate the Hessian inversion. However, the mentioned methods suffer from an additional factor of $\log(\epsilon^{-1})$ in both sample complexity and batch size. Hence, there exists a gap in complexity analysis between stochastic bilevel and single-level optimization when employing the same stochastic gradient estimator in the aforementioned methods.

To address the nonlinear characteristics of ∇H within functions f and g and avoid relying on the stochastic Neumann approximation, recent works (Arbel & Mairal, 2022; Dagréou et al., 2022) have employed the decoupling approach. This approach breaks down the hypergradient computation into three gradient estimates, as outlined in (5)-(7). When utilizing the framework presented in (Arbel & Mairal, 2022; Dagréou et al., 2022), several studies have indicated that stochastic bilevel algorithms exhibit similar sample complexity when compared to their single-level counterparts. For instance, in the context of **general expectation setting**, by implementing multiple SGD iterations in subroutines and using large batchsizes of order $\mathcal{O}(\epsilon^{-1})$, AmIGO from (Arbel & Mairal, 2022) achieves the same sample complexity $\mathcal{O}(\epsilon^{-2})$ to SGD for smooth nonconvex single-level optimization, that is required to get an ϵ-stationary point, defined as $\mathbb{E}\|\nabla H(x)\|^2 \leq \epsilon$ (Ghadimi & Lan, 2013). When implementing the classical single-level SGD to the mentioned framework, we obtain the SOBA algorithm (Dagréou et al., 2022). However, by the result in Appendix D of (Dagréou et al., 2022), SOBA achieves a sample complexity of $\mathcal{O}(\epsilon^{-2.5})$ under standard smoothness assumptions. Thus, the gap in complexity analysis between stochastic bilevel and single-level optimization, using the SGD gradient estimator, is on the order of $\mathcal{O}(\epsilon^{-0.5})$. Recently, this gap has been effectively addressed by MA-SOBA (Chen et al., 2023), which builds upon the SOBA algorithm by incorporating an additional standard momentum (also referred to as moving average) into the update of the UL variable.

Despite the simplicity and power of MA-SOBA, we lack a comprehensive understanding of Question 1 regarding other stochastic gradient estimators. Specifically, if we additionally assume that the stochastic gradient satisfies a mean-squared smoothness property, as commonly assumed in the existing literature (Yang et al., 2021; Khanduri et al., 2021a), the lower bound for nonconvex stochastic optimization is improved to $\mathcal{O}(\epsilon^{-1.5})$ (Arjevani et al., 2023). It is natural to ask the following question:

Question 2. *How to develop a fully single-loop algorithm for solving stochastic bilevel optimization problems that achieves an optimal sample complexity $\mathcal{O}(\epsilon^{-1.5})$ under the bounded variance and mean-squared smoothness?*

In the **finite-sum setting**, as indicated by the result of SABA in Appendix D of (Dagréou et al., 2022), there also exists a gap of order $\mathcal{O}(n+m)^{1/3})$ between stochastic bilevel and single-level optimization in complexity analysis, when using the SAGA gradient estimator. In a recent work (Dagréou et al., 2023), the authors introduce SRBA, which is a bilevel extension of the well-known SARAH algorithm (Nguyen et al., 2017a). They demonstrate that SRBA achieves a better sample complexity of $\mathcal{O}((n+m)^{1/2}\epsilon^{-1})$, matching the lower bound they established for bilevel optimization. Unfortunately, the current analysis of SRBA relies on the assumption of higher-order smoothness for both the UL and LL functions to achieve optimality. It is also worth noting that SRBA in (Dagréou et al., 2023) utilizes a double-loop structure. Consequently, natural questions arise:

Question 3. *Is it possible to fill the gap between stochastic bilevel and single-level optimization when using the SAGA? How to develop a fully single-loop algorithm for solving stochastic bilevel optimization problems that achieves an optimal sample complexity $\mathcal{O}((n+m)^{\frac{1}{2}}\epsilon^{-1})$ under standard smoothness assumptions in the finite-sum setting?*

Table 1. Comparison of our methods with closely related works for **nonconvex-strongly-convex BLO** under standard smoothness assumptions, without relying on high-order smoothness. The $\tilde{\mathcal{O}}$ notation hides a factor of $\log(\epsilon^{-1})$. The sample complexity corresponds to the number of calls made to stochastic gradients and Hessian (Jocobian)-vector products required to get an ϵ-stationary point, i.e., $\mathbb{E}\|\nabla H(x)\|^2 \leq \epsilon$. * : This result can be found in Appendix D of (Dagréou et al., 2022).

Setting	Method	Stochastic Estimators	Sample Complexity	Batch Size
Expection (Mean-squared smoothness)	MRBO (Yang et al., 2021)	STORM	$\tilde{\mathcal{O}}(\epsilon^{-1.5})$	$\tilde{\mathcal{O}}(1)$
	SUSTAIN (Khanduri et al., 2021b)	STORM	$\tilde{\mathcal{O}}(\epsilon^{-1.5})$	$\tilde{\mathcal{O}}(1)$
	VRBO (Yang et al., 2021)	SARAH	$\tilde{\mathcal{O}}(\epsilon^{-1.5})$	$\tilde{\mathcal{O}}(\epsilon^{-0.5})$
	SRMBA (Ours)	STORM	$\tilde{\mathcal{O}}(\epsilon^{-1.5})$	$\mathcal{O}(1)$
	SPABA (Ours)	PAGE	$\mathcal{O}(\epsilon^{-1.5})$	$\mathcal{O}(1)$
Finite-Sum	SABA (Dagréou et al., 2022)	SAGA	$\mathcal{O}((n+m)\epsilon^{-1})^*$	$\mathcal{O}(1)$
	MA-SABA (Ours)	SAGA + x-Momentum	$\mathcal{O}((n+m)^{\frac{2}{3}}\epsilon^{-1})$	$\mathcal{O}(1)$
	SPABA (Ours)	PAGE	$\mathcal{O}((n+m)^{\frac{1}{2}}\epsilon^{-1})$	$\mathcal{O}(1)$

1.1. Main Contribution

The purpose of this work is to understand these theoretical questions. Our contributions are summarized below.

- **Bridging the gap between stochastic bilevel and single-level optimization when using the SAGA.** We first introduce a single-loop stochastic bilevel algorithm, named MA-SABA, that achieves a sample complexity of $\mathcal{O}((n+m)^{\frac{2}{3}}\epsilon^{-1})$ without the need for high-order smoothness. It is worth noting that MA-SABA is based on SABA and inspired by MA-SOBA by integrating an additional standard momentum into the update of the UL variable.

- **Achieving the optimal sample complexity in both the finite-sum and general expectation setting.** We propose a fully single-loop and sample-efficient stochastic bilevel algorithm, called SPABA, that achieves an optimal sample complexity of $\mathcal{O}((n+m)^{\frac{1}{2}}\epsilon^{-1})$ under standard smoothness assumptions in the finite-sum scenario. Additionally, it attains optimal sample complexity of $\mathcal{O}(\epsilon^{-1.5})$ under the bounded variance and mean-squared smoothness in the general expectation context. Technically, SPABA is an adaptation of the PAGE algorithm in (Li et al., 2021) to the bilevel setting.

- **Convergence rate and complexity analysis.** It is often difficult to analyze biased stochastic algorithms. We provide a general and unified convergence rates and complexity analysis based on biased stochastic gradient estimator such as STORM and PAGE, which either match or improve the state-of-the-art sample complexity results.

- Finally, numerical experiments demonstrate the superior efficiency of our proposed methods in bilevel optimization.

1.2. Additional Related Work

In the section we give a brief review of some recent works that are directly related to ours. A summary of the comparison of the proposed methods with closely related works is provided in Table 1.

Lower Bounds for Stochastic Bilevel Optimization. When $H(x)$ be convex or strongly convex, the study (Ji & Liang, 2022) has provided lower complexity bounds for deterministic bilevel optimization, that are larger than the corresponding optimal complexities of minimax optimization. For non-convex stochastic bilevel optimization, since nonconvex optimization can be regarded as a specific instance of a bilevel problem, it is natural to consider that lower bounds for nonconvex stochastic optimization also apply as lower bounds for bilevel counterparts. Therefore, the $\mathcal{O}(\epsilon^{-2})$ complexity is a lower bound for non-convex stochastic bilevel optimization in general expectation setting (Arjevani et al., 2023). Such complexity is attained by SGD in nonconvex stochastic optimization (Ghadimi & Lan, 2013). If we additionally assume that the stochastic gradient satisfies a mean-squared smoothness property, the lower bound is improved to $\mathcal{O}(\epsilon^{-1.5})$ (Arjevani et al., 2023), which is attained in nonconvex stochastic optimization by SPIDER (Fang et al., 2018), Spiderboost (Wang et al., 2018), SARAH (Pham et al., 2020), and PAGE (Li et al., 2021). Moreover, with the additional assumption of Lipschitz continuity, STORM (Cutkosky & Orabona, 2019) can also reach this complexity level. In the nonconvex finite-sum setting, if we assume that the objective function is averagely L-smooth, the lower bound becomes $\Omega(n^{1/2}\epsilon^{-1})$ (Fang et al., 2018; Li et al., 2021). Such complexity has been achieved by SARAH (Nguyen et al., 2017a;b; Pham et al., 2020), SPIDER (Fang et al., 2018), and PAGE (Li et al., 2021).

Discussion under Stronger Smoothness Conditions. Some studies have been conducted based on stronger smoothness conditions, such as SOBA and SABA in (Dagréou et al., 2022). Indeed, when the UL and LL objective functions possess high-order smoothness, their study illustrates that SABA, an adaptation of the SAGA algorithm (Defazio et al., 2014), exhibits a sample complexity of $\mathcal{O}((n+m)^{2/3}\epsilon^{-1})$. This is consistent with the sample complexity of SAGA in the single-level counterpart. Recently, leveraging on high-order smoothness, SRBA (Dagréou et al., 2023), an adaptation of the SARAH algorithm to the bilevel setting, achieves the same complexity $\mathcal{O}((n+m)^{1/2}\epsilon^{-1})$ as single-level SARAH. It is unclear whether a gap exists between stochastic bilevel and single-level optimization when utilizing the SARAH gradient estimator under standard smoothness assumptions.

2. The Proposed Stochastic Bilevel Algorithms

2.1. Overview of the Framework in (Arbel & Mairal, 2022; Dagréou et al., 2022)

In this section, we provide an overview of the algorithm design. First, we review the decoupling method employed in (Arbel & Mairal, 2022; Dagréou et al., 2022). To handle the nonlinear characteristics of ∇H within functions f and g, the authors in (Dagréou et al., 2022) introduce an extra variable $z \in \mathbb{R}^{d_y}$ to effectively decouple the nonlinear structure in ∇H. This allows us to utilize $\nabla_1 f(x,y) - \nabla_{12}^2 g(x,y)z$ to approximate the hypergradient $\nabla H(x)$, where y represents an approximate solution to the LL problem, while z serves as an inexact solution to the linear system $\left[\nabla_{22}^2 g(x,y)\right] z - \nabla_2 f(x,y) = 0$, which can also be seen as optimizing the following quadratic problem:

$$\min_z \frac{1}{2}\langle \nabla_{22}^2 g(x,y)z, z\rangle - \langle \nabla_2 f(x,y), z\rangle. \tag{4}$$

In summary, to solve the upper-level optimization problem $\min H(x)$, we decompose the search direction (or hypergradient estimate) of x into three steps, as follows:

$$D_x(x,y,z) = \nabla_1 f(x,y) - \nabla_{12}^2 g(x,y)z, \tag{5}$$
$$D_y(x,y,z) = \nabla_2 g(x,y), \tag{6}$$
$$D_z(x,y,z) = \nabla_{22}^2 g(x,y)z - \nabla_2 f(x,y). \tag{7}$$

Notably, all search directions are linear within functions f and g. The latter two directions align with two strongly convex optimization problems: the lower-level optimization problems (2) and (4). In addition, as detailed in Section 2.1 of (Liu et al., 2023), the search directions presented in (5-7) precisely correspond to the KKT condition of the equality-constrained optimization reformulation of (1):

$$\min_{x,y} f(x,y) \quad \text{s.t.} \quad \nabla_2 g(x,y) = 0.$$

Consequently, z can be interpreted as the dual multiplier.

Algorithm 1 Pseudocode for a generic Decoupling stochastic Bilevel Optimizer (DecBO)

```
Input: Initializations $(x_{-1}, y_{-1}, z_{-1})$ and $(x_0, y_0, z_0)$, number of total iterations $K$, step size $\{\alpha_k, \beta_k, \gamma_k\}$;
for $k = 0$ to $K-1$ do
    Sample $\mathcal{S}_k^f$ for $f$ and $\mathcal{S}_k^g$ for $g$;
    Construct an unbiased or biased estimator $v_k^x$ of $D_x(x_k, y_k, z_k)$ in (5) using $\mathcal{S}_k^f, \mathcal{S}_k^g$ and past gradient estimators;
    Update
        $x_{k+1} \leftarrow x_k - \alpha_k v_k^x;$   (8)
    Construct an unbiased or biased estimator $v_k^y$ of $D_y(x_k, y_k, z_k)$ in (6) using $\mathcal{S}_k^g$ and past gradient estimators;
    Update
        $y_{k+1} \leftarrow y_k - \beta_k v_k^y;$   (9)
    Construct an unbiased or biased estimator $v_k^z$ of $D_z(x_k, y_k, z_k)$ in (7) using $\mathcal{S}_k^f, \mathcal{S}_k^g$ and past gradient estimators;
    Update
        $z_{k+1} \leftarrow z_k - \gamma_k v_k^z.$   (10)
end for
```

Now, we provide a comprehensive description of the framework in (Arbel & Mairal, 2022; Dagréou et al., 2022), referred to as the Decoupling stochastic Bilevel Optimizer (DecBO). In each iteration, we sample $\mathcal{S}_k^f$ for $f(x,y)$ and $\mathcal{S}_k^g$ for $g(x,y)$. We then construct unbiased or biased stochastic gradient estimators, denoted as v_k^x, v_k^y and v_k^z, for $D_x(x_k,y_k,z_k)$, $D_y(x_k,y_k,z_k)$ and $D_z(x_k,y_k,z_k)$ in equations (5)-(7), respectively. These gradient estimators are constructed using the samples from $\mathcal{S}_k^f$ and $\mathcal{S}_k^g$, as well as past gradient estimators. We provide a pseudo code to illustrate this (see Algorithm 1).

The proposed framework opens opportunities for developing new algorithms in stochastic bilevel optimization. These algorithms can integrate diverse stochastic gradient estimation techniques from stochastic single-level optimization. For example, the aforementioned unbiased or biased gradient estimators can be efficiently constructed by combining variance-reduced gradient estimators like SAGA, SVRG, SPIDER or SARAH with momentum. Alternatively, one can utilize accelerated variance-reduced gradient estimators such as STORM or PAGE. We focus in this work on the loopless variance-reduced estimators because they share handy theoretical properties. As a result, the framework DecBO also benefits from a loopless structure. In the subsequent sections, we delve into the study of three such techniques.

2.2. MA-SABA: Bridging the Gap between Stochastic Bilevel and Single-level Optimization when Using the SAGA

For the finite-sum setting, we present MA-SABA, which is based on SABA (Dagréou et al., 2022) and inspired by MA-SOBA (Chen et al., 2023) by integrating an additional standard momentum into the update of the UL variable.

The SAGA method (Defazio et al., 2014) achieves variance reduction by updating historical gradients and performing gradient correction. Define two memory variables $w_{k,i} = (w_{k,i}^x, w_{k,i}^y, w_{k,i}^z)$ for $i \in [n]$ and $w_{k,j} = (w_{k,j}^x, w_{k,j}^y, w_{k,j}^z)$ for $j \in [m]$ corresponding to calls to f and g, respectively. At each iteration k, we draw two random independent indices $i \in [n]$ and $j \in [m]$ uniformly, for $i' \neq i$, do

$$\begin{cases} (w_{k+1,i}^x, w_{k+1,i}^y, w_{k+1,i}^z) = (x_k, y_k, z_k), \\ (w_{k+1,i'}^x, w_{k+1,i'}^y, w_{k+1,i'}^z) = (w_{k,i'}^x, w_{k,i'}^y, w_{k,i'}^z), \end{cases}$$

and similarly for $(w_{k+1,j}^x, w_{k+1,j}^y, w_{k+1,j}^z)$.

At each iteration k, we randomly select $i_k \in [n]$ and $j_k \in [m]$. In order to facilitate gradient correction, for $u_k := (x_k, y_k, z_k)$, we define two operations

$$\begin{aligned} \mathcal{M}_f(\phi, k, u, w) :=& \phi_{i_k}(x_k, y_k) - \phi_{i_k}(w_{k,i_k}^x, w_{k,i_k}^y) \\ &+ \sum_{i=1}^{n} \frac{\phi_i(w_{k,i}^x, w_{k,i}^y)}{n}, \\ \mathcal{M}_g(\phi, k, u, w) :=& \phi_{j_k}(u_k) - \phi_{j_k}(w_{k,j_k}) + \sum_{j=1}^{m} \frac{\phi_j(w_{k,j})}{m}. \end{aligned}$$

Then we update x_k using an additional standard momentum. The specific form of the iteration directions of MA-SABA are as follows:

$$\begin{aligned} v_k^y =& \mathcal{M}_g(\nabla_2 G, k, u, w), \\ v_k^z =& \mathcal{M}_g(\nabla_{22}^2 G z, k, u, w) - \mathcal{M}_f(\nabla_2 F, k, u, w), \\ v_k^x =& (1 - \rho_{k-1}) v_{k-1}^x + \rho_{k-1} \mathcal{M}_f(\nabla_1 F, k-1, u, w) \\ & - \rho_{k-1} \mathcal{M}_g(\nabla_{12}^2 G z, k-1, u, w). \end{aligned}$$

2.3. SPABA: Stochastic ProbAbilistic Bilevel Algorithm

Now we introduce SPABA, an adaptation of the PAGE algorithm in (Li et al., 2021) to the bilevel setting. To start, we present the algorithm description within the finite sum setting. During each iteration, we sample $I \subset [n]$ for f and $J \subset [m]$ for g, with a minibatch size of b. The PAGE method is utilized for stochastic gradient estimators in all three directions as follows:

$$\begin{aligned} v_k^y &= v_k(\nabla_2 G; b), \\ v_k^z &= v_k(\nabla_{22}^2 G z; b) - v_k(\nabla_2 F; b), \\ v_k^x &= v_k(\nabla_1 F; b) - v_k(\nabla_{12}^2 G z; b), \end{aligned}$$

where $\phi(u_k; b) = \frac{1}{b} \sum_{i' \in I} \phi_{i'}(u_k)$ and

$$v_k(\phi; b) = \begin{cases} \phi(u_k) & \text{w.p. } p, \\ v_{k-1}^x + \phi(u_k; b) - \phi(u_{k-1}; b) & \text{w.p. } 1 - p. \end{cases}$$

Recall that PAGE uses the vanilla minibatch SGD update with probability (w.p.) p, and reuses the previous gradient with a momentum-based minibatch SGD w.p. $1 - p$.

Furthermore, similar to PAGE, SPABA is adaptable to the general expectation setting by replacing the full gradient with another vanilla minibatch SGD using a minibatch size of τ'. Refer to Section D.1 for more details.

2.4. SRMBA: Stochastic Recursive Momentum Bilevel Algorithm

The STORM method (Cutkosky & Orabona, 2019) does not require the maintenance of anchor points or the use of large batches. Next, we propose SRMBA, which is a combination of the idea of STORM and the framework DecBO in Algorithm 1.

At each iteration k, we randomly select ξ and ζ for the functions f and g, respectively. Define $D_k^y = D_y(x_k, y_k, z_k; \zeta)$, $D_k^z = D_z(x_k, y_k, z_k; \xi, \zeta)$ and $D_k^x = D_x(x_k, y_k, z_k; \xi, \zeta)$. The iteration directions of SRMBA take the specific form as follows:

$$\begin{aligned} v_k^y &= \rho_k^y D_k^z + (1 - \rho_k^y)\left(D_k^y - D_{k-1}^y + v_{k-1}^y\right), \\ v_k^z &= \rho_k^z D_k^z + (1 - \rho_k^z)\left(D_k^z - D_{k-1}^z + v_{k-1}^z\right), \\ v_k^x &= \rho_k^x D_k^x + (1 - \rho_k^x)\left(D_k^x - D_{k-1}^x + v_{k-1}^x\right). \end{aligned}$$

3. Complexity Analysis

In this section, we will present the theoretical results for MA-SABA, SPABA and SRMBA, which either match or improve the state-of-the-art sample complexity results.

We say that $\bar{x}$ is a ϵ-stationary point if $\mathbb{E}\|\nabla H(\bar{x})\|^2 \leq \epsilon$. The sample complexity corresponds to the total number of calls made to stochastic gradients and Hessian (Jocobian)-vector products required to get an ϵ-stationary point.

3.1. Structure Assumptions

In order to provide convergence rates and complexity analysis, one usually needs the following standard assumptions depending on the setting (Ghadimi & Wang, 2018; Guo et al., 2021; Yang et al., 2021; Khanduri et al., 2021b; Chen et al., 2021; Arbel & Mairal, 2022; Dagréou et al., 2022; Hong et al., 2023; Chen et al., 2023; Huang, 2023).

Assumption 3.1. (1) ∇f is Lipschitz continuous in (x, y) with Lipschitz constant L^f; (2) There exists $C^f > 0$, such that $\|\nabla_2 f(x, y^*(x))\| \le C^f$ for any x.

Assumption 3.2. (1) ∇g and $\nabla^2 g$ are L_1^g and L_2^g Lipschitz continuous in (x, y), respectively; (2) $g(x, \cdot)$ is μ-strongly convex for any x.

Such assumptions are classical and sufficient to ensure the Lipschitz continuity of $y^*(x)$ and $z^*(x)$, the boundedness of $z^*(x)$, and the L-smoothness of $H(x)$. Next, we discuss assumptions made on the stochastic oracles.

Assumption 3.3. (**Bounded Variance**) In the general expectation setting, there exist positive constants σ_f, $\sigma_{g,1}$ and $\sigma_{g,2}$ such that

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

Furthermore, to achieve a better sample complexity results, we need to adopt the mean-squared smoothness assumption in (Arbel & Mairal, 2022; Chen et al., 2023).

Assumption 3.4. (**Mean-Squared Smoothness**) Stochastic functions $\nabla F(x, y; \xi)$, $\nabla G(x, y; \zeta)$ and $\nabla^2 G(x, y; \zeta)$ are L^f, L_1^g and L_2^g Lipschitz continuous in (x, y), respectively.

3.2. Convergence Analysis

We provide a general and unified convergence rates and complexity analysis and then illustrate it through the proposed methods. Let us identify what the crucial steps are. A clearer exposition of the analytical process is provided in Figures 3 and 4 of Appendix B.

General approach. One of the most important steps is to establish a recursive estimate often generated by two or three consecutive iterates:

$$\alpha_k \mathbb{E}\left[\|\nabla H(x_k)\|^2\right] \le L_k - L_{k+1} + \Delta_k, \tag{11}$$

where L_k, L_{k+1} and Δ_k are all nonnegative quantities, α_k is the step size used for updating x_k. Denote $\theta = \min_{k\in[K]}\{\alpha_k\}$. By induction, we have

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla H(x_k)\|^2\right] \le \frac{L_0}{K\theta} + \frac{\sum_{k=0}^{K-1}\Delta_k}{K\theta}.$$

This allows us to estimate the convergence rates of the underlying algorithm.

Usually, the recursive estimate (11) is derived through a series of recursive inequalities in conditional expectation:

$$\mathbb{E}\big[\widetilde{D}_{k+1} \,|\, \mathcal{F}_k\big] + \Lambda_k \le \omega_k \widetilde{D}_k + \Omega_k, \tag{12}$$

where $\widetilde{D}_k$, Λ_k, Ω_k are all nonnegative quantities, and $\omega_k \in [0, 1]$ is a contraction factor. We can now divide the proof of the recursive estimate (11) into four main steps:

(1) We begin by bounding the descent of $H(x)$ as follows:

$$\begin{aligned}\frac{\alpha_k}{2}\mathbb{E}\big[\,\|\nabla H(x_k)\|^2\,\big] \le\ & \mathbb{E}[H(x_k)] - \mathbb{E}[H(x_{k+1})]\\ &+ \Big(\frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2}\Big)\mathbb{E}\big[\,\|v_k^x\|^2\,\big]\\ &+ \frac{\alpha_k}{2}\mathbb{E}\big[\,\|\nabla H(x_k) - v_k^x\|^2\,\big],\end{aligned} \tag{13}$$

which is a recursive inequality, as demonstrated in (12). It is established in Lemma D.7 by the L^H-smoothness of $H(x)$;

(2) Considering the presence of a mean-squared error $\mathbb{E}\big[\,\|\nabla H(x_k) - v_k^x\|^2\,\big]$ in the right-hand side of (13), we study the descent of the mentioned mean-squared error. When integrating a standard momentum or a variation of momentum, such as those found in PAGE and STORM, into the update of x_k, we can establish a recursive inequality in the form of (12) for $\widetilde{D}_k := \|\nabla H(x_k) - v_k^x\|^2$. This inequality is derived from two or three consecutive iterates. For example, this result is proven in Lemma E.1 when standard momentum is utilized in v_k^x. It's important to highlight that the contraction factor $\omega_k = 1 - \rho_k$, where ρ_k is the "momentum" parameter and will tend to approach 0 in the subsequent setting. If there is no momentum, it is only possible to obtain an upper bound for the mean-squared error.

(3) To gain better control over the Ω_k-type terms in the descent of the mean-squared error $\mathbb{E}\big[\,\|\nabla H(x_k) - v_k^x\|^2\,\big]$, it is essential to investigate the descent of the approximate errors $\mathbb{E}\big[\,\|y_k - y^*(x_k)\|^2\,\big]$ and $\mathbb{E}\big[\,\|z_k - z^*(x_k)\|^2\,\big]$. By leveraging the strong convexity of the LL problem and the quadratic problem (4), one can readily derive recursive inequalities akin to (12) for the approximate errors. Additionally, $(1 - \omega_k)$ exhibits a similar order of magnitude as the step sizes for both y and z; please refer to Lemma E.2 for an illustration.

(4) In all the recursive inequalities mentioned above, the remaining terms include only the variances of the stochastic gradient estimators, such as $\mathbb{E}\big[\,\|D_x(x_k, y_k, z_k) - D_k^x\|^2\,\big]$. If the stochastic gradient estimators used lack variance reduction properties, like SGD, it is only feasible to attain a constant upper bound, even when we consider Assumption 3.3. To further reduce sampling complexity, one can integrate unbiased or biased variance reduction techniques into the algorithm. For example, MA-SABA aligns its sampling complexity with that of single-level optimization using SAGA. For an illustration, please refer to Lemma E.3.

3.3. Convergence Results

In this section, we provide the convergence results for the proposed stochastic bilevel algorithms. The detailed proofs of the results are deferred to the appendix.

We first provide the theoretical analysis of MA-SABA leading to a sample complexity in $\mathcal{O}((n+m))^{2/3}\epsilon^{-1})$ under standard smoothness assumptions in the finite-sum setting. This result bridges the gap between stochastic bilevel and single-level optimization when using the SAGA.

Theorem 3.5. ***(Convergence Rate of MA-SABA.)*** *Fix an iteration $K > 1$ and assume that Assumptions 3.1 to 3.2 and 3.4 hold. Then there exist positive constants c_1, c_2, c_3 and c_4 such that if $\alpha_k = c_1(n+m)^{-2/3}$, $\beta_k = c_2(n+m)^{-2/3}$, $\gamma_k = c_3(n+m)^{-2/3}$, $\rho_k = c_4(n+m)^{-2/3}$, the iterates in MA-SABA satisfy*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla H(x_k)\|^2\right] = \mathcal{O}\left((n+m)^{\frac{2}{3}}K^{-1}\right).$$

Remark 3.6. ***(Sample Complexity of MA-SABA.)*** *To achieve the ϵ-stationary point, the sampling complexity of MA-SABA is $\mathcal{O}((n+m))^{2/3}\epsilon^{-1})$, which is analogous to the sample complexity of SAGA in the nonconvex finite-sum setting.*

Next, we present the theoretical analysis of SPABA in both the finite-sum and general expectation scenarios.

Theorem 3.7. ***(Convergence Rate of SPABA in Finite-Sum Setting.)*** *Fix an iteration $K > 1$ and assume that Assumptions 3.1 to 3.2 and 3.4 hold. Then there exist positive constants c, c_β, and c_γ, such that if*

$$\alpha_k \le \frac{c}{1+\sqrt{\frac{1-p}{pb}}}, \quad \beta_k = c_\beta\alpha_k, \quad \gamma_k = c_\gamma\alpha_k,$$

the iterates in SPABA satisfy

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla H(x_k)\|^2\right] = \mathcal{O}\left(\frac{1+\sqrt{\frac{1-p}{pb}}}{K}\right).$$

Remark 3.8. ***(Sample Complexity of SPABA in Finite-Sum setting.)*** *If we take $p = b/(n+m+b)$ and $b = \mathcal{O}((n+m)^{1/2})$, then the sample complexity of SPABA is $\mathcal{O}((n+m))^{1/2}\epsilon^{-1})$. This implies that there is no gap between stochastic bilevel and single-level optimization in the context of PAGE implementation. The lower bound established in (Dagréou et al., 2023) for bilevel optimization indicates that SPABA attains optimal sample complexity in the finite-sum setting when $m = \mathcal{O}(n)$ and $\epsilon = \mathcal{O}(n^{-1/2})$.*

Theorem 3.9. ***(Convergence Rate of SPABA in Expectation Setting.)*** *Fix an iteration $K > 1$ and assume that Assumptions 3.1 to 3.4 hold. Choose minibatch size τ' and $b < \tau'$, the probability $p \in (0,1]$. Then there exist positive constants c, c_β, c_γ and σ, such that if*

$$\alpha_k \le \frac{c}{1+\sqrt{\frac{1-p}{pb}}}, \quad \beta_k = c_\beta\alpha_k, \quad \gamma_k = c_\gamma\alpha_k,$$

the iterates in SPABA satisfy

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla H(x_k)\|^2\right] = \mathcal{O}\left(\frac{1+\sqrt{\frac{1-p}{pb}}}{K} + \frac{1}{Kp\tau'} + \frac{\sigma}{\tau'}\right).$$

Remark 3.10. ***(Sample Complexity of SPABA in Expectation Setting.)*** *If we take $p = b/(n+m+b)$, $\tau' = \mathcal{O}(\epsilon^{-1})$ and $b \le \sqrt{\tau'}$, then the sample complexity of SPABA is $\mathcal{O}(\epsilon^{-1.5})$. This means that there is no gap between stochastic bilevel and single-level optimization when implementing PAGE. And SPABA achieves optimal sample complexity in the general expectation scenario.*

Finally, we state the convergence rate and sample complexity of SRMBA, an adaptation of the STORM method to the bilevel setting.

Theorem 3.11. ***(Convergence Rate of SRMBA in Expectation Setting.)*** *Fix an iteration $K > 1$ and assume that Assumptions 3.1 to 3.4 hold. Then there exist positive constants η, c_β, c_γ, c_x, c_y and c_z such that if*

$$\alpha_k = \frac{1}{(\eta+k)^{1/3}}, \quad \beta_k = c_\beta\alpha_k, \quad \gamma_k = c_\gamma\alpha_k;$$
$$\rho_k^x = c_x\alpha_k^2, \quad \rho_k^y = c_y\alpha_k^2, \quad \rho_k^z = c_z\alpha_k^2,$$

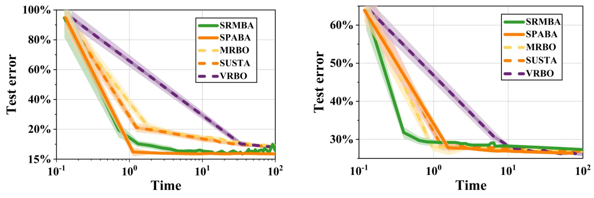
the iterates in SRMBA satisfy

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla H(x_k)\|^2\right] = \mathcal{O}\left(\frac{\log(K-1)}{K^{2/3}}\right).$$

Remark 3.12. ***(Sample Complexity of SRMBA in Expectation Setting.)*** *Theorem 3.11 implies that the sample complexity of SRMBA is $\mathcal{O}(\epsilon^{-1.5}\log(\epsilon^{-1}))$, which is analogous to the sample complexity of STORM in the nonconvex optimization. This tells us that there is no gap between stochastic bilevel and single-level optimization when implementing STORM.*

4. Numerical Experiments

While our contribution is mostly theoretical, we conducted a series of experiments to compare our proposed algorithms (SRMBA, SPABA, and MA-SABA) with their corresponding counterparts, namely, AmIGO (Arbel & Mairal, 2022), SUSTAIN (Khanduri et al., 2021b), SABA (Dagréou et al., 2022), SOBA (Dagréou et al., 2022), SRBA (Dagréou et al., 2023), MRBO (Yang et al., 2021), and VRBO (Yang et al., 2021). Further elaboration on these experiments is available in the Appendix.



(a) Datacleaning (b) Logistic regression

Figure 1. **Left:** Compare SRMBA and SPABA with other accelerated algorithms in a data hypercleaning experiment on the MINST dataset. **Right:** Compare SRMBA and SPABA with other accelerated algorithms in a hyperparameter selection experiment on the covtype dataset.

4.1. Data Hyper-Cleaning

The first learning task we perform is data hyper-cleaning conducted on the MNIST dataset [1] (Franceschi et al., 2017). The dataset is divided into a training set $\left(d_i^{\text{train}}, y_i^{\text{train}}\right)$, a validation set $\left(d_j^{\text{val}}, y_j^{\text{val}}\right)$ and a test set. The training set comprises 20,000 samples, the validation set contains 5,000 samples and the test set encompasses 10,000 samples. The target values y range from 0 to 9, while the samples d are of dimension 784. Within the training set, each sample is subject to corruption with a probability $\tilde{p}$: a sample d_i is deemed corrupted when its label y_i is replaced by a random label from the set $\{0, \ldots, 9\}$. Samples within the validation and test sets remain uncorrupted. The objective of data cleaning is to train a multinomial logistic regression model on the training set and ascertain a weight per training sample, ideally diminishing to 0 for corrupted samples. This is formalized by the bilevel optimization problem with $f(\lambda, \theta) = \frac{1}{m}\sum_{j=1}^{m} \ell\left(\theta d_j^{\text{val}}, y_j^{\text{val}}\right)$ and $g(\lambda, \theta) = \frac{1}{n}\sum_{i=1}^{n} \sigma\left(\lambda_i\right) \ell\left(\theta d_i^{\text{train}}, y_i^{\text{train}}\right) + C_r\|\theta\|^2$, where ℓ is the cross entropy loss and σ is the sigmoid function. We report in Figure 1(a) the test error, i.e., the percentage of incorrect predictions on the testing data. We utilize a corruption probability of $\tilde{p} = 0.7$ (sample corruption rate) for this experiment. In this scenario, SPABA demonstrates the most favorable performance.

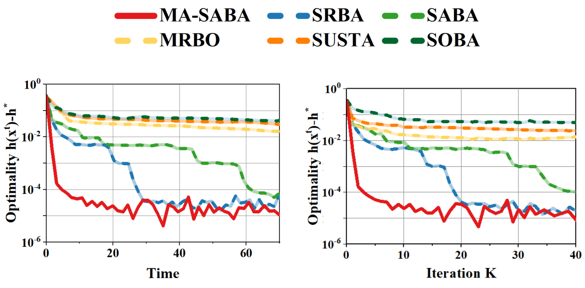


Figure 2. Comparison of MA-SABA with competitors in a hyperparameter selection experiment. The results indicate that MA-SABA outperforms other methods in terms of both time and iteration. Solid lines depict our proposed methods, whereas dashed lines represent competitors.

4.2. Hyperparameter Selection

We address hyperparameter selection for determining regularization parameters in ℓ^2 logistic regression. Let $\left(\left(d_i^{\text{train}}, y_i^{\text{train}}\right)\right) 1 \leq i \leq n$ and $\left(\left(d_j^{\text{val}}, y_j^{\text{val}}\right)\right) 1 \leq j \leq m$ denote the training and validation sets, respectively. In this context, the LL variable θ corresponds to the model parameters, while the UL variable λ represents the regularization parameter. The functions f and g for bilevel optimization are defined as follows: $f(\lambda, \theta) = \frac{1}{m}\sum_{j=1}^{m} \varphi\left(y_j^{\text{val}}\left\langle d_j^{\text{val}}, \theta\right\rangle\right)$ and $g(\lambda, \theta) = \frac{1}{n}\sum_{i=1}^{n} \varphi\left(y_i^{\text{train}}\left\langle d_i^{\text{train}}, \theta\right\rangle\right) + \frac{1}{2}\sum_{k=1}^{p} e^{\lambda_k}\theta_k^2$ where $\varphi(u) = \log\left(1 + e^{-u}\right)$. In this experiment, two datasets, namely IJCNN1 and covtype, are employed, corresponding to the algorithms MA-SABA, SPABA and SRMBA, respectively. In Figure 1(b), the test error is presented alongside the corresponding running time. It is observed that SRMBA exhibits the shortest runtime, while SPABA achieves the highest accuracy promptly. In the hyperparameter selection experiment, the suboptimality gap is depicted in Figure 2 for each method. The lowest values are attained by MA-SABA, indicating its superior performance. MA-SABA reaches a considerably high final value, significantly outperforming other methods.

5. Conclusion

In this work we propose a loopless and sample-efficient stochastic bilevel algorithm, named SPABA, achieving optimal sample complexity in both the finite-sum and expectation settings. Technically, SPABA is an adaptation of the PAGE algorithm in (Li et al., 2021) within the proposed framework in (Arbel & Mairal, 2022; Dagréou et al., 2022). More importantly, the complexity analysis of SPABA can be easily generalized to other stochastic gradient estimators. In fact, it already leads to MA-SABA and SRMBA that is an adaptation of STORM to the bilevel setting. It's worth noting that certain stochastic bilevel algorithms, like VRBO employing a double-loop variance reduction technique known as SARAH, are not covered by our convergence analysis framework. We acknowledge this limitation and plan to address it as part of our future work.

[1] http://yann.lecun.com/exdb/mnist/

6. Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Arbel, M. and Mairal, J. Amortized implicit differentiation for stochastic bilevel optimization. In *The Tenth International Conference on Learning Representations*, 2022.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199 (1-2):165–214, 2023.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Chen, T., Sun, Y., and Yin, W. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *NeurIPS*, volume 34, pp. 25294–25307, 2021.

Chen, X., Xiao, T., and Balasubramanian, K. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *arXiv preprint arXiv:2306.12067*, 2023.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Dagréou, M., Ablin, P., Vaiter, S., and Moreau, T. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *NeurIPS*, volume 35, pp. 26698–26710, 2022.

Dagréou, M., Moreau, T., Vaiter, S., and Ablin, P. A lower bound and a near-optimal algorithm for bilevel empirical risk minimization. *arXiv e-prints*, pp. arXiv–2302, 2023.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

Franceschi, L., Donini, M., Frasconi, P., and Pontil, M. Forward and reverse gradient-based hyperparameter optimization. In *ICML*, pp. 1165–1173, 2017.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *ICML*, pp. 1568–1577, 2018.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

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

Grazzi, R., Pontil, M., and Salzo, S. Convergence properties of stochastic hypergradients. In *International Conference on Artificial Intelligence and Statistics*, pp. 3826–3834. PMLR, 2021.

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

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Huang, F. On momentum-based gradient methods for bilevel optimization with nonconvex lower-level. *arXiv preprint arXiv:2303.03944*, 2023.

Ji, K. and Liang, Y. Lower bounds and accelerated algorithms for bilevel optimization. *JMLR*, 23:1–56, 2022.

Ji, K., Lee, J. D., Liang, Y., and Poor, H. V. Convergence of meta-learning with task-specific adaptation over partial parameters. *Advances in Neural Information Processing Systems*, 33:11490–11500, 2020.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *ICML*, pp. 4882–4892, 2021.

Khanduri, P., Zeng, S., Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A momentum-assisted single-timescale stochastic approximation algorithm for bilevel optimization. *arXiv preprint arXiv:2102.07367*, 2021a.

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. *Advances in neural information processing systems*, 34:30271–30283, 2021b.

Lam, S. K., Pitrou, A., and Seibert, S. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pp. 1–6, 2015.

Li, Z., Bao, H., Zhang, X., and Richtárik, P. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, pp. 6286–6295. PMLR, 2021.

Liang, H., Zhang, S., Sun, J., He, X., Huang, W., Zhuang, K., and Li, Z. Darts+: Improved differentiable architecture search with early stopping. *arXiv preprint arXiv:1909.06035*, 2019.

Liu, D. C. and Nocedal, J. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45:503–528, 1989.

Liu, H., Simonyan, K., and Yang, Y. Darts: Differentiable architecture search. In *ICLR*, 2018.

Liu, R., Gao, J., Zhang, J., Meng, D., and Lin, Z. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE TPAMI*, 44(12):10045–10067, 2021.

Liu, R., Liu, Y., Yao, W., Zeng, S., and Zhang, J. Averaged method of multipliers for bi-level optimization without lower-level strong convexity. In *ICML*, pp. 21839–21866, 2023.

Lorraine, J., Vicol, P., and Duvenaud, D. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pp. 1540–1552. PMLR, 2020.

MacKay, M., Vicol, P., Lorraine, J., Duvenaud, D., and Grosse, R. Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions. *arXiv preprint arXiv:1903.03088*, 2019.

Moreau, T., Massias, M., Gramfort, A., Ablin, P., Bannier, P.-A., Charlier, B., Dagréou, M., Dupre la Tour, T., Durif, G., Dantas, C. F., et al. Benchopt: Reproducible, efficient and collaborative optimization benchmarks. *Advances in Neural Information Processing Systems*, 35:25404–25421, 2022.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pp. 2613–2621. PMLR, 2017a.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv preprint arXiv:1705.07261*, 2017b.

Pedregosa, F. Hyperparameter optimization with approximate gradient. In *ICML*, pp. 737–746, 2016.

Pham, N. H., Nguyen, L. M., Phan, D. T., and Tran-Dinh, Q. Proxsarah: An efficient algorithmic framework for stochastic composite nonconvex optimization. *The Journal of Machine Learning Research*, 21(1):4455–4502, 2020.

Qu, G. and Li, N. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

Von Stackelberg, H. *The theory of the market economy*. William Hodge, 1952.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv*, 2018, 2018.

Yang, H. and Bell, M. G. Transport bilevel programming problems: recent methodological advances. *Transportation Research Part B: Methodological*, 35(1):1–4, 2001.

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

Zhang, Y., Khanduri, P., Tsaknakis, I., Yao, Y., Hong, M., and Liu, S. An introduction to bi-level optimization: Foundations and applications in signal processing and machine learning. *arXiv preprint arXiv:2308.00788*, 2023.

A. Appendix

The appendix is organized as follows:

- We present a unified framework for converagence analysis and highlight the proof sketch of Theorems in Section B.
- Additional experimental results are provided in Section C.
- Algorithms and general lemmas are provided in Section D.
- Proof details are provided in Sections E to H.
- The algorithm description and proof for MA-SOBA-q are provided in Section I.

B. Convergence Analysis Framework and Proof Sketches for Theorems

To analyze complexity, we introduce a general and unified convergence analysis method. In this section, we provide a more detailed exposition. Furthermore, we illustrate this by proving Theorems in this paper.

B.1. Convergence Analysis Framework

Our convergence analysis relies on the following recursive inequality

$$\alpha_k \mathbb{E}\left[\|\nabla H(x_k)\|^2\right] \le L_k - L_{k+1} + \Delta_k, \tag{14}$$

where L_k, L_{k+1}, and Δ_k are all positive terms, α_k is the step size used for updating the UL variable x_k. The term L_k is referred to as the potential function or Lyapunov function.

To derive (14), there are two crucial considerations: first, how to construct an appropriate L_k; and second, analyzing the descent of each element within L_k. Although the Lyapunov does not possess a uniform form, we start from the descent of the total UL objective and analyze layer-by-layer the elements it should comprise and their respective descents. This will be presented in B.2.

Assuming that, through the analysis of the decreasing properties of the elements in Lyapunov function and the selection of appropriate parameters, we have obtained (14). The next customary step is to define $\theta = \min_{k\in[K]}\{\alpha_k\}$, and by induction, we have

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla H(x_k)\|^2\right] \le \frac{L_0}{K\theta} + \frac{\sum_{k=0}^{K-1}\Delta_k}{K\theta},$$

which characterizes the convergence of the algorithm.

B.2. Proof Sketches for Theorems

In this section, we will present proof sketches for the theorems, utilizing a four-step layer-by-layer analysis to derive a recursive inequality similar to (14). Additionally, **this analysis showcases how to close the gap between stochastic bilevel and single-level optimization under classical assumptions and how to effectively handle biased stochastic estimations to attain superior complexity results.**

Usually, the recursive estimate (11) is derived through a series of recursive inequalities in conditional expectation:

$$\mathbb{E}\big[\widetilde{D}_{k+1} \,|\, \mathcal{F}_k\big] + \Lambda_k \le \omega_k \widetilde{D}_k + \Omega_k, \tag{15}$$

where $\widetilde{D}_k$, Λ_k, Ω_k are all nonnegative quantities, and $\omega_k \in [0,1]$ is a contraction factor.

Step 1: Originating from the descent of the total UL objective.

Under the classical assumptions, where the LL objective is of class $C_L^{1,1}$ space and strongly convex, and the UL objective is of class $C_L^{2,2}$ space, we can deduce that $H(x)$ is L^H-smooth. Consequently, we arrive at the following significant inequality

$$\frac{\alpha_k}{2}\mathbb{E}\left[\|\nabla H(x_k)\|^2\right] \le \mathbb{E}\left[H(x_k)\right] - \mathbb{E}\left[H(x_{k+1})\right] + \left(\frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2}\right)\mathbb{E}\left[\|v_k^x\|^2\right] + \underbrace{\frac{\alpha_k}{2}\mathbb{E}\left[\|\nabla H(x_k) - v_k^x\|^2\right]}, \tag{16}$$

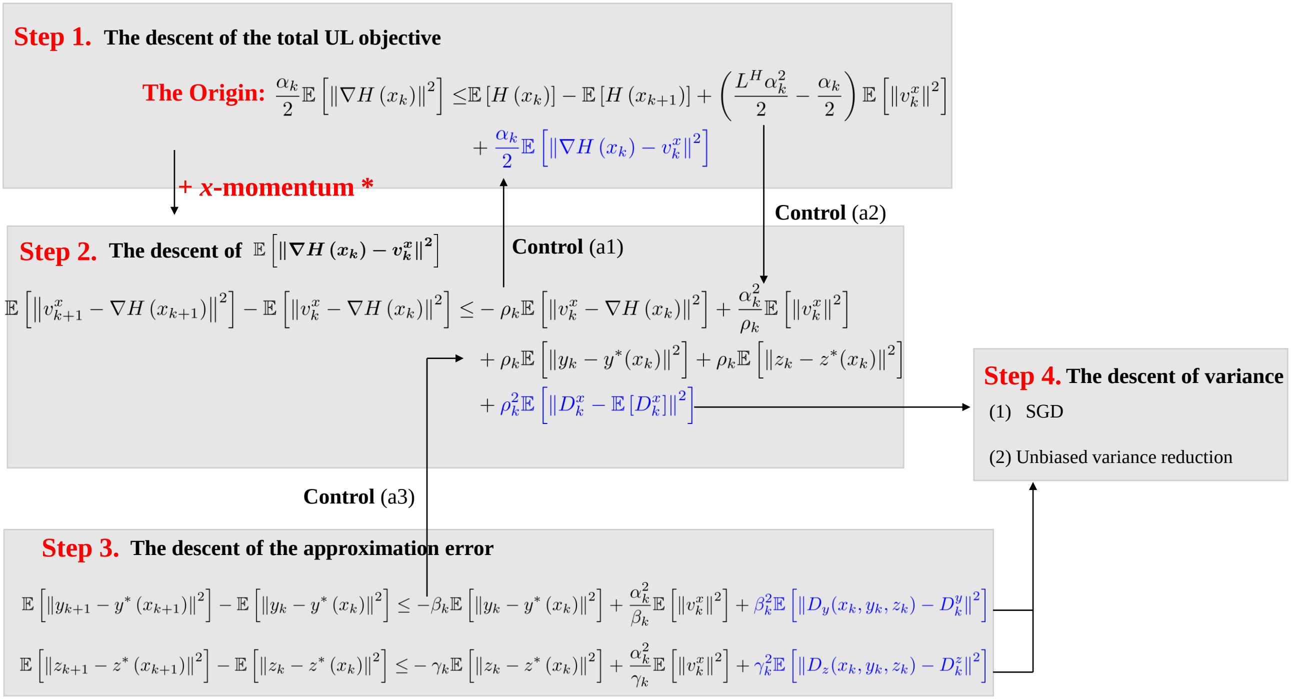

Step 1. The descent of the total UL objective
The Origin: $\frac{\alpha_k}{2}\mathbb{E}\left[\|\nabla H(x_k)\|^2\right] \leq \mathbb{E}[H(x_k)] - \mathbb{E}[H(x_{k+1})] + \left(\frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2}\right)\mathbb{E}\left[\|v_k^x\|^2\right] + \frac{\alpha_k}{2}\mathbb{E}\left[\|\nabla H(x_k) - v_k^x\|^2\right]$
+ x-momentum *
Control (a1)
Control (a2)
Step 2. The descent of $\mathbb{E}\left[\|\nabla H(x_k) - v_k^x\|^2\right]$
$\mathbb{E}\left[\|v_{k+1}^x - \nabla H(x_{k+1})\|^2\right] - \mathbb{E}\left[\|v_k^x - \nabla H(x_k)\|^2\right] \leq -\rho_k\mathbb{E}\left[\|v_k^x - \nabla H(x_k)\|^2\right] + \frac{\alpha_k^2}{\rho_k}\mathbb{E}\left[\|v_k^x\|^2\right] + \rho_k\mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right] + \rho_k\mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right] + \rho_k^2\mathbb{E}\left[\|D_k^x - \mathbb{E}[D_k^x]\|^2\right]$
Step 4. The descent of variance
(1) SGD
(2) Unbiased variance reduction
Control (a3)
Step 3. The descent of the approximation error
$\mathbb{E}\left[\|y_{k+1} - y^*(x_{k+1})\|^2\right] - \mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right] \leq -\beta_k\mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right] + \frac{\alpha_k^2}{\beta_k}\mathbb{E}\left[\|v_k^x\|^2\right] + \beta_k^2\mathbb{E}\left[\|D_y(x_k, y_k, z_k) - D_k^y\|^2\right]$
$\mathbb{E}\left[\|z_{k+1} - z^*(x_{k+1})\|^2\right] - \mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right] \leq -\gamma_k\mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right] + \frac{\alpha_k^2}{\gamma_k}\mathbb{E}\left[\|v_k^x\|^2\right] + \gamma_k^2\mathbb{E}\left[\|D_z(x_k, y_k, z_k) - D_k^z\|^2\right]$

Figure 3. Proof sketch (using unbiased estimations for the iteration directions of y and z)
(*): Fortunately, with the momentum applied to x, we achieve the second step, namely the descent of $\mathbb{E}[\|\nabla H(x_k) - v_k^x\|^2]$. Control(a1) demonstrates how Step 2 manages $\alpha_k\mathbb{E}[\|\nabla H(x_k) - v_k^x\|^2]$. Control(a2) and Control(a3) illustrate that step 2 does not introduce new, uncontrollable terms. Each term can be managed by inequalities found in either Step 1 or 2. The blue section in the figure indicating Step 4 highlights the variance terms critically influencing the convergence rate and complexity. This necessitates further examination in Step 4, employing either SGD or its variants with variance reduction.

which is pressented in Lemma D.7. This inequality characterizes the descent of the total UL objective. It forms the foundation of our analytical framework. With this lemma, we can achieve the same handling of the effect of α_k^2/β_k as in Lemma 3.9 in (Dagréou et al., 2022), even though we introduce a new term $\mathbb{E}\left[\|\nabla H(x_k) - v_k^x\|^2\right]$. We will address this term in the next step.

Step 2: The descent of $\mathbb{E}\left[\widetilde{D}_{k+1} \mid \mathcal{F}_k\right] = \mathbb{E}\left[\|\nabla H(x_k) - v_k^x\|^2\right]$.

In order to manage the newly introduced term, we have undertaken a pivotal step – the incorporation of **momentum** into the iteration direction of x_k to facilitate the term's descent. In this study, we utilize three distinct momentum methods: Moving-average, STORM, and PAGE.

Upon observing the right-hand side of the inequalities, what needs further bounding are $\mathbb{E}[\|y_k - y^*(x_k)\|^2]$ and $\mathbb{E}[\|z_k - z^*(x_k)\|^2]$, which will be the focus of our next step.

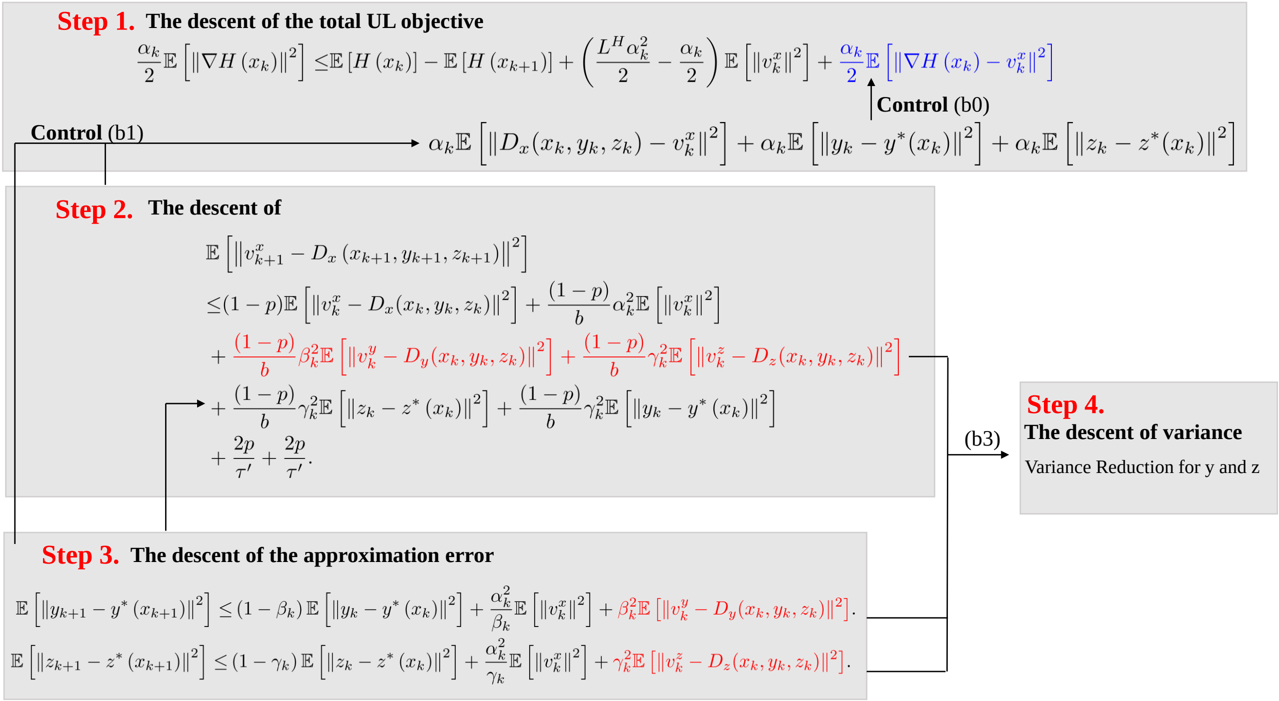
Step 3: The descent of the approximation error(better control over the Ω_k-type terms).

To gain better control over the Ω_k-type terms in the descent of the mean-squared error, in this step, we provide their descent to effectively control the approximate errors $\mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right]$ and $\mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right]$.

Step 4: The descent of variance.

In all the aforementioned inequalities, the remaining terms consist of only

$$\mathbb{E}\left[\|D_y(x_k, y_k, z_k) - D_k^y\|^2\right], \quad \mathbb{E}\left[\|D_x(x_k, y_k, z_k) - D_k^x\|^2\right], \quad \mathbb{E}\left[\|D_z(x_k, y_k, z_k) - D_k^z\|^2\right],$$



Step 1. **The descent of the total UL objective**

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

Control (b0)

Control (b1)

$$\alpha_k\mathbb{E}\left[\left\|D_x(x_k,y_k,z_k)-v_k^x\right\|^2\right]+\alpha_k\mathbb{E}\left[\left\|y_k-y^*(x_k)\right\|^2\right]+\alpha_k\mathbb{E}\left[\left\|z_k-z^*(x_k)\right\|^2\right]$$

Step 2. **The descent of**

$$\begin{aligned}
&\mathbb{E}\left[\left\|v_{k+1}^x-D_x\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^2\right]\\
\leq&(1-p)\mathbb{E}\left[\left\|v_k^x-D_x(x_k,y_k,z_k)\right\|^2\right]+\frac{(1-p)}{b}\alpha_k^2\mathbb{E}\left[\left\|v_k^x\right\|^2\right]\\
&+\frac{(1-p)}{b}\beta_k^2\mathbb{E}\left[\left\|v_k^y-D_y(x_k,y_k,z_k)\right\|^2\right]+\frac{(1-p)}{b}\gamma_k^2\mathbb{E}\left[\left\|v_k^z-D_z(x_k,y_k,z_k)\right\|^2\right]\\
&+\frac{(1-p)}{b}\gamma_k^2\mathbb{E}\left[\left\|z_k-z^*\left(x_k\right)\right\|^2\right]+\frac{(1-p)}{b}\gamma_k^2\mathbb{E}\left[\left\|y_k-y^*\left(x_k\right)\right\|^2\right]\\
&+\frac{2p}{\tau'}+\frac{2p}{\tau'}.
\end{aligned}$$

(b3)

Step 4.
The descent of variance
Variance Reduction for y and z

Step 3. **The descent of the approximation error**

$$\mathbb{E}\left[\left\|y_{k+1}-y^*\left(x_{k+1}\right)\right\|^2\right]\leq(1-\beta_k)\mathbb{E}\left[\left\|y_k-y^*\left(x_k\right)\right\|^2\right]+\frac{\alpha_k^2}{\beta_k}\mathbb{E}\left[\left\|v_k^x\right\|^2\right]+\beta_k^2\mathbb{E}\left[\|v_k^y-D_y(x_k,y_k,z_k)\|^2\right].$$

$$\mathbb{E}\left[\left\|z_{k+1}-z^*\left(x_{k+1}\right)\right\|^2\right]\leq(1-\gamma_k)\mathbb{E}\left[\left\|z_k-z^*\left(x_k\right)\right\|^2\right]+\frac{\alpha_k^2}{\gamma_k}\mathbb{E}\left[\left\|v_k^x\right\|^2\right]+\gamma_k^2\mathbb{E}\left[\|v_k^z-D_z(x_k,y_k,z_k)\|^2\right].$$

Figure 4. Proof sketch (using biased estimations for the iteration directions of y and z). In this analytical framework, we begin by dissecting $\mathbb{E}[\|\nabla H(x_k)-v_k^x\|^2]$ into two segments, $\mathbb{E}[\|D_x(x_k,y_k,z_k)-v_k^x\|^2]$ and $\mathbb{E}[\|y_k-y^*(x_k)\|^2]+\mathbb{E}[\|z_k-z^*(x_k)\|^2$, utilizing Control(b0). These segments are subsequently regulated by Step 2 and Step 3, referred to as Control(b1). The unique aspect here is that the inequalities applied in steps 2 and 3 are specifically designed for biased estimations. This adaptation enables the integration of an expanded selection of variance reduction methods to efficiently manage the red section.

representing the variance of the stochastic estimations. If we apply stochastic gradient descent to x, y, and z, while assuming Assumption 3.3, we can achieve a constant upper bound. This ultimately yields sampling complexity results equivalent to those obtained with single-layer optimization using SGD, as demonstrated by the results of MA-SOBA-q in sectionI.

To further reduce sampling complexity, we can incorporate unbiased variance reduction techniques into the algorithm. Just as we introduced our MA-SABA, it aligns its sampling complexity with that of single-level optimization with SAGA.

Extra: Utilizing Momentum-Based Biased Variance Reduction.

Our framework is meticulously designed to adeptly address biased estimations, recognizing the efficacy of targeted variance reduction strategies in yielding superior results. A pivotal distinction of our approach is the specialized adaptation of Step 2 and Step 3, meticulously crafted to accommodate biased estimations. By harnessing the capabilities of this framework, we unlock the potential to develop stochastic algorithms specifically engineered for bilevel optimization, thereby achieving markedly lower sampling complexities.

Therefore, by selecting appropriate step sizes and coefficients for the Lyapunov functions to scale the inequality, we can derive a recursive inequality of the form similar to (11).

Methods and Conclusions	Step 1	Step 2	Step 3	Step 4
MA-SABA (Th3.5)	Lemma D.7	Lemma E.1	Lemma E.2	Lemma E.5
SPABA (Th3.7)	Lemma D.7	Lemma F.4(2)	Lemma F.3	Lemma F.4(1)(3)
SPABA (Th3.9)	Lemma D.7	Lemma G.1(2)	Lemma F.3	Lemma G.1(1)(3)
SRMBA (Th3.11)	Lemma D.7	Lemma H.2(2)	Lemma F.3	Lemma H.2(1)(3)

Table 2. Lemmas Aligned with Each Step in Theorem Proofs

C. Additional experimental results

All experiments were conducted in Python, utilizing the Benchopt package (Moreau et al., 2022), JAX (Bradbury et al., 2018), and Numba (Lam et al., 2015) for efficient implementation of stochastic methods. For each problem, oracles for a given function f were employed, providing the quantities $\left(f(x,y), \nabla_1 f(x,y), \nabla^2_{22} f(x,y)z, \nabla^2_{12} f(x,y)z\right)$ to avoid redundant computation of intermediate results.

The experiments were executed using Python 3.8 on a system equipped with an Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz and an NVIDIA A100 GPU with 40GB of memory.

C.1. Hyperparameter selection on covtype dataset

Similar to (Dagréou et al., 2022), we conducted an additional experiment involving the selection of the best regularization parameter for an ℓ^2-regularized multinomial logistic regression problem on the covtype dataset[2]. This dataset comprises 581,012 samples with $p = 54$ features and encompasses $C = 7$ classes. Specifically, we utilized $n = 371,847$ training samples, $m = 92,962$ validation samples, and $n_{\text{test}} = 116,203$ test samples.

In this experiment, we fitted a multiclass logistic regression model on this dataset, with one hyperparameter per class. Thus, if $\left(d_i^{\text{train}}, y_i^{\text{train}}\right)_{i\in[n]}$ and $\left(d_j^{\text{val}}, y_j^{\text{val}}\right)_{j\in[m]}$ represent the training and validation samples, respectively, we solve the following bilevel optimization:

$$
\begin{aligned}
f(\lambda,\theta) &= \frac{1}{m}\sum_{j=1}^{m} \ell\left(\theta d_j^{\text{val}}, y_j^{\text{val}}\right) \text{ and } \\
g(\lambda,\theta) &= \frac{1}{n}\sum_{i=1}^{n} \ell\left(\theta d_i^{\text{train}}, y_i^{\text{train}}\right) + \sum_{c=1}^{C} e^{\lambda_c} \sum_{i=1}^{p} \theta_{i,c}^2,
\end{aligned}
$$

where $\lambda \in \mathbb{R}^C$ is the UL variable and $\theta \in \mathbb{R}^{p\times C}$ is the LL variable.

Hyper-parameter setting for algorithm. For SPABA, the probability $p = 0.5$, the step-sizes are chosen as $\alpha_k = 0.2/0.01, \gamma_k = \beta_k = 0.2$. For MA-SABA, the step-sizes are chosen as $\alpha_k = 0.2, \beta_k = 0.2/0.0001, \gamma_k = \beta_k$ and $\rho_k = 0.2$. For SRMBA, the step-sizes are chosen as $\alpha_k = \frac{5}{k^{1/3}}, \beta_k = \frac{0.2}{k^{1/3}}, \gamma_k = \frac{0.002}{k^{1/3}}$ and $\rho^x_k = \rho^y_k = \rho^z_k = \frac{0.5}{k^{2/3}}$. Other algorithms choose their step sizes according to the optimal strategy in (Dagréou et al., 2022).

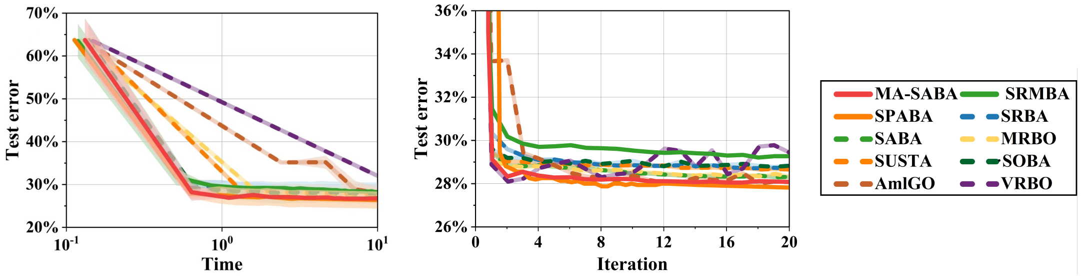


Figure 5. Comparison of MA-SABA, SPABA, and SRMBA with other stochastic bilevel optimization methods in a hyperparameter selection experiment. The result reveals that MA-SABA achieves the best performance in terms of both time and iteration. The dashed lines represent other stochastic bilevel optimization methods, while the solid lines denote the proposed methods.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_covtype.html

C.2. Hyperparameters selection on IJCNN1

In this experiment, we select the regularization parameters for a multiregularized logistic regression model, where we have one hyperparameter per feature:

$$f(\lambda, \theta) = \frac{1}{m}\sum_{i=1}^{m} \varphi\left(y_i^{\text{val}}\left\langle d_i^{\text{val}}, \theta\right\rangle\right) \text{ and}$$

$$g(\lambda, \theta) = \frac{1}{n}\sum_{i=1}^{n} \varphi\left(y_i^{\text{train}}\left\langle d_i^{\text{train}}, \theta\right\rangle\right) + \frac{1}{2}\theta^\top \operatorname{diag}\left(e^{\lambda_1}, \ldots, e^{\lambda_p}\right)\theta,$$

where λ, θ are the UL and LL variables, respectively. The parametrization choice, using e^λ rather than λ, ensures that there are no constraints placed on the variable λ. It is a classical approach in the bilevel optimization literature (Pedregosa, 2016; Ji et al., 2021; Grazzi et al., 2021).

In these experiments, as in (Dagréou et al., 2022), we employ Just-In-Time (JIT) compilation using the Numba package (Lam et al., 2015) to reduce Python overhead in the iteration loop. Additionally, to evaluate $H(\lambda)$, we utilize L-BFGS (Liu & Nocedal, 1989) to compute $y^*(x_k)$ and subsequently evaluate the function $H(x_k) = f(x_k, y^*(x_k))$.

Hyper-parameter setting for algorithm. For SPABA, the probability $p = 0.5$, the step-sizes are chosen as $\alpha_k = 0.2/0.01, \gamma_k = \beta_k = 0.2$. For MA-SABA, the step-sizes are chosen as $\alpha_k = 0.5, \beta_k = 0.5, \gamma_k = 0.4$ and $\rho_k = 0.2$. Other algorithms choose their step sizes according to the optimal strategy in (Dagréou et al., 2022).

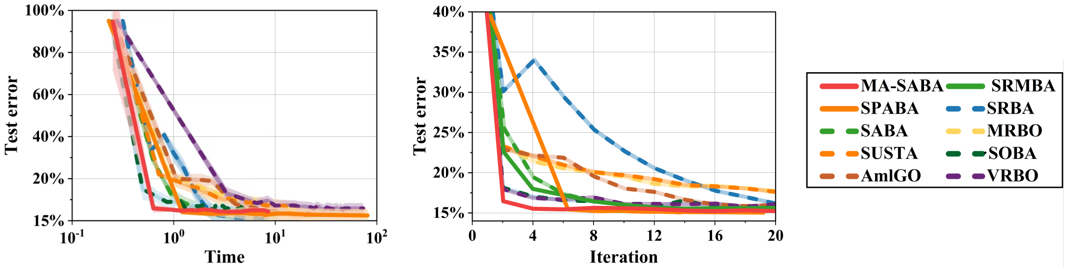


Figure 6. Comparison of MA-SABA, SPABA, and SRMBA with other stochastic bilevel optimization methods in a data hyper-cleaning experiment. It demonstrates that MA-SABA achieves superior performance in both time and iteration. The dashed lines represent other stochastic bilevel optimization methods, while the solid lines depict the proposed methods.

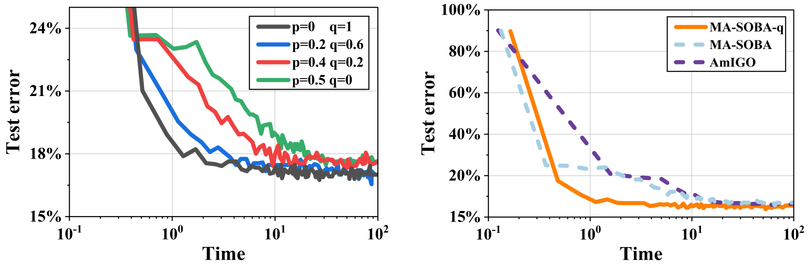


Figure 7. **Left:** Compare by selecting different q and p in MA-SOBA-q with data hyper-cleaning on MINST. **Right:** Compare of MA-SOBA-q with other acceleration algorithms on hyper-cleaning on MINST.

C.3. Data hyper-cleaning

Following the experimental setup in (Dagréou et al., 2022), we identified the optimal value for the regularization parameter C_r as 0.2 through a manual search, aiming to achieve the highest final test accuracy. It's worth noting that in this case, we were unable to utilize Just-In-Time (JIT) compilation from Numba due to the incompatibility of the softmax function from Scipy with Numba at the time of the experiment.

Figure 6 presents additional convergence curves with different methods. MA-SABA consistently emerges as the fastest algorithm to reach its final accuracy. Generally, the error decreases rapidly until it reaches a final value. Moreover, in Figure 7, we test the impact of the parameter q on the algorithm MA-SOBA-q's performance. We observe that as q increases starting from $q = 0$, the convergence speed of the algorithm also accelerates, aligning with our theoretical expectations.

Hyper-parameter setting for algorithm. For MA-SABA, the step-sizes are chosen as $\alpha_k = 0.005/0.0002, \beta_k = 0.005, \gamma_k = 0.01$ and $\rho_k = 0.2$. For SRMBA, the step-sizes are chosen as $\alpha_k = \frac{500}{k^{1/3}}, \beta_k = \frac{0.2}{k^{1/3}}, \gamma_k = \frac{0.02}{k^{1/3}}$ and $\rho_k^x = \rho_k^y = \rho_k^z = \frac{5}{k^{2/3}}$. Other algorithms choose their step sizes according to the optimal strategy in (Dagréou et al., 2022). In Figure 7, for MA-SOBA-q in Section I, the step-sizes and the batch-sizes are chosen as: $\alpha_k = 0.1/0.001, \beta_k = \gamma_k = 0.1, S = 1000(p = 0, q = 1)$; $\alpha_k = 0.1/0.001, \beta_k = \gamma_k = 0.1, S = 1000(p = 0.2, q = 0.6)$; $\alpha_k = 0.1/0.001, \beta_k = \gamma_k = 0.1, S = 1000(p = 0.4, q = 0.2)$; $\alpha_k = 0.1/0.001, \beta_k = \gamma_k = 0.1, S = 1000(p = 0.5, q = 0)$.

D. Algorithms and General lemmas

In this section, we present the specific forms of the algorithms MA-SABA, SPABA, and SRMBA and provide some general conclusions that are useful for the proof.

D.1. Algorithms

The red portion highlights the momentum introduced into the iteration direction of x_k.

Algorithm 2 MA-SABA

```
Input: Initializations $(x_{-1}, y_{-1}, z_{-1})$, $(x_0, y_0, z_0)$, and $v_{-1}^x$, number of total iterations $K$, step size $\{\alpha_k, \beta_k, \gamma_k\}$, momentum parameter $\rho_k$;
for $k = 0$ to $K - 1$ do
    Sample $i \in [n]$ for $f$ and $j \in [m]$ for $g$;
    $v_k^x = (1 - \rho_{k-1}) v_{k-1}^x + \rho_{k-1} D_{k-1}^x$;
    $x_{k+1} = x_k - \alpha_k v_k^x$;
    $D_k^x = \nabla_1 F_i(x_k, y_k) - \nabla_1 F_i(w_{k,i}^x, w_{k,i}^y) + \frac{1}{n}\sum_{i=1}^n \nabla_1 F_i(w_{k,i}^x, w_{k,i}^y) - \nabla_{12}^2 G_j(x_k, y_k) z_k + \nabla_{12}^2 G_j(w_{k,j}^x, w_{k,j}^y) w_{k,j}^z + \frac{1}{m}\sum_{j=1}^m \nabla_{12}^2 G_j(w_{k,j}^x, w_{k,j}^y) w_{k,j}^z$
    $v_k^y = \nabla_2 G_j(x_k, y_k) - \nabla_2 G_j(w_{k,j}^x, w_{k,j}^y) + \frac{1}{m}\sum_{j=1}^m \nabla_2 G_j(w_{k,j}^x, w_{k,j}^y)$;
    $y_{k+1} = y_k - \beta_k v_k^y$;
    $v_k^z = \nabla_{22}^2 G_j(x_k, y_k) z_k - \nabla_{22}^2 G_j(w_{k,j}^x, w_{k,j}^y) w_{k,j}^z + \frac{1}{m}\sum_{j=1}^m \nabla_{22}^2 G_j(w_{k,j}^x, w_{k,j}^y) w_{k,j}^z - \nabla_2 F_i(x_k, y_k) + \nabla_2 F_i(w_{k,i}^x, w_{k,i}^y) - \frac{1}{n}\sum_{i=1}^n \nabla_2 F_i(w_{k,i}^x, w_{k,i}^y)$;
    $z_{k+1} = z_k - \gamma_k v_k^z$.
end for
```

D.2. General lemmas

In this section, we present general conclusions that will be used, including an important lemma on the descent of $H(x_k)$

Lemma D.1. ***(Lipschitz continuity of** $y^*(x)$**)***

Under the Assumptions 3.2, $y^*(x)$ is L_{y^*}-Lipschitz continuous, where $L_{y^*} = \frac{L_1^g}{\mu}$.

Proof. See Lemma A.1 in (Dagréou et al., 2023). □

Lemma D.2. ***(Lipschitz continuity of** $z^*(x)$**)***

Under the Assumptions 3.1 and 3.2, $z^*(x)$ is L_{z^*} Lipschitz continuous, where $L_{z^*} = \left(\frac{L^f}{\mu} + \frac{C^f L_2^g}{\mu^2}\right)\left(1 + \frac{L_1^g}{\mu}\right)$.

Proof. See Lemma A.1 in (Dagréou et al., 2023). □

Lemma D.3. ***(boundness of** $z^*(x)$**)***

Algorithm 3 SPABA

```
Input: Initializations $(v_{-1}^x, v_{-1}^y, v_{-1}^z)$, $(x_{-1}, y_{-1}, z_{-1})$ and $(x_0, y_0, z_0)$, number of total iterations $K$, step size $\{\alpha_k, \beta_k, \gamma_k\}$, minibatch size $b$, constant $R$;
for $k = 0$ to $K-1$ do
    Sample $I$ for $f$ and $J$ for $g$, with minibatch size $|I| = |J| = b$;
    $v_k^x = \begin{cases} \frac{1}{n}\sum_{i\in[n]} \nabla_1 F_i(x_k, y_k) - \frac{1}{m}\sum_{j\in[m]} \nabla_{12}^2 G_j(x_k, y_k) z_k, & \text{with probability } p, \\ v_{k-1}^x + \frac{1}{b}\sum_{i\in I} (\nabla_1 F_i(x_k, y_k) - \nabla_1 F_i(x_{k-1}, y_{k-1})) & \\ \quad -\frac{1}{b}\sum_{j\in J} (\nabla_{12}^2 G_j(x_k, y_k) z_k - \nabla_{12}^2 G_j(x_{k-1}, y_{k-1}) z_{k-1}); & \text{with probability } 1-p; \end{cases}$
    $x_{k+1} = x_k - \alpha_k v_k^x$;
    $v_k^y = \begin{cases} \frac{1}{m}\sum_{j\in[m]} \nabla_2 G_j(x_k, y_k), & \text{with probability } p, \\ v_{k-1}^y + \frac{1}{b}\sum_{j\in J} (\nabla_2 G_j(x_k, y_k) - \nabla_2 G_j(x_{k-1}, y_{k-1})), & \text{with probability } 1-p; \end{cases}$
    $y_{k+1} = y_k - \beta_k v_k^y$;
    $v_k^z = \begin{cases} \frac{1}{m}\sum_{j\in[m]} \nabla_{22}^2 G_j(x_k, y_k) z_k - \frac{1}{n}\sum_{i\in[n]} \nabla_2 F_i(x_k, y_k), & \text{with probability } p, \\ v_{k-1}^z + \frac{1}{b}\sum_{j\in J} (\nabla_{22}^2 G_j(x_k, y_k) z_k - \nabla_{22}^2 G_j(x_{k-1}, y_{k-1}) z_{k-1}) & \\ \quad -\frac{1}{b}\sum_{i\in I} (\nabla_2 F_i(x_k, y_k) - \nabla_2 F_i(x_{k-1}, y_{k-1})), & \text{with probability } 1-p; \end{cases}$
    $z_{k+1} = \mathrm{Proj}_{\mathbb{B}(R)}(z_k - \gamma_k v_k^z)$.
end for
```

Algorithm 4 SRMBA

```
Input: Initializations $(x_{-1}, y_{-1}, z_{-1})$, $(x_0, y_0, z_0)$, and $v_{-1}^x$, number of total iterations $K$, step size $\{\alpha_k, \beta_k, \gamma_k\}$, momentum parameter $\{\rho_k^x, \rho_k^y, \rho_k^z\}$,constant $R$ ;
for $k = 0$ to $K-1$ do
    Sample $\xi$ for $f$ and $\zeta$ for $g$;
    $D_k^x = \nabla_1 F(x_k, y_k; \xi) - \nabla_{12}^2 G(x_k, y_k; \zeta) z_k$;
    $v_k^x = \rho_k^x D_k^x + (1 - \rho_k^x)(v_{k-1}^x + D_k^x - D_{k-1}^x)$;
    $x_{k+1} = x_k - \alpha_k v_k^x$;
    $D_k^y = \nabla_2 G(x_k, y_k; \zeta)$;
    $v_k^y = \rho_k^y D_k^y + (1 - \rho_k^y)(v_{k-1}^y + D_k^y - D_{k-1}^y)$;
    $y_{k+1} = y_k - \beta_k v_k^y$;
    $D_k^z = \nabla_{22}^2 G(x_k, y_k; \zeta) z_k - \nabla_2 F(x_k, y_k; \xi)$;
    $v_k^z = \rho_k^z D_k^z + (1 - \rho_k^z)(v_{k-1}^z + D_k^z - D_{k-1}^z)$;
    $z_{k+1} = \mathrm{Proj}_{\mathbb{B}(R)}(z_k - \gamma_k v_k^z)$.
end for
```

Under the Assumptions 3.1 and 3.2, $z^*(x)$ is bounded by R, i.e., for each x, we have

$$\|z^*(x)\| \leq \frac{C^f}{\mu} \triangleq R.$$

Proof. See Lemma B.2 in (Chen et al., 2023). □

Lemma D.4. ***(smoothness of function** H**)***

Suppose Assumptions 3.1 and 3.2 hold, the function $H(x)$ is L^H-smooth, where

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

Proof. See Lemma 2.2 in (Ghadimi & Wang, 2018). □

Lemma D.5. *Suppose Assumptions 3.1 and 3.2 hold. Then the following inequalities hold:*

$$\mathbb{E}\left[\|D_y(x_k,y_k,z_k)\|^2\right] \le \left(L_1^g\right)^2 \mathbb{E}\left[\|y_k - y^*\left(x_k\right)\|^2\right], \tag{17}$$

$$\mathbb{E}\left[\|D_z(x_k,y_k,z_k)\|^2\right] \le L_z^2 \mathbb{E}\left[\|z_k - z^*\left(x_k\right)\|^2\right] + L_z^2 \mathbb{E}\left[\|y_k - y^*\left(x_k\right)\|^2\right], \tag{18}$$

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

Proof. Proof of (17): Based on the fact that $D_y(x,y,z) = \nabla_2 g(x,y)$ and $D_y(x_k, y^*(x_k), z^*(x_k)) = \nabla_2 g(x_k, y^*(x_k)) = 0$, we have

$$\begin{aligned}\mathbb{E}\left[\|D_y(x_k,y_k,z_k)\|^2\right] &= \mathbb{E}\left[\|D_y(x_k,y_k,z_k) - D_y(x_k,y^*(x_k),z^*(x_k))\|^2\right]\\ &= \mathbb{E}\left[\|\nabla_2 g(x_k,y_k) - \nabla_2 g(x_k,y^*(x_k))\|^2\right]\\ &\le \left(L_1^g\right)^2 \mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right],\end{aligned}$$

where the last inequality utilizes the fact that ∇g is L_1^g-Lipschitz continuous, as stated in Assumption 3.2.

Proof of (18): Based on the fact that $D_z(x,y,z) = \nabla_{22}^2 g(x,y)z - \nabla_2 f(x,y)$ and $D_z(x_k,y^*(x_k),z^*(x_k)) = \nabla_{22}^2 g(x_k,y^*(x_k))z^*(x_k) - \nabla_2 f(x_k,y^*(x_k)) = 0$, we have

$$\begin{aligned}\mathbb{E}\left[\|D_z(x_k,y_k,z_k)\|^2\right] &= \mathbb{E}\left[\|D_z(x_k,y_k,z_k) - D_z(x_k,y^*(x_k),z^*(x_k))\|^2\right]\\ &= \mathbb{E}\left[\left\|\nabla_{22}^2 g(x_k,y_k)z_k - \nabla_2 f(x_k,y_k) - \nabla_{22}^2 g(x_k,y^*(x_k))z^*(x_k) + \nabla_2 f(x_k,y^*(x_k))\right\|^2\right]\\ &\le 3\mathbb{E}\left[\left\|\nabla_{22}^2 g(x_k,y_k)z_k - \nabla_{22}^2 g(x_k,y_k)z^*(x_k)\right\|^2\right]\\ &\quad + 3\mathbb{E}\left[\left\|\nabla_{22}^2 g(x_k,y_k)z^*(x_k) - \nabla_{22}^2 g(x_k,y^*(x_k))z^*(x_k)\right\|^2\right]\\ &\quad + 3\mathbb{E}\left[\|\nabla_2 f(x_k,y^*(x_k)) - \nabla_2 f(x_k,y_k)\|^2\right]\\ &\le 3\left(L_1^g\right)^2 \mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right] + 3R^2\left(L_2^g\right)^2 \mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right] + 3\left(L^f\right)^2 \mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right]\\ &= 3\left(L_1^g\right)^2 \mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right] + 3\left(R^2\left(L_2^g\right)^2 + \left(L^f\right)^2\right) \mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right]\\ &\le L_z^2 \mathbb{E}\left[\|z_k - z^*\left(x_k\right)\|^2\right] + L_z^2 \mathbb{E}\left[\|y_k - y^*\left(x_k\right)\|^2\right],\end{aligned}$$

where the validity of the second inequality is based on the application of Lemma D.3, along with the assumptions that ∇f is L^f-Lipschitz continuous as stated in Assumption 3.1, and $\nabla^2 g$ is L_2^g-Lipschitz continuous as mentioned in Assumption 3.2. The last inequality is due to the fact that $L_z^2 = \max\{3\left(L_1^g\right)^2, 3R^2\left(L_2^g\right)^2 + 3\left(L^f\right)^2\}$. □

Lemma D.6. *Suppose Assumptions 3.1 and 3.2 hold. Then we have*

$$\mathbb{E}\left[\|D_x(x_k,y_k,z_k) - \nabla H\left(x_k\right)\|^2\right] \quad \le 3\left(\left(L^f\right)^2 + \left(L_2^g R\right)^2\right) \mathbb{E}\left[\|y_k - y^*\left(x_k\right)\|^2\right] + 3\left(L_1^g\right)^2 \mathbb{E}\left[\|z_k - z^*\left(x_k\right)\|^2\right].$$

Proof. Using the unbiasedness of D_k^x and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}\left[\|D_x(x_k,y_k,z_k) - \nabla H\left(x_k\right)\|^2\right] \quad &= \quad \mathbb{E}\left[\left\|\nabla_1 f\left(x_k,y_k\right) + \nabla_{12}^2 g\left(x_k,y_k\right)z_k - \nabla_1 f\left(x_k,y^*\left(x_k\right)\right)\right.\right.\\ &\qquad \left.\left. - \nabla_{12}^2 g\left(x_k,y_k\right)z^*\left(x_k\right) + \nabla_{12}^2 g\left(x_k,y_k\right)z^*\left(x_k\right) - \nabla_{12}^2 g\left(x_k,y^*\left(x_k\right)\right)z^*\left(x_k\right)\right\|^2\right]\\ &\le \quad 3\left(\mathbb{E}\left[\|\nabla_1 f\left(x_k,y_k\right) - \nabla_1 f\left(x_k,y^*\left(x_k\right)\right)\|^2\right] + \mathbb{E}\left[\left\|\nabla_{12}^2 g\left(x_k,y_k\right)\left(z_k - z^*\left(x_k\right)\right)\right\|^2\right]\right.\\ &\qquad \left. + \mathbb{E}\left[\left\|\nabla_{12}^2 g\left(x_k,y_k\right) - \nabla_{12}^2 g\left(x_k,y^*\left(x_k\right)\right)\right\|^2 \|z^*\left(x_k\right)\|^2\right]\right).\end{aligned}$$

The three terms on the right-hand side of the above inequality can be bounded by utilizing Assumption 3.3 and Lemma D.3. Thus, the lemma is proven. □

Lemma D.7. *Suppose Assumptions 3.1 and 3.2 hold. Then we have*

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

Proof. The L^H-smoothness of $H(x)$ in Lemma D.4 implies

$$\begin{aligned}
H\left(x_{k+1}\right)-H\left(x_k\right) & \leq \left\langle\nabla H\left(x_k\right), x_{k+1}-x_k\right\rangle+\frac{L^H}{2}\left\|x_{k+1}-x_k\right\|^2 \\
& = -\alpha_k\left\langle\nabla H\left(x_k\right), v_k\right\rangle+\frac{L^H}{2} \alpha_k^2\left\|v_k^x\right\|^2 \\
& = -\frac{\alpha_k}{2}\left[\left\|\nabla H\left(x_k\right)\right\|^2+\left\|v_k\right\|^2-\left\|\nabla H\left(x_k\right)-v_k^x\right\|^2\right]+\frac{L^H}{2} \alpha_k^2\left\|v_k^x\right\|^2 \\
& = -\frac{\alpha_k}{2}\left\|\nabla H\left(x_k\right)\right\|^2+\left(\frac{L^H}{2} \alpha_k^2-\frac{\alpha_k}{2}\right)\left\|v_k^x\right\|^2+\frac{\alpha_k}{2}\left\|\nabla H\left(x_k\right)-v_k^x\right\|^2,
\end{aligned}$$

where the second equation uses the fact that $\langle a, b\rangle=\frac{1}{2}\left(\|a\|^2+\|b\|^2-\|a-b\|^2\right)$. Taking expectation on both sides, we know (D.7) holds. □

E. Proof of Theorem 3.5

Lemma E.1. *Suppose Assumptions 3.1 and 3.2 hold. Then we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|v_{k+1}^x-\nabla H\left(x_{k+1}\right)\right\|^2\right] \leq & \left(1-\rho_k\right) \mathbb{E}\left[\left\|v_k^x-\nabla H\left(x_k\right)\right\|^2\right]+\frac{2\left(L^H\right)^2 \alpha_k^2}{\rho_k} \mathbb{E}\left[\left\|v_k^x\right\|^2\right] \\
& +2 \rho_k \mathbb{E}\left[\left\|E[D_k^x]-\nabla H(x_k)\right\|^2\right]+\rho_k^2 \mathbb{E}\left[\left\|D_k^x-\mathbb{E}\left[D_k^x\right]\right\|^2\right],
\end{aligned}$$

where we require that $0 \leq \rho_k \leq 1$.

Proof. Due to iteratively updating v_k^x, we have

$$\begin{aligned}
& \mathbb{E}\left[\left\|v_{k+1}^x-\nabla H\left(x_{k+1}\right)\right\|^2\right] \\
= & \mathbb{E}\left[\left\|\left(1-\rho_k\right) v_k^x+\rho_k D_k^x-\nabla H\left(x_{k+1}\right)\right\|^2\right] \\
= & \mathbb{E}\left[\|\left(1-\rho_k\right)\left(v_k^x-\nabla H\left(x_k\right)\right)-\rho_k \nabla H\left(x_k\right)+\rho_k \mathbb{E}\left[D_k^x\right]-\rho_k \mathbb{E}\left[D_k^x\right]+\rho_k D_k^x+\nabla H\left(x_k\right)-\nabla H\left(x_{k+1}\right)\|^2\right] \\
= & \mathbb{E}\left[\left\|\left(1-\rho_k\right)\left(v_k^x-\nabla H\left(x_k\right)\right)-\rho_k \nabla H\left(x_k\right)+\rho_k \mathbb{E}\left[D_k^x\right]+\nabla H\left(x_k\right)-\nabla H\left(x_{k+1}\right)\right\|^2\right] \\
& +\rho_k^2 \mathbb{E}\left[\left\|D_k^x-\mathbb{E}\left[D_k^x\right]\right\|^2\right] \\
\leq & \left(1-\rho_k\right) \mathbb{E}\left[\left\|v_k^x-\nabla H\left(x_k\right)\right\|^2\right]+\rho_k \mathbb{E}\left[\left\|\mathbb{E}\left[D_k^x\right]-\nabla H\left(x_k\right)+\frac{\nabla H\left(x_k\right)-\nabla H\left(x_{k+1}\right)}{\rho_k}\right\|^2\right] \\
& +\rho_k^2 \mathbb{E}\left[\left\|D_k^x-\mathbb{E}\left[D_k^x\right]\right\|^2\right] \\
\leq & \left(1-\rho_k\right) \mathbb{E}\left[\left\|v_k^x-\nabla H\left(x_k\right)\right\|^2\right]+2 \rho_k \mathbb{E}\left[\left\|\mathbb{E}\left[D_k^x\right]-\nabla H\left(x_k\right)\right\|^2\right] \\
& +\frac{2 \alpha_k^2\left(L^H\right)^2}{\rho_k} \mathbb{E}\left[\left\|v_k^x\right\|^2\right]+\rho_k^2 \mathbb{E}\left[\left\|D_k^x-\mathbb{E}\left[D_k^x\right]\right\|^2\right],
\end{aligned}$$

where the third equation uses the unbiasedness of D_k^x, the first inequality is due to the convexity of $\|\cdot\|^2$, and the second inequality uses the L^H-smoothness of H. □

Lemma E.2. *Suppose Assumption 3.1 and 3.2 hold and the step size satisfy*

$$\beta_k \leq 1/(\mu+L_1^g), \quad \gamma_k \leq 1/(10 \mu).$$

Then we have

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

$$\mathbb{E}\left[\|z_{k+1}-z^*\left(x_{k+1}\right)\|^2\right]-\mathbb{E}\left[\|z_k-z^*\left(x_k\right)\|^2\right]\leq-\gamma_k\mu\mathbb{E}\left[\|z_k-z^*\left(x_k\right)\|^2\right]+8\Delta\gamma_k\mathbb{E}\left[\|y_k-y^*(x_k)\|^2\right]$$
$$+2\gamma_k^2\mathbb{E}\left[\|D_z(x_k,y_k,z_k)-D_k^z\|^2\right]+\frac{3L_{z^*}^2\alpha_k^2}{\gamma_k\mu}\mathbb{E}\left[\|v_k^x\|^2\right],$$

where $\Delta=\left(\left(L_2^gR\right)^2+\left(L^f\right)^2\right)/\mu$.

Proof. **Inequality for** y

We use Young inequality to start by expanding the square

$$\begin{aligned}\|y_{k+1}-y^*\left(x_{k+1}\right)\|^2 &= \|y_{k+1}-y^*\left(x_k\right)+y^*\left(x_k\right)-y^*\left(x_{k+1}\right)\|^2\\ &= \|y_{k+1}-y^*\left(x_k\right)\|^2+\|y^*\left(x_k\right)-y^*\left(x_{k+1}\right)\|^2+2\left\langle y_{k+1}-y^*\left(x_k\right),y^*\left(x_k\right)-y^*\left(x_{k+1}\right)\right\rangle\\ &\leq \left(1+\beta_k\mu\right)\|y_{k+1}-y^*\left(x_k\right)\|^2+\left(1+\frac{1}{\beta_k\mu}\right)\|y^*\left(x_k\right)-y^*\left(x_{k+1}\right)\|^2\\ &\leq \left(1+\beta_k\mu\right)\|y_{k+1}-y^*\left(x_k\right)\|^2+\left(1+\frac{1}{\beta_k\mu}\right)L_{y^*}^2\alpha_k^2\|v_k^x\|^2,\end{aligned}$$

where the last inequality is due to Lemma D.1.

Taking the expectation conditionally on x_k, y_k, z_k yields

$$E_k\left[\|y_{k+1}-y^*\left(x_{k+1}\right)\|^2\right]\quad\leq\quad\left(1+\beta_k\mu\right)E_k\left[\|y_{k+1}-y^*\left(x_k\right)\|^2\right]+\left(1+\frac{1}{\beta_k\mu}\right)L_{y^*}^2\alpha_k^2E_k[\|v_k^x\|^2]. \tag{19}$$

For the first member, we have

$$\begin{aligned}E_k\left[\|y_{k+1}-y^*\left(x_k\right)\|^2\right] &= E_k\left[\|y_k-y^*\left(x_k\right)-\beta_kv_k^y\|^2\right]\\ &= E_k\left[\|y_k-\beta_kD_y(x_k,y_k,z_k)-y^*\left(x_k\right)-\beta_k\left(v_k^y-D_y(x_k,y_k,z_k)\right)\|^2\right]\\ &= E_k\left[\|y_k-\beta_kD_y(x_k,y_k,z_k)-y^*\left(x_k\right)\|^2\right]+E_k\left[\|\beta_k\left(v_k^y-D_y(x_k,y_k,z_k)\right)\|^2\right]\\ &\quad +2\ E_k\left[\left\langle y_k-\beta_kD_y(x_k,y_k,z_k)-y^*\left(x_k\right),\beta_k\left(v_k^y-D_y(x_k,y_k,z_k)_k^*\right)\right\rangle\right]\\ &= E_k\left[\|y_k-\beta_kD_y(x_k,y_k,z_k)-y^*\left(x_k\right)\|^2\right]+E_k\left[\|\beta_k\left(v_k^y-D_y(x_k,y_k,z_k)\right)\|^2\right]\\ &\leq \left(1-\beta_k\mu\right)^2\|y_k-y^*\left(x_k\right)\|^2+\beta_k^2E_k\left[\|v_k^y-D_y(x_k,y_k,z_k)\|^2\right],\end{aligned}$$

where the first inequality holds because D_k^y is an unbiased estimate of $D_y(x_k,y_k,z_k)$. The first inequality utilizes Lemma 10 in (Qu & Li, 2017) which requires that g is strongly convex and Lipschitz smooth. Plugging it into (19) and taking the total expectation, we have

$$\begin{aligned}\mathbb{E}\left[\|y_{k+1}-y^*\left(x_{k+1}\right)\|^2\right] &\leq \left(1+\beta_k\mu\right)\left(1-\beta_k\mu\right)^2\mathbb{E}[\|y_k-y^*\left(x_k\right)\|^2]\\ &\quad +\left(1+\beta_k\mu\right)\beta_k^2\mathbb{E}\left[\|v_k^y-D_y(x_k,y_k,z_k)\|^2\right]+\left(1+\frac{1}{\beta_k\mu}\right)L_{y^*}^2\alpha_k^2\mathbb{E}[\|v_k^x\|^2]\\ &\leq \left(1-\beta_k\mu\right)\mathbb{E}[\|y_k-y^*\left(x_k\right)\|^2+2\beta_k^2\mathbb{E}\left[\|v_k^y-D_y(x_k,y_k,z_k)\|^2\right]+\frac{2L_{y^*}^2\alpha_k^2}{\beta_k\mu}\mathbb{E}[\|v_k^x\|^2],\end{aligned}$$

where the last inequality is due to $\beta_k \le 1/(\mu + L_1^g)$.

Inequality for z

Similar to the analysis of $\mathbb{E}\left[\|y(x_k) - y^*(x_k)\|^2\right]$, we analyze the auxiliary variable z

$$\begin{aligned} \|z_{k+1} - z^*(x_{k+1})\|^2 &= \|z_{k+1} - z^*(x_k) + z^*(x_k) - z^*(x_{k+1})\|^2 \\ &\le \left(1 + \frac{\gamma_k \mu}{2}\right) \|z_{k+1} - z^*(x_k)\|^2 + \left(1 + \frac{2}{\gamma_k \mu}\right) \|z^*(x_k) - z^*(x_{k+1})\|^2 . \end{aligned} \tag{20}$$

For the second term, taking total expectation and utilizing the Lipschitz continuity of $z * (x)$, we have

$$\mathbb{E}\left[\|z^*(x_k) - z^*(x_{k+1})\|^2\right] \le L_{z^*}^2 \alpha_k^2 \mathbb{E}\left[\|v_k^x\|^2\right]. \tag{21}$$

The analysis of the first term is more complex. Based on the definition of z_{k+1} and the fact that D_k^z is an unbiased estimate of $D_z(x_k, y_k, z_k)$, we have

$$\begin{aligned} E_k\left[\|z_{k+1} - z^*(x_k)\|^2\right] &= E_k\left[\left\|z_k - \gamma_k v_z^k - z^*(x_k)\right\|^2\right] \\ &= E_k\left[\|z_k - \gamma_k D_z(x_k, y_k, z_k) - z^*(x_k) - \gamma_k\left(v_k^z - D_z(x_k, y_k, z_k)\right)\|^2\right] \\ &= E_k\left[\|z_k - \gamma_k D_z(x_k, y_k, z_k) - z^*(x_k)\|^2\right] + \gamma_k^2 E_k\left[\|v_k^z - D_z(x_k, y_k, z_k)\|^2\right]. \end{aligned} \tag{22}$$

According to the definition of D_z and $D_z(x_k, y^*(x_k), z^*(x_k)) = 0$, we have

$$\begin{aligned} & E_k\left[\|z_k - \gamma_k D_z(x_k, y_k, z_k) - z^*(x_k)\|^2\right] \\ &= E_k\left[\left\|z_k - \gamma_k\left[\nabla_{22}^2 g(x_k, y_k) z_k - \nabla_2 f(x_k, y_k)\right] - z^*(x_k)\right\|^2\right] \\ &= E_k\left[\|z_k - z^*(x_k) - \gamma_k \nabla_{22}^2 g(x_k, y_k) z_k + \gamma_k \nabla_{22}^2 g(x_k, y_k) z_k^* - \gamma_k \nabla_{22}^2 g(x_k, y_k) z_k^* + \gamma_k \nabla_{22}^2 g(x_k, y^*(x_k)) z_k^*\right. \\ &\quad \left. + \gamma_k \nabla_2 f(x_k, y_k) - \gamma_k \nabla_2 f(x_k, y^*(x_k))\|^2\right] \\ &= E_k [\| \ (I - \gamma_k \nabla_{22}^2 g(x_k, y_k))(z_k - z^*(x_k)) \\ &\quad \left. + \gamma_k\left[(\nabla_{22}^2 g(x_k, y^*(x_k)) - \nabla_{22}^2 g(x_k, y_k)) z^*(x_k) + \nabla_2 f(x_k, y_k) - \nabla_2 f(x_k, y^*(x_k))\right]\|^2\right] \\ &\le \left(1 + \frac{\gamma_k \mu}{3}\right) E_k\left[\left\|(I - \gamma_k \nabla_{22}^2 g(x_k, y_k))(z_k - z^*(x_k))\right\|^2\right] \\ &\quad + \left(2 + \frac{6}{\gamma_k \mu}\right) \gamma_k^2\left[E_k\|\nabla_{22}^2 g(x_k, y^*(x_k)) - \nabla_{22}^2 g(x_k, y_k)\|^2 \|z^*(x_k)\|^2 + E_k\|\nabla_2 f(x_k, y_k) - \nabla_2 f(x_k, y^*(x_k))\|^2\right] \\ &\le \left(1 + \frac{\gamma_k \mu}{3}\right)(1 - \gamma_k \mu)^2 E_k\left[\|z_k - z^*(x_k))\|^2\right] + \left(2 + \frac{6}{\gamma_k \mu}\right) \gamma_k^2\left(\left(L_2^g R\right)^2 + (L^f)^2\right) E_k[\|y_k - y^*(x_k)\|^2]. \end{aligned} \tag{23}$$

Combining (20), (21), (22), and (23) and taking the total expectation, we have

$$\begin{aligned} \mathbb{E}\left[\|z_{k+1} - z^*(x_{k+1})\|^2\right] &\le \left(1 + \frac{\gamma_k \mu}{2}\right)\left(1 + \frac{\gamma_k \mu}{3}\right)(1 - \gamma_k \mu)^2 \mathbb{E}\left[\|z_k - z^*(x_k))\|^2\right] \\ &\quad + \left(1 + \frac{\gamma_k \mu}{2}\right)\left(2 + \frac{6}{\gamma_k \mu}\right) \gamma_k^2\left(\left(L_2^g R\right)^2 + (L^f)^2\right) \mathbb{E}[\|y_k - y^*(x_k)\|^2] \\ &\quad + \left(1 + \frac{\gamma_k \mu}{2}\right) \gamma_k^2 \mathbb{E}\left[\|v_k^z - D_z(x_k, y_k, z_k)\|^2\right] \\ &\quad + \left(1 + \frac{2}{\gamma_k \mu}\right) L_{z^*}^2 \alpha_k^2 \mathbb{E}\left[\|v_k^x\|^2\right] \\ &\le (1 - \gamma_k \mu) \mathbb{E}\left[\|z_k - z^*(x_k))\|^2\right] + 8\gamma_k \Delta \mathbb{E}[\|y_k - y^*(x_k)\|^2] \\ &\quad + 2\gamma_k^2 \mathbb{E}\left[\|v_k^z - D_z(x_k, y_k, z_k)\|^2\right] + \frac{3 L_{z^*}^2 \alpha_k^2}{\gamma_k \mu} \mathbb{E}\left[\|v_k^x\|^2\right]. \end{aligned}$$

For convenience of expression, let's denote $\Delta = \left(\left(L_2^g R\right)^2 + \left(L^f\right)^2 \right) / \mu$, and the last inequality is based on the choice of $\gamma_k \leq 1/(10\mu)$. □

To facilitate the discussion, we define $S_k = E_{k,f}^z + E_{k,f}^y + E_{k,f}^x + E_{k,g}^z + E_{k,g}^y + E_{k,g}^x$, where

$$E_{k,f}^z = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\|z_k - w_{k,i}^z\|^2],$$

and similarly $E_{k,f}^z$, $E_{k,f}^y$,$E_{k,f}^x$, $E_{k,g}^z$, $E_{k,g}^y$ and $E_{k,g}^x$. Additionally, let

$$\tau = \min\left\{ \frac{1}{2n}, \frac{1}{2m} \right\}.$$

Lemma E.3. *Suppose Assumption 3.1, 3.2 and 3.4 hold, there exist positive constants L_x', L_x'', L_z' and L_z'' such that*

$$\begin{aligned}
&\mathbb{E}\left[\|D_y(x_k, y_k, z_k) - D_k^y\|^2\right] \leq \left(L_1^g\right)^2 S_k, \\
&\mathbb{E}\left[\|D_z(x_k, y_k, z_k) - D_k^z\|^2\right] \leq L_z' S_k + L_z'' \left(\mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right] + \mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right] \right), \\
&\mathbb{E}\left[\|D_x(x_k, y_k, z_k) - D_k^x\|^2\right] \leq L_x' S_k + L_x'' \left(\mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right] + \mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right] \right).
\end{aligned}$$

where

$$\begin{aligned}
L_x' = L_z' &= \max\{16\left(L_2^g\right)^2 R^2, 16\left(L_1^g\right)^2, 2(L^f)^2\}, \\
L_x'' = L_z'' &= \max\{24\left(L_1^g\right)^2, 24R^2\left(L_2^g\right)^2\}.
\end{aligned}$$

Proof. Assuming we sample i and j from $[n]$ and $[m]$ at iteration k respectively, then we have

$$\begin{aligned}
&E_k\left[\|D_y\left(x_k, y_k, z_k\right) - D_k^y\|^2\right] \\
&= E_k[\|\nabla_2 g\left(x_k, y_k\right) - \nabla_2 G_j\left(x_k, y_k\right) + \nabla_2 G_j\left(w_{k,j}^x - w_{k,j}^y\right) - \frac{1}{m}\sum_{j=1}^{m} \nabla_2 G_j\left(w_{k,j}^x, w_{k,j}^y\right)\|^2] \\
&\leq E_k\left[\left\|\nabla_2 G_j\left(x_k, y_k\right) - \nabla_2 G_j\left(w_{k,j}^x, w_{k,j}^y\right)\right\|^2\right] \\
&= \frac{1}{m}\sum_{j=1}^{m} E_k\left[\left\|\nabla_2 G_j\left(x_k, y_k\right) - \nabla_2 G_j\left(w_{k,j}^x, w_{k,j}^y\right)\right\|^2\right] \\
&\leq \left(L_1^g\right)^2 \left(\frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\left[\left\|x_k - w_{k,j}^x\right\|^2\right] + \frac{1}{m}\mathbb{E}\left[\left\|y_k - w_{k,j}^y\right\|^2\right] \right),
\end{aligned}$$

where the first inequality uses the fact that $E[(X - E[X])^2] \leq E[X^2]$, the second inequality uses the Lipschitz continuity of ∇G_j. Taking the total expectation and by the definition of S_k, we can obtain

$$\mathbb{E}\left[\|D_y(x_k, y_k, z_k) - D_k^y\|^2\right] = \left(L_1^g\right)^2 \left(E_{k,g}^x + E_{k,g}^y\right) \leq \left(L_1^g\right)^2 S_k.$$

For x, we have

$$\begin{aligned}
&E_k\left[\|\nabla_x\left(x_k, y_k, z_k\right)-\nabla_k^x\|^2\right] \\
=&E_k\Big[\|\nabla_{22}^2 g\left(x_k, y_k\right) z_k-\nabla_2 f\left(x_k, y_k\right)-\nabla_{22}^2 G_j\left(x_k, y_k\right) z_k+\nabla_{22}^2 G_j\left(w_{k,j}^x, w_{k,j}^y\right) w_{k,j}^z \\
&\quad-\frac{1}{m} \sum_{j=1}^m \nabla_{22}^2 G_j\left(w_{k,j}^x, w_{k,j}^y\right) w_{k,j}^z+\nabla_2 F_i\left(x_k, y_k\right)-\nabla_2 F_i\left(w_{k,i}^x, w_{k,i}^y\right)+\frac{1}{n} \sum_{i=1}^n \nabla_2 F_i\left(w_{k,i}^x, w_{k,i}^y\right)\|^2\Big] \\
&\leq E_k\left[\|\nabla_{22}^2 G_j\left(w_{k,j}^x, w_{k,j}^y\right) w_{k,j}^z-\nabla_{22}^2 G_j\left(x_k, y_k\right) z_k+\nabla_2 F_i\left(x_k, y_k\right)-\nabla_2 F_i\left(w_{k,i}^x, w_{k,i}^y\right) w_{k,i}^z\|^2\right] \\
&\leq 2 E_k\left[\left\|\nabla_{22}^2 G_j\left(w_{k,j}^x, w_{k,j}^y\right) w_{k,j}^z-\nabla_{22}^2 G_j\left(x_k, y_k\right) z_k\right\|^2\right]+2 E_k\left[\left\|\nabla F_i\left(x_k, y_k\right)-\nabla F_i\left(w_{k,i}^x, w_{k,i}^y\right)\right\|^2\right] .
\end{aligned}$$

Taking the total expectation of the first term and we have

$$\begin{aligned}
&\mathbb{E}\left[E_k\left[\left\|\nabla_{22}^2 G_j\left(w_{k,j}^x, w_{k,j}^y\right) w_{k,j}^z-\nabla_{22}^2 G_j\left(x_k, y_k\right) z_k\right\|^2\right]\right] \\
=&\mathbb{E}[E_k[\|\nabla_{22}^2 G_j\left(x_k, y_k\right) z_k-\nabla_{22}^2 G_j\left(x_k, y_k\right) z^*\left(x_k\right)+\nabla_{22}^2 G_j\left(x_k, y_k\right) z^*\left(x_k\right)-\nabla_{22}^2 G_j\left(x_k, y^*\left(x_k\right)\right) z^*\left(x_k\right) \\
&+\nabla_{22}^2 G_j\left(x_k, y^*\left(x_k\right)\right) z^*\left(x_k\right)-\nabla_{22}^2 G_j\left(w_{k,j}^x, w_{k,j}^y\right) z^*\left(x_k\right) \\
&+\nabla_{22}^2 G_j\left(w_{k,j}^x, w_{k,j}^y\right) z^*\left(x_k\right)-\nabla_{22}^2 G_j\left(w_{k,j}^x, w_{k,j}^y\right) w_{k,j}^z\|^2]] \\
\leq& 4\left(L_1^g\right)^2 \mathbb{E}\left[\|z_k-z^*\left(x_k\right)\|^2\right]+4 R^2\left(L_2^g\right)^2 \mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right] \\
&+4\left(L_2^g\right)^2 R^2\left(\frac{1}{m} \sum_{j=1}^m \mathbb{E}\left[\left\|x_k-w_{k,j}^x\right\|^2\right]+2 E\left[\|y^*\left(x_k\right)-y_k\|^2\right]+\frac{2}{m} \sum_{j=1}^m E\left[\left\|y_k-w_{k,j}^k\right\|^2\right]\right) \\
&+4\left(L_1^g\right)^2\left(2 \mathbb{E}\left[\|z^*\left(x_k\right)-z_k\|^2\right]+\frac{2}{m} \sum_{j=1}^m \mathbb{E}\left[\left\|z_k-w_{k,j}^z\right\|^2\right]\right) \\
=&12\left(L_1^g\right)^2 \mathbb{E}\left[\|z_k-z^*\left(x_k\right)\|^2\right]+12 R^2\left(L_2^g\right)^2 \mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right] \\
&+4\left(L_2^g\right)^2 R^2 E_{k,g}^x+8\left(L_2^g\right)^2 R^2 E_{k,g}^y+8\left(L_1^g\right)^2 E_{k,g}^z
\end{aligned}$$

where the inequality is due to Assumption 3.4 and Lemma D.3.

The second term can be bounded as

$$\mathbb{E}\left[\left\|\nabla F_i\left(x_k, y_k\right)-\nabla F_i\left(w_{k,i}^x, w_{k,i}^y\right)\right\|^2\right] \leq (L^f)^2(E_{k,f}^x+E_{k,f}^y).$$

Combining the above inequalities, we have

$$\begin{aligned}
E_k\left[\|D_x\left(x_k, y_k, z_k\right)-D_k^x\|^2\right] \leq& 24\left(L_1^g\right)^2 \mathbb{E}\left[\|z_k-z^*\left(x_k\right)\|^2\right]+24 R^2\left(L_2^g\right)^2 \mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right] \\
&+8\left(L_2^g\right)^2 R^2 E_{k,g}^x+16\left(L_2^g\right)^2 R^2 E_{k,g}^y+16\left(L_1^g\right)^2 E_{k,g}^z+2(L^f)^2(E_{k,f}^x+E_{k,f}^y) \\
\leq& L_x^{\prime} S_k+L_x^{\prime\prime}\left(\mathbb{E}\left[\|y_k-y^*(x_k)\|^2\right]+\mathbb{E}\left[\|z_k-z^*(x_k)\|^2\right]\right),
\end{aligned}$$

where

$$L_x^{\prime}=\max\{16\left(L_2^g\right)^2 R^2, 16\left(L_1^g\right)^2, 2(L^f)^2\}, \quad L_x^{\prime\prime}=\max\{24\left(L_1^g\right)^2, 24 R^2\left(L_2^g\right)^2\}.$$

Similarly, we can obtain the inequality for $\mathbb{E}\left[\|D_z(x_k, y_k, z_k)-D_k^z\|^2\right]$. □

Lemma E.4. *For the error between the iterates and the memories, we have the following inequalities:*

$$\begin{aligned}
E^x_{k+1,f} &\le \left(1-\frac{1}{2n}\right)E^x_{k,f} + (2n+1)\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right],\\
E^x_{k+1,g} &\le \left(1-\frac{1}{2m}\right)E^x_{k,g} + (2m+1)\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right],\\
E^y_{k+1,f} &\le \left(1-\frac{1}{2n}\right)E^y_{k,f} + \beta_k^2\mathbb{E}\,\|D_k^y\|^2 + 2n\beta_k^2\mathbb{E}\left[\|D_y\left(x_k,y_k,z_k\right)\|^2\right],\\
E^y_{k+1,g} &\le \left(1-\frac{1}{2m}\right)E^y_{k,g} + \beta_k^2\mathbb{E}\,\|D_k^y\|^2 + 2m\beta_k^2\mathbb{E}\left[\|D_y\left(x_k,y_k,z_k\right)\|^2\right],\\
E^z_{k+1,f} &\le \left(1-\frac{1}{2n}\right)E^z_{k,f} + \gamma_k^2\mathbb{E}\,\|D_k^z\|^2 + 2n\gamma_k^2\mathbb{E}\left[\|D_z\left(x_k,y_k,z_k\right)\|^2\right],\\
E^z_{k+1,g} &\le \left(1-\frac{1}{2m}\right)E^z_{k,g} + \gamma_k^2\mathbb{E}\,\|D_k^z\|^2 + 2m\gamma_k^2\mathbb{E}\left[\|D_z\left(x_k,y_k,z_k\right)\|^2\right].
\end{aligned}$$

Proof. According to the definition of x_k^i, we have

$$\begin{aligned}
E_k\left[\left\|x_{k+1}-w^x_{k+1,i}\right\|^2\right] &= \frac{1}{n}E_k\left[\|x_{k+1}-x_k\|^2\right] + \frac{n-1}{n}E_k\left[\left\|x_{k+1}-w^x_{k,i}\right\|^2\right]\\
&= \frac{\alpha_k^2}{n}E\left[\|v_k^x\|^2\right] + \frac{n-1}{n}E_k\left[\left\|x_{k+1}-w^x_{k,i}\right\|^2\right].
\end{aligned}$$

For the second term, we use the Young's inequality, then

$$\begin{aligned}
E_k\left[\left\|x_{k+1}-w^x_{k,i}\right\|^2\right] &= E_k\left[\left\|x_k-\alpha_k v_k^x - w^x_{k,i}\right\|^2\right]\\
&= E_k\left[\left\|x_k-w^x_{k,i}\right\|^2 + \alpha_k^2\left\|v_k^x\right\|^2 - 2\alpha_k\left\langle v_k^x, x_k - w^x_{k,i}\right\rangle\right]\\
&\le E_k\left[\left\|x_k-w^x_{k,i}\right\|^2 + \alpha_k^2\left\|v_k^x\right\|^2 + \frac{\alpha_k}{2n\alpha_k}\left\|x_k-w^x_{k,i}\right\|^2 + 2n\alpha_k^2\left\|v_k^x\right\|^2\right].
\end{aligned}$$

Thus, we obtain

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

Taking the full expectation yields the desired inequality in the lemma. Similarly, we can obtain the result regarding $E^x_{k+1,g}$. The proof for $E^y_{k+1,f}$, $E^y_{k+1,g}$, $E^z_{k+1,f}$ and $E^z_{k+1,g}$ can be found in Lemma C.5 of (Dagréou et al., 2022). □

Lemma E.5. *Suppose Assumption 3.1, 3.2 and 3.4 hold, if $4\beta_k^2(L_1^g)^2 + 4\gamma_k^2L_z' \le \tau/2$, then*

$$S_{k+1} \le \left(1-\frac{\tau}{2}\right)S_k + (P_1\gamma_k^2 + P_2\beta_k^2)\mathbb{E}\left[\|y_k - y^*\left(x_k\right)\|^2\right] + P_3\gamma_k^2\mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right] + P_4\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right],$$

where

$$\begin{gathered}
P_1 = (2(m+n)+4)L_z^2 + 4L_z'', \quad P_2 = (2(m+m)+4)\left(L_1^g\right)^2,\\
P_3 = (2(m+n)+4)L_z^2 + 4L_z'', \quad P_4 = 2(m+n)+2.
\end{gathered}$$

Proof. By adding the inequalities in Lemma E.4, we obtain

$$\begin{aligned}
S_{k+1} \leq\; & (1-\tau)S_k + \mathbb{E}\Big[2\beta_k^2\left[\|D_k^y\|^2\right] + 2\gamma_k^2\left[\|D_k^z\|^2\right]\Big) \\
& +2(m+n)\left(\beta_k^2\mathbb{E}\left[\|D_y(x_k,y_k,z_k)\|^2\right] + \gamma_k^2\mathbb{E}\left[\|D_z(x_k,y_k,z_k)\|^2\right]\right) \\
& +2(m+n+1)\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right] \\
\leq\; & (1-\tau)S_k + 2\beta_k^2\left(2\mathbb{E}\left[\|D_y(x_k,y_k,z_k)\|^2\right] + 2\left(L_1^g\right)^2 S_k\right) \\
& +2\gamma_k^2\left(2\mathbb{E}\left[\|D_z(x_k,y_k,z_k)\|^2\right] + 2L_z' S_k + 2L_z''\mathbb{E}[\|y_k - y^*(x_k)\|^2] + 2L_z''\mathbb{E}[\|z_k - z^*(x_k)\|^2]\right) \\
& +2(m+n)\left(\beta_k^2\mathbb{E}\left[\|D_y(x_k,y_k,z_k)\|^2\right] + \gamma_k^2\mathbb{E}\left[\|D_z(x_k,y_k,z_k)\|^2\right]\right) + 2(m+n+1)\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right] \\
=\; & \left(1-\tau+4\beta_k^2\left(L_1^g\right)^2 + 4\gamma_k^2 L_z'\right)S_k + (2(m+n)+4)\beta_k^2\mathbb{E}\left[\|D_y(x_k,y_k,z_k)\|^2\right] \\
& +(2(m+n)+4)\gamma_k^2\mathbb{E}\left[\|D_z(x_k,y_k,z_k)\|^2\right] + 4\gamma_k^2 L_z''\left(\mathbb{E}[\|y_k - y^*(x_k)\|^2] + \mathbb{E}[\|z_k - z^*(x_k)\|^2]\right) \\
& +(2(m+n)+2)\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right] \\
\leq\; & \left(1-\tau+4\beta_k^2\left(L_1^g\right)^2 + 4\gamma_k^2 L_z'\right)S_k + (2(m+n)+2)\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right] \\
& +\left[(2(m+m)+4)\beta_k^2\left(L_1^g\right)^2 + (2(m+n)+4)\gamma_k^2 L_z^2 + 4\gamma_k^2 L_z''\right]\mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right] \\
& +\left[(2(m+n)+4)\gamma_k^2 L_z^2 + 4\gamma_k^2 L_z''\right]\mathbb{E}\left[\|z_k - z^*\left(x_k\right)\|^2\right],
\end{aligned}$$

where the second and third inequalities use Lemma E.3 and Lemma D.5, respectively.

Suppose $4\beta_k^2\left(L_1^g\right)^2 + 4\gamma_k^2 L' \leq \tau/2$, we have

$$\begin{aligned}
S_{k+1} \leq\; & \left(1-\frac{\tau}{2}\right)S_k + (2(m+n)+2)\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right] \\
& +\left[(2(m+m)+4)\beta_k^2\left(L_1^g\right)^2 + (2(m+n)+4)\gamma_k^2 L_z^2 + 4\gamma_k^2 L_z''\right]\mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right] \\
& +\left[(2(m+n)+4)\gamma_k^2 L_z^2 + 4\gamma_k^2 L_z''\right]\mathbb{E}\left[\|z_k - z^*\left(x_k\right)\|^2\right].
\end{aligned}$$

□

Theorem E.6. ***(Restatement of Theorem 3.5)***

Fix an iteration $K > 1$ and assume that Assumptions 3.1 to 3.2 and 3.4 hold. Let the step sizes be $\alpha_k = c_1 N^{-2/3}$, $\beta_k = c_2 N^{-2/3}$, $\gamma_k = c_3 N^{-2/3}$, $\rho_k = c_4 N^{-2/3}$. Take c_1, c_2, c_3 and c_4 satisfy

$$\begin{aligned}
c_2 &\leq \min\left\{\frac{\mu}{16c''}, \sqrt{\frac{c'}{16(L_1^g)^2}}\right\}, \\
c_3 &\leq \min\left\{\sqrt{\frac{c'}{16L_z'}}, \sqrt{\frac{\mu c_2}{16c''}}, \frac{\mu}{16\Delta}c_2\right\}, \\
c_4 &\leq \min\left\{\sqrt{\frac{\mu c_3}{8L_x''}}, \frac{2}{3}c_3, \frac{\mu}{12(L_1^g)^2}c_3\right\}, \\
c_1 &\leq \min\left\{\frac{1}{32c''}, \frac{1}{2L^H}, \frac{\mu}{16L_{y^*}^2}c_2, \frac{\mu}{48L_{z^*}^2}c_3, \frac{1}{64(L^H)^2}c_4, 2c_4\right\},
\end{aligned}$$

where $c' = 2$ and $c'' = \max\left\{6L_z^2 + 4L_z'', 6\left(L_1^g\right)^2, 4\right\}$ are constants that make $\tau \leq c'N^{-1}$ and $P_1, P_2, P_3, P_4 \leq c''N$ hold true, respectively. Then the iterates in MA-SABA satisfy

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla H\left(x_k\right)\|^2\right] = \mathcal{O}\left(N^{\frac{2}{3}}K^{-1}\right).$$

Proof. First, we introduce the notation $N = n + m$ and set $c' = 2$, $c'' = \max\left\{6L_z^2 + 4L_z'', 6\left(L_1^g\right)^2, 4\right\}$. From Lemma E.5, it is known that $\tau \leq c'N^{-1}$ and $P_1, P_2, P_3, P_4 \leq c''N$ hold true (see the original text lines 1145, 1317-1318). Then, we consider the Lyapunov function

$$L_k = \mathbb{E}\left[H\left(x_k\right)\right] + A\mathbb{E}\left[\left\|y_k - y^*\left(x_k\right)\right\|^2\right] + B\mathbb{E}\left[\left\|z_k - z^*\left(x_k\right)\right\|^2\right] + C\mathbb{E}[\|\nabla H(x_k) - v_k^x\|^2] + DS_k. \tag{24}$$

Using Lemma D.6 Lemma D.7, Lemma E.2, Lemma E.1 and Lemma E.5, we get

$$\begin{aligned}
L_{k+1} - L_k =& \mathbb{E}\left[H\left(x_{k+1}\right)\right] - E\left[H\left(x_k\right)\right] + A\left(\mathbb{E}\left[\left\|y_{k+1} - y^*\left(x_{k+1}\right)\right\|^2\right] - \mathbb{E}\left[\left\|y_k - y^*\left(x_k\right)\right\|^2\right]\right) \\
&+ B\left(\mathbb{E}\left[\left\|z_{k+1} - z^*\left(x_{k+1}\right)\right\|^2\right] - \mathbb{E}\left[\left\|z_k - z^*\left(x_k\right)\right\|^2\right]\right) \\
&+ C\left(\mathbb{E}[\|\nabla H(x_{k+1}) - v_{k+1}^x\|^2] - \mathbb{E}[\|\nabla H(x_k) - v_k^x\|^2]\right) \\
&+ D\left(S_{k+1} - S_k\right) \\
\leq& -\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] - \frac{\tau}{2}DS_k \\
&+ \left(\frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2} + A\frac{2L_{y^*}^2\alpha_k^2}{\beta_k\mu} + B\frac{3L_{z^*}^2\alpha_k^2}{\gamma_k\mu} + C\frac{2\left(L^H\right)^2\alpha_k^2}{\rho_k} + P_4\alpha_k^2 D\right)\mathbb{E}\left[\|v_k^x\|^2\right] \\
&+ \left(-A\beta_k\mu + 8\Delta B\gamma_k + 6C\Delta\rho_k + (P_1\gamma_k^2 + P_2\beta_k^2)D\right)\mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right] \\
&+ \left(-B\gamma_k\mu + 6C\left(L_1^g\right)^2\rho_k + P_3\gamma_k^2 D\right)\mathbb{E}\left[\left\|z_k - z^*\left(x_k\right)\right\|^2\right] \\
&+ \left(\frac{\alpha_k}{2} - C\rho_k\right)\mathbb{E}\left[\left\|\nabla H\left(x_k\right) - v_k^x\right\|^2\right] \\
&+ 2A\beta_k^2\mathbb{E}\left[\|D_y(x_k, y_k, z_k) - D_k^y\|^2\right] + 2B\gamma_k^2\mathbb{E}\left[\|D_z(x_k, y_k, z_k) - D_k^z\|^2\right] + C\rho_k^2\mathbb{E}\left[\|D_k^x - \mathbb{E}\left[D_k^x\right]\|^2\right],
\end{aligned}$$

For the variance terms in the above inequality, using Lemma E.3, we have

$$\begin{aligned}
L_{k+1} - L_k \leq& -\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] \\
&+ \left(-\frac{\tau}{2}D + 2A\left(L_1^g\right)^2\beta_k^2 + 2B\gamma_k^2 L_z' + C\rho_k^2 L_x'\right)S_k \\
&+ \left(\frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2} + A\frac{2L_{y^*}^2\alpha_k^2}{\beta_k\mu} + B\frac{3L_{z^*}^2\alpha_k^2}{\gamma_k\mu} + C\frac{2\left(L^H\right)^2\alpha_k^2}{\rho_k} + P_4\alpha_k^2 D\right)\mathbb{E}\left[\|v_k^x\|^2\right] \\
&+ \left(-A\beta_k\mu + 8\Delta B\gamma_k + 6C\Delta\rho_k + (P_1\gamma_k^2 + P_2\beta_k^2)D + 2B\gamma_k^2 L_z'' + C\rho_k^2 L_x''\right)\mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right] \\
&+ \left(-B\gamma_k\mu + 6C\left(L_1^g\right)^2\rho_k + P_3\gamma_k^2 D + 2B\gamma_k^2 L_z'' + C\rho_k^2 L_x''\right)\mathbb{E}\left[\left\|z_k - z^*\left(x_k\right)\right\|^2\right] \\
&+ \left(\frac{\alpha_k}{2} - C\rho_k\right)\mathbb{E}\left[\left\|\nabla H\left(x_k\right) - v_k^x\right\|^2\right],
\end{aligned}$$

We choose the coefficients of the Lyapunov function as $A = 1$, $B = 1$, $C = 1$, $D = N^{-1/3}$, and the step sizes $\alpha_k = c_1 N^{-2/3}$, $\beta_k = c_2 N^{-2/3}$, $\gamma_k = c_3 N^{-2/3}$, $\rho_k = c_4 N^{-2/3}$.

Based on our choice of

$$\begin{aligned}
c_2 &\le \min\left\{\frac{\mu}{16c''}, \sqrt{\frac{c'}{16(L_1^g)^2}}\right\},\\
c_3 &\le \min\left\{\sqrt{\frac{c'}{16L_z'}}, \sqrt{\frac{\mu c_2}{16c''}}, \frac{\mu}{16\Delta}c_2\right\},\\
c_4 &\le \min\left\{\sqrt{\frac{\mu c_3}{8L_x''}}, \frac{2}{3}c_3, \frac{\mu}{12(L_1^g)^2}c_3\right\},\\
c_1 &\le \min\left\{\frac{1}{32c''}, \frac{1}{2L^H}, \frac{\mu}{16L_{y^*}^2}c_2, \frac{\mu}{48L_{z^*}^2}c_3, \frac{1}{64(L^H)^2}c_4, 2c_4\right\},
\end{aligned}$$

we proceed with the following derivation:

Since $c_2 \le \frac{\mu}{16c''}$ and $c_3 \le \sqrt{\frac{\mu c_2}{16c''}}$, it follows that $c_3 \le \sqrt{\frac{\mu c_2}{16c''}} \le \frac{\mu}{16c''} \le \frac{\mu}{4c''}$;

Since $c_3 \le \sqrt{\frac{\mu c_2}{16c''}}$ and $c'' = \max\left\{6L_z^2 + 4L_z'', 6\left(L_1^g\right)^2, 4\right\} \ge 4L_z''$, it follows that $c_3 \le \sqrt{\frac{\mu c_2}{16c''}} \le \sqrt{\frac{\mu c_2}{16 \cdot 4L_z''}} \le \sqrt{\frac{\mu c_2}{32L_z''}}$;

Since $c_2 \le \frac{\mu}{16c''}$, $c_3 \le \sqrt{\frac{\mu c_2}{32L_z''}}$ and $c'' \ge 4L_z''$, it follows that $c_3 \le \sqrt{\frac{\mu c_2}{32c''}} \le \sqrt{\frac{\mu^2}{32 \cdot 16 \cdot L_z'' c''}} \le \frac{\mu}{16L_z''}$;

Since $c_4 \le \frac{2}{3}c_3$, $c_3 \le \min\left\{\sqrt{\frac{c'}{16L_z'}}, \frac{\mu}{16\Delta}c_2, \sqrt{\frac{\mu c_2}{32L_z''}}\right\}$, $L_x' = L_z'$ and $L_x'' = L_z''$, it follows that $c_4 \le \min\left\{\sqrt{\frac{c'}{8L_x'}}, \frac{\mu c_2}{24\Delta}, \sqrt{\frac{\mu c_2}{16L_x''}}\right\}$.

Therefore, we have

$$\begin{aligned}
c_2 &\le \min\left\{\frac{\mu}{16c''}, \sqrt{\frac{c'}{16(L_1^g)^2}}\right\},\\
c_3 &\le \min\left\{\frac{\mu}{16L_z''}, \sqrt{\frac{c'}{16L_z'}}, \frac{\mu}{4c''}, \sqrt{\frac{\mu c_2}{32L_z''}}, \sqrt{\frac{\mu c_2}{16c''}}, \frac{\mu}{16\Delta}c_2\right\},\\
c_4 &\le \min\left\{\sqrt{\frac{c'}{8L_x'}}, \frac{\mu c_2}{24\Delta}, \sqrt{\frac{\mu c_2}{16L_x''}}, \frac{\mu c_3}{12(L_1^g)^2}, \sqrt{\frac{\mu c_3}{8L_x''}}\right\},\\
c_1 &\le \min\left\{\frac{1}{32c''}, \frac{1}{2L^H}, \frac{\mu}{16L_{y^*}^2}c_2, \frac{\mu}{48L_{z^*}^2}c_3, \frac{1}{64(L^H)^2}c_4, 2c_4\right\},
\end{aligned}$$

it can be deduced that

$$\begin{aligned}
&\alpha_k \le \min\left\{\frac{1}{2L^H}, \frac{\mu}{16L_{y^*}^2}\beta_k, \frac{\mu}{48L_{z^*}^2}\gamma_k, \frac{1}{64(L^H)^2}\rho_k, 2\rho_k\right\},\\
&\gamma_k \le \min\left\{\frac{\mu}{16\Delta}\beta_k, \frac{\mu}{16L_z''}\right\}, \quad \gamma_k^2 \le \frac{\mu}{32L_z''}\beta_k,\\
&\rho_k \le \min\left\{\frac{\mu}{24\Delta}\beta_k, \frac{\mu}{12(L_1^g)^2}\gamma_k\right\}, \quad \rho_k^2 \le \min\left\{\frac{\mu}{8L_x''}\gamma_k, \frac{\mu}{16L_x''}\beta_k\right\}\\
&c_1 \le \frac{1}{32c''}, \quad c_2 \le \sqrt{\frac{c'}{8\left(L_1^g\right)^2}}, \quad c_3 \le \min\left\{\sqrt{\frac{c'}{16L_z'}}, \frac{\mu}{4c''}\right\}, \quad c_4 \le \sqrt{\frac{c'}{8L_x'}}, \quad c''(c_2^2 + c_3^2) \le \frac{c_2\mu}{8},\\
&\quad 4\beta_k^2(L_1^g)^2 + 4\gamma_k^2 L_z' \le \tau/2.
\end{aligned}$$

Then we have the following set of inequalities established:

$$\begin{cases} -\frac{\tau}{2}D + 2A\left(L_1^g\right)^2\beta_k^2 + 2B\gamma_k^2 L_z' + C\rho_k^2 L_x' \le 0, \\ \frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2} + A\frac{2L_{y^*}^2\alpha_k^2}{\beta_k\mu} + B\frac{3L_{z^*}^2\alpha_k^2}{\gamma_k\mu} + C\frac{2\left(L^H\right)^2\alpha_k^2}{\rho_k} + P_4\alpha_k^2 D \le 0, \\ -A\beta_k\mu + 8\Delta B\gamma_k + 6C\Delta\rho_k + (P_1\gamma_k^2 + P_2\beta_k^2)D + 2B\gamma_k^2 L_z'' + C\rho_k^2 L_x'' \le 0, \\ -B\gamma_k\mu + 6C\left(L_1^g\right)^2\rho_k + P_3\gamma_k^2 D + 2B\gamma_k^2 L_z'' + C\rho_k^2 L_x'' \le 0, \\ \frac{\alpha_k}{2} - C\rho_k \le 0. \end{cases}$$

To make the proof more comprehensive, we will verify the validity of each inequality one by one.

- **Inequality 1:**

$$\begin{aligned} & -\frac{\tau}{2}D + 2A\left(L_1^g\right)^2\beta_k^2 + 2B\gamma_k^2 L_z' + C\rho_k^2 L_x' \\ & \le -\frac{c'}{2}N^{-1-\frac{1}{3}} + 2\left(L_1^g\right)^2 c_2^2 N^{-\frac{4}{3}} + 2c_3^2 L_z' N^{-\frac{4}{3}} + c_4^2 L_x' N^{-\frac{4}{3}} \\ & \le -\frac{c'}{4}N^{-\frac{4}{3}} + 2c_3^2 L_z' N^{-\frac{4}{3}} + c_4^2 L_x' N^{-\frac{4}{3}} \\ & \le -\frac{c'}{8}N^{-\frac{4}{3}} + c_4^2 L_x' N^{-\frac{4}{3}} \\ & \le 0, \end{aligned}$$

where the justification for the four inequalities holding true are, respectively, $\tau \le c'N^{-1}$, $c_2 \le \sqrt{\frac{c'}{8\left(L_1^g\right)^2}}$, $c_3 \le \sqrt{\frac{c'}{16L_z'}}$, and $c_4 \le \sqrt{\frac{c'}{8L_x'}}$.

- **Inequality 2:**

$$\begin{aligned} & \frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2} + A\frac{2L_{y^*}^2\alpha_k^2}{\beta_k\mu} + B\frac{3L_{z^*}^2\alpha_k^2}{\gamma_k\mu} + C\frac{2\left(L^H\right)^2\alpha_k^2}{\rho_k} + P_4\alpha_k^2 D \\ & = \frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2} + \frac{2L_{y^*}^2\alpha_k^2}{\beta_k\mu} + \frac{3L_{z^*}^2\alpha_k^2}{\gamma_k\mu} + \frac{2\left(L^H\right)^2\alpha_k^2}{\rho_k} + P_4\alpha_k^2 N^{-\frac{1}{3}} \\ & \le -\frac{\alpha_k}{4} + \frac{2L_{y^*}^2\alpha_k^2}{\beta_k\mu} + \frac{3L_{z^*}^2\alpha_k^2}{\gamma_k\mu} + \frac{2\left(L^H\right)^2\alpha_k^2}{\rho_k} + P_4\alpha_k^2 N^{-\frac{1}{3}} \\ & \le -\frac{\alpha_k}{8} + \frac{3L_{z^*}^2\alpha_k^2}{\gamma_k\mu} + \frac{2\left(L^H\right)^2\alpha_k^2}{\rho_k} + P_4\alpha_k^2 N^{-\frac{1}{3}} \\ & \le -\frac{\alpha_k}{16} + \frac{2\left(L^H\right)^2\alpha_k^2}{\rho_k} + P_4\alpha_k^2 N^{-\frac{1}{3}} \\ & \le -\frac{\alpha_k}{32} + P_4\alpha_k^2 N^{-\frac{1}{3}} \\ & \le -\frac{c_1}{32}N^{-\frac{2}{3}} + c''c_1^2 N^{1-\frac{1}{3}-\frac{4}{3}} \\ & \le 0, \end{aligned}$$

where the justification for the six inequalities holding true are, respectively, $\alpha_k \le \frac{1}{2L^H}$, $\alpha_k \le \frac{\mu}{16L_{y^*}^2}\beta_k$, $\alpha_k \le \frac{\mu}{48L_{z^*}^2}\gamma_k$, $\alpha_k \le \frac{1}{64(L^H)^2}\rho_k$, $P_4 \le c''N$, and $c_1 \le \frac{1}{32c''}$.

- **Inequality 3:**

$$\begin{aligned}
&-A\beta_k\mu + 8\Delta B\gamma_k + 6C\Delta\rho_k + (P_1\gamma_k^2 + P_2\beta_k^2)D + 2B\gamma_k^2 L_z'' + C\rho_k^2 L_x'' \\
&= -\beta_k\mu + 8\Delta\gamma_k + 6\Delta\rho_k + (P_1\gamma_k^2 + P_2\beta_k^2)D + 2\gamma_k^2 L_z'' + \rho_k^2 L_x'' \\
&\leq -\frac{\beta_k\mu}{2} + 6\Delta\rho_k + (P_1\gamma_k^2 + P_2\beta_k^2)D + 2\gamma_k^2 L_z'' + \rho_k^2 L_x'' \\
&\leq -\frac{\beta_k\mu}{4} + (P_1\gamma_k^2 + P_2\beta_k^2)D + 2\gamma_k^2 L_z'' + \rho_k^2 L_x'' \\
&\leq -\frac{\beta_k\mu}{8} + 2\gamma_k^2 L_z'' + \rho_k^2 L_x'' \\
&\leq -\frac{\beta_k\mu}{16} + \rho_k^2 L_x'' \\
&\leq 0
\end{aligned}$$

where the justification for the five inequalities holding true are, respectively, $\gamma_k \leq \frac{\mu}{16\Delta}\beta_k$, $\rho_k \leq \frac{\mu}{24\Delta}\beta_k$ $c''(c_2^2 + c_3^2) \leq \frac{c_2\mu}{8}$, $\gamma_k^2 \leq \frac{\mu}{32L_z''}\beta_k$, and $\rho_k^2 \leq \frac{\mu}{16L_x''}\beta_k$.

To prevent any confusion, we additionally note that the second inequality arises because $P_1, P_2 \leq c''N$ and $c''(c_2^2 + c_3^2) \leq \frac{c_2\mu}{4}$ ensure that

$$\begin{aligned}
-\frac{\beta_k\mu}{4} + (P_1\gamma_k^2 + P_2\beta_k^2)D &\leq -\frac{c_2\mu}{4}N^{-\frac{2}{3}} + (c''c_3^2 + c''c_2^2)N^{1-\frac{4}{3}-\frac{1}{3}} \\
&= \left(-\frac{c_2\mu}{4} + (c''c_3^2 + c''c_2^2)\right)N^{-\frac{2}{3}} \\
&\leq -\frac{c_2\mu}{8}N^{-\frac{2}{3}} \\
&= -\frac{\beta_k\mu}{8}.
\end{aligned}$$

The condition $c''(c_2^2 + c_3^2) \leq \frac{c_2\mu}{8}$ is also reasonable. This can be achieved, for instance, by requiring that the coefficients of the step sizes adhere to $c_2 \leq \frac{\mu}{16c''}$ and $c_3^2 \leq \frac{\mu c_2}{16c''}$.

- **Inequality 4:**

$$\begin{aligned}
&-B\gamma_k\mu + 6C\left(L_1^g\right)^2\rho_k + P_3\gamma_k^2 D + 2B\gamma_k^2 L_z'' + C\rho_k^2 L_x'' \\
&= -\gamma_k\mu + 6\left(L_1^g\right)^2\rho_k + P_3\gamma_k^2 D + 2\gamma_k^2 L_z'' + \rho_k^2 L_x'' \\
&\leq -\frac{\gamma_k\mu}{2} + P_3\gamma_k^2 D + 2\gamma_k^2 L_z'' + \rho_k^2 L_x'' \\
&\leq -\frac{\gamma_k\mu}{4} + 2\gamma_k^2 L_z'' + \rho_k^2 L_x'' \\
&\leq -\frac{\gamma_k\mu}{8} + \rho_k^2 L_x'' \\
&\leq 0,
\end{aligned}$$

where the justification for the four inequalities holding true are, respectively, $\rho_k \leq \frac{\mu}{12(L_1^g)^2}\gamma_k$, $c_3 \leq \frac{\mu}{4c''}$, $\gamma_k \leq \frac{\mu}{16L_z''}$, and $\rho_k^2 \leq \frac{\mu}{8L_x''}\gamma_k$. For a complete proof, the detailed process by which the second inequality holds is as follows:

$$\begin{aligned}
-\frac{\gamma_k\mu}{2} + P_3\gamma_k^2 D &\leq -\frac{c_3\mu}{2}N^{-\frac{2}{3}} + c''c_3^2 N^{1-\frac{4}{3}-\frac{1}{3}} \\
&= \left(-\frac{c_3\mu}{2} + c''c_3^2\right)N^{-\frac{2}{3}} \\
&\leq -\frac{c_3\mu}{4}N^{-\frac{2}{3}} \\
&= -\frac{\gamma_k\mu}{4}.
\end{aligned}$$

- **Inequality 5:** Given that $C = 1$ and $\alpha_k \leq 2\rho_k$, we can affirm that the last inequality holds true, which is:

$$\frac{\alpha_k}{2} - C\rho_k = \frac{\alpha_k}{2} - \rho_k \leq 0.$$

Up to this point, we've confirmed that each inequality in the system holds.

Consequently, the inequality of the difference in the Lyapunov function can be simplified to

$$L_{k+1} - L_k \leq -\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right]$$

Summing and rearranging the above expressions yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] \leq \frac{L_0}{\alpha_k K} = \mathcal{O}\left(\frac{N^{2/3}}{K}\right).$$

□

F. Proof of Theorems3.7

Corollary F.1. *Suppose Assumptions 3.1 and 3.2 hold. Then we have*

$$\begin{aligned}\mathbb{E}\left[H\left(x_{k+1}\right)\right] \quad \leq \quad & \mathbb{E}\left[H\left(x_k\right)\right] - \frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] + \left(\frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2}\right)\mathbb{E}\left[\left\|v_k^x\right\|^2\right] + \alpha_k\mathbb{E}\left[\left\|D_x(x_k, y_k, z_k) - v_k^x\right\|^2\right] \\ & + 3\alpha_k\left(\left(L^f\right)^2 + \left(L_2^g R\right)^2\right)\mathbb{E}\left[\left\|y_k - y^*\left(x_k\right)\right\|^2\right] + 3\alpha_k\left(L_1^g\right)^2\mathbb{E}\left[\left\|z_k - z^*\left(x_k\right)\right\|^2\right]\end{aligned}$$

Proof. By combining Lemmas D.7 and D.6, the proof can be established. □

Lemma F.2. *If ϕ is α-strongly convex and β-smooth, then*

$$\langle\nabla\phi(x) - \nabla\phi(y), x - y\rangle \geq \frac{\alpha\beta}{\alpha+\beta}\|x-y\|^2 + \frac{1}{\alpha+\beta}\|\nabla\phi(x) - \nabla\phi(y)\|^2.$$

Proof. See Lemma C.2. in (Khanduri et al., 2021b). □

Lemma F.3. *Suppose Assumption 3.1 and 3.2 hold and the step sizes satisfy*

$$\beta_k, \gamma_k \leq \min\left\{\frac{\mu + L_1^g}{\mu L_1^g}, \frac{1}{\mu + L_1^g}\right\}.$$

Then we have

$$\begin{aligned}\mathbb{E}\left[\left\|y_{k+1} - y^*\left(x_{k+1}\right)\right\|^2\right] \leq & \left(1 - \frac{\mu L_1^g\beta_k}{2(\mu + L_1^g)}\right)\mathbb{E}\left[\left\|y_k - y^*\left(x_k\right)\right\|^2\right] - \frac{1}{\mu + L_1^g}\beta_k\mathbb{E}\left[\left\|D_y(x_k, y_k, z_k)\right\|^2\right] \\ & + 4\frac{\mu + L_1^g}{\mu L_1^g}\beta_k^2\mathbb{E}\left[\left\|v_k^y - D_y(x_k, y_k, z_k)\right\|^2\right] \\ & + \frac{3(\mu + L_1^g)L_{y^*}^2\alpha_k^2}{\mu L_1^g\beta_k}\mathbb{E}\left[\left\|v_k^x\right\|^2\right].\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left\|z_{k+1} - z^*\left(x_{k+1}\right)\right\|^2\right] \leq & \left(1 - \frac{\mu L_1^g\gamma_k}{2(\mu + L_1^g)}\right)\mathbb{E}\left[\left\|z_k - z^*\left(x_k\right)\right\|^2\right] - \frac{1}{\mu + L_1^g}\gamma_k\mathbb{E}\left[\left\|D_z(x_k, y_k, z_k)\right\|^2\right] \\ & + 4\frac{\mu + L_1^g}{\mu L_1^g}\gamma_k^2\mathbb{E}\left[\left\|v_k^z - D_z(x_k, y_k, z_k)\right\|^2\right] \\ & + \frac{3(\mu + L_1^g)L_{z^*}^2\alpha_k^2}{\mu L_1^g\gamma_k}\mathbb{E}\left[\left\|v_k^x\right\|^2\right].\end{aligned}$$

The proof of this lemma is similar to that of Lemma E.2. The main difference is that $v_k^{\cdot}$ is no longer an unbiased estimate of $D_{\cdot}(x_k, y_k, z_k)$. Below, we present the specific proof process.

Proof. **Inequality for y.**

By utilizing the Young's inequality and the L_{y^*}-Lipschitz continuity of $y^*(x)$, we have

$$\begin{aligned}\|y_{k+1} - y^*(x_{k+1})\|^2 &= \|y_{k+1} - y^*(x_k) + y^*(x_k) - y^*(x_{k+1})\|^2 \\ &\leq (1+\delta_k)\|y_{k+1} - y^*(x_k)\|^2 + \left(1 + \frac{1}{\delta_k}\right)\|y^*(x_k) - y^*(x_{k+1})\|^2 \\ &\leq (1+\delta_k)\|y_{k+1} - y^*(x_k)\|^2 + \left(1 + \frac{1}{\delta_k}\right) L_{y^*}^2 \alpha_k^2 \|v_k^x\|^2\end{aligned}$$

Taking the expectation conditionally on x_k, y_k, z_k yields

$$E_k\left[\|y_{k+1} - y^*(x_{k+1})\|^2\right] \leq (1+\delta_k) E_k\left[\|y_{k+1} - y^*(x_k)\|^2\right] + \left(1 + \frac{1}{\delta_k}\right) L_{y^*}^2 \alpha_k^2 E_k[\|v_k^x\|^2]. \tag{25}$$

For the first term, once again employing Young's inequality, we have

$$\begin{aligned}E_k\left[\|y_{k+1} - y^*(x_k)\|^2\right] &= E_k\left[\|y_k - y^*(x_k) - \beta_k v_k^y\|^2\right] \\ &= E_k\left[\|y_k - \beta_k D_y(x_k, y_k, z_k) - y^*(x_k) - \beta_k\left(v_k^y - D_y(x_k, y_k, z_k)\right)\|^2\right] \\ &\leq (1+2\delta_k) E_k\left[\|y_k - \beta_k D_y(x_k, y_k, z_k) - y^*(x_k)\|^2\right] \\ &\quad + \left(1 + \frac{1}{2\delta_k}\right) E_k\left[\|\beta_k\left(v_k^y - D_y(x_k, y_k, z_k)\right)\|^2\right],\end{aligned}$$

Utilizing Lemma F.2, we can thus establish the following inequality

$$\begin{aligned}E_k\left[\|y_k - \beta_k D_y(x_k, y_k, z_k) - y^*(x_k)\|^2\right] =& E_k\left[\|y_k - y^*(x_k)\|^2\right] + E_k\left[\|\beta_k D_y(x_k, y_k, z_k)\|^2\right] \\ & - 2E_k\beta_k\langle D_y(x_k, y_k, z_k), y_k - y^*(x_k)\rangle \\ \leq& \left(1 - 2\beta_k\frac{\mu L_1^g}{\mu + L_1^g}\right) E_k\left[\|y_k - y^*(x_k)\|^2\right] \\ & + \left(\beta_k^2 - 2\beta_k\frac{1}{\mu + L_1^g}\right) E_k\left[\|D_y(x_k, y_k, z_k)\|^2\right].\end{aligned}$$

Plugging it into (25) and taking the total expectation, we have

$$\begin{aligned}\mathbb{E}\left[\|y_{k+1} - y^*(x_{k+1})\|^2\right] \leq& (1+\delta_k)(1+2\delta_k)\left(1 - 2\beta_k\frac{\mu L_1^g}{\mu + L_1^g}\right)\mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right] \\ & + (1+\delta_k)(1+2\delta_k)\left(\beta_k^2 - 2\beta_k\frac{1}{\mu + L_1^g}\right)\mathbb{E}\left[\|D_y(x_k, y_k, z_k)\|^2\right] \\ & + (1+\delta_k)\left(1 + \frac{1}{2\delta_k}\right)\beta_k^2\mathbb{E}\left[\|v_k^y - D_y(x_k, y_k, z_k)\|^2\right] \\ & + \left(1 + \frac{1}{\delta_k}\right) L_{y^*}^2 \alpha_k^2 \mathbb{E}\left[\|v_k^x\|^2\right],\end{aligned}$$

We choose the parameter δ_k and the step size β_k to satisfy

$$\delta_k = \frac{\mu L_1^g}{\mu + L_1^g}\beta_k, \quad \beta_k \leq \min\left\{\frac{\mu + L_1^g}{\mu L_1^g}, \frac{1}{\mu + L_1^g}\right\},$$

Consequently, the lemma concerning y is proven.

Inequality for z.

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

$$\begin{aligned}\left\|z_{k+1}-z^*\left(x_{k+1}\right)\right\|^2 &= \left\|z_{k+1}-z^*\left(x_k\right)+z^*\left(x_k\right)-z^*\left(x_{k+1}\right)\right\|^2 \\ &\leq \left(1+\delta_k'\right)\left\|z_{k+1}-z^*\left(x_k\right)\right\|^2+\left(1+\frac{1}{\delta_k'}\right)\left\|z^*\left(x_k\right)-z^*\left(x_{k+1}\right)\right\|^2 \\ &\leq \left(1+\delta_k'\right)\left\|z_{k+1}-z^*\left(x_k\right)\right\|^2+\left(1+\frac{1}{\delta_k'}\right)L_{z^*}^2\alpha_k^2\left\|v_k^x\right\|^2,\end{aligned}$$

where based on the definition of z_{k+1} and the contractivity of projection, we have

$$\begin{aligned}\left\|z_{k+1}-z^*\left(x_k\right)\right\|^2 &= \left\|\operatorname{Proj}_{\mathbb{B}(R)}(z_k-\gamma_k v_k^z)-z^*\left(x_k\right)\right\|^2=\left\|\operatorname{Proj}_{\mathbb{B}(R)}(z_k-\gamma_k v_k^z)-\operatorname{Proj}_{\mathbb{B}(R)}(z^*\left(x_k\right))\right\|^2 \\ &\leq \left\|z_k-\gamma_k v_k^z-z^*\left(x_k\right)\right\|^2=\left\|z_k-\gamma_k D_z(x_k,y_k,z_k)-z^*\left(x_k\right)-\gamma_k\left(v_k^z-D_z(x_k,y_k,z_k)\right)\right\|^2 \\ &\leq \left(1+2\delta_k'\right)\left\|z_k-\gamma_k D_z(x_k,y_k,z_k)-z^*\left(x_k\right)\right\|^2 \\ &\quad +\left(1+\frac{1}{2\delta_k'}\right)\left\|\gamma_k\left(v_k^z-D_z(x_k,y_k,z_k)\right)\right\|^2\end{aligned}$$

This leads to the result when substituted back into the previous equation

$$\begin{aligned}\left\|z_{k+1}-z^*\left(x_{k+1}\right)\right\|^2 \leq\ & \left(1+\delta_k'\right)\left(1+2\delta_k'\right)\left\|z_k-\gamma_k D_z(x_k,y_k,z_k)-z^*\left(x_k\right)\right\|^2 \\ &+\left(1+\delta_k'\right)\left(1+\frac{1}{2\delta_k'}\right)\left\|\gamma_k\left(v_k^z-D_z(x_k,y_k,z_k)\right)\right) \\ &+\left(1+\frac{1}{\delta_k'}\right)L_{z^*}^2\alpha_k^2\left\|v_k^x\right\|^2,\end{aligned}$$

For the first term, since the function $\phi(z)=\frac{1}{2}\langle\nabla_{22}^2 g(x,y)z,z\rangle-\langle\nabla_2 f(x,y),z\rangle$ is L_1^g-strongly convex and μ-smooth, we have

$$\begin{aligned}&\left\|z_k-\gamma_k D_z(x_k,y_k,z_k)-z^*\left(x_k\right)\right\|^2 \\ &=\left\|z_k-z^*\left(x_k\right)\right\|^2+\gamma_k^2\left\|D_z(x_k,y_k,z_k)\right\|^2-\gamma_k\left\langle z_k-z^*\left(x_k\right),D_z(x_k,y_k,z_k)\right\rangle \\ &\leq\left(1-2\gamma_k\frac{\mu L_1^g}{\mu+L_1^g}\right)\left\|z_k-z^*\left(x_k\right)\right\|^2+\left(\gamma_k^2-\frac{2\gamma_k}{\mu+L_1^g}\right)\left\|D_z(x_k,y_k,z_k)\right\|^2,\end{aligned}$$

Rearranging the above inequalities and taking the totaol expectation yields.

$$\begin{aligned}\mathbb{E}\left[\left\|z_{k+1}-z^*\left(x_{k+1}\right)\right\|^2\right]\leq&\left(1+\delta_k'\right)\left(1+2\delta_k'\right)\left(1-2\gamma_k\frac{\mu L_1^g}{\mu+L_1^g}\right)\mathbb{E}\left[\left\|z_k-z^*\left(x_k\right)\right\|^2\right] \\ &+\left(1+\delta_k'\right)\left(1+2\delta_k'\right)\left(\gamma_k^2-2\gamma_k\frac{1}{\mu+L_1^g}\right)\mathbb{E}\left[\left\|D_z(x_k,y_k,z_k)\right\|^2\right] \\ &+\left(1+\delta_k'\right)\left(1+\frac{1}{2\delta_k'}\right)\gamma_k^2\mathbb{E}\left[\left\|v_k^z-D_z(x_k,y_k,z_k)\right\|^2\right] \\ &+\left(1+\frac{1}{\delta_k'}\right)L_{z^*}^2\alpha_k^2\mathbb{E}\left[\left\|v_k^x\right\|^2\right],\end{aligned}$$

We choose the parameter δ_k' and the step size γ_k to satisfy

$$\delta_k'=\frac{\mu L_1^g}{\mu+L_1^g}\gamma_k,\quad \gamma_k\leq\min\left\{\frac{\mu+L_1^g}{\mu L_1^g},\frac{1}{\mu+L_1^g}\right\},$$

Consequently, the lemma concerning z is proven. □

Lemma F.4. *Under the Assumption 3.1, 3.2, 3.3 and 3.4, We have the following inequalities established:*

$$\begin{aligned}
(1)\quad \mathbb{E}\left[\left\|v_{k+1}^{y}-D_y\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^2\right] \leq &(1-p)\mathbb{E}\left[\|v_k^y-D_y(x_k,y_k,z_k)\|^2\right]+\frac{(1-p)}{b}(L_1^g)^2\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right]\\
&+\frac{2(1-p)}{b}(L_1^g)^2\beta_k^2\mathbb{E}\left[\|v_k^y-D_y(x_k,y_k,z_k)\|^2\right]\\
&+\frac{2(1-p)}{b}(L_1^g)^2\beta_k^2\left(L^f\right)^2\mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right].
\end{aligned}$$

$$\begin{aligned}
(2)\quad \mathbb{E}\left[\left\|v_{k+1}^{x}-D_x\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^2\right] \leq &(1-p)\mathbb{E}\left[\|v_k^x-D_x(x_k,y_k,z_k)\|^2\right]\\
&+\left(2(L^f)^2+4R^2(L_2^g)^2\right)\frac{(1-p)}{b}\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right]\\
&+\left(2(L^f)^2+4R^2(L_2^g)^2\right)\frac{(1-p)}{b}\beta_k^2\mathbb{E}\left[\|v_k^y-D_y(x_k,y_k,z_k)\|^2\right]\\
&+\left(2(L^f)^2+4R^2(L_2^g)^2\right)\frac{(1-p)}{b}\beta_k^2\left(L^f\right)^2\mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right]\\
&+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2\mathbb{E}\left[\|v_k^z-D_z(x_k,y_k,z_k)\|^2\right]\\
&+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2L_z^2\mathbb{E}\left[\|z_k-z^*\left(x_k\right)\|^2\right]\\
&+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2L_z^2\mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right]
\end{aligned}$$

$$\begin{aligned}
(3)\quad \mathbb{E}\left[\left\|v_{k+1}^{z}-D_z\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^2\right] \leq &(1-p+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2)\mathbb{E}\left[\|v_k^z-D_z(x_k,y_k,z_k)\|^2\right]\\
&+\left(2(L^f)^2+4R^2(L_2^g)^2\right)\frac{(1-p)}{b}\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right]\\
&+\left(2(L^f)^2+4R^2(L_2^g)^2\right)\frac{(1-p)}{b}\beta_k^2\mathbb{E}\left[\|v_k^y-D_y(x_k,y_k,z_k)\|^2\right]\\
&+\left(2(L^f)^2+4R^2(L_2^g)^2\right)\frac{(1-p)}{b}\beta_k^2\left(L^f\right)^2\mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right]\\
&+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2L_z^2\mathbb{E}\left[\|z_k-z^*\left(x_k\right)\|^2\right]\\
&+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2L_z^2\mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right]
\end{aligned}$$

Proof. **Proof of (1).**

By the definition of v_{k+1}^y, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|v_{k+1}^y - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\
=&p\mathbb{E}\left[\left\|\frac{1}{m}\sum_{j\in[m]}\nabla_2 G_j\left(x_{k+1}, y_{k+1}\right) - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\
&+(1-p)\mathbb{E}\left[\left\|v_k^y + \frac{1}{b}\sum_{j\in J}\left[\nabla_2 G_j\left(x_{k+1}, y_{k+1}\right) - \nabla_2 G_j\left(x_k, y_k\right)\right] - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\
=&(1-p)\mathbb{E}\left[\left\|v_k^y + \frac{1}{b}\sum_{j\in J}\left[\nabla_2 G_j\left(x_{k+1}, y_{k+1}\right) - \nabla_2 G_j\left(x_k, y_k\right)\right] - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\
=&(1-p)\mathbb{E}\left[\left\|v_k^y - D_y(x_k, y_k, z_k) + \frac{1}{b}\sum_{j\in J}\left[\nabla_2 G_j\left(x_{k+1}, y_{k+1}\right) - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right) + D_y(x_k, y_k, z_k) - \nabla_2 G_j\left(x_k, y_k\right)\right]\right\|^2\right] \\
=&(1-p)\mathbb{E}\left[\left\|v_k^y - D_y(x_k, y_k, z_k)\right\|^2\right] \\
&+(1-p)\mathbb{E}\left[\left\|\frac{1}{b}\sum_{j\in J}\left(\nabla_2 G_j\left(x_{k+1}, y_{k+1}\right) - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right) + D_y(x_k, y_k, z_k) - \nabla_2 G_j\left(x_k, y_k\right)\right)\right\|^2\right] \\
\leq&(1-p)\mathbb{E}\left[\left\|v_k^y - D_y(x_k, y_k, z_k)\right\|^2\right] + (1-p)\mathbb{E}\left[\left\|\frac{1}{b}\sum_{j\in J}\left(\nabla_2 G_j\left(x_{k+1}, y_{k+1}\right) - \nabla_2 G_j\left(x_k, y_k\right)\right)\right\|^2\right],
\end{aligned}
$$

where the last equation uses the fact that

$$
\mathbb{E}\left[\frac{1}{b}\sum_{j\in J}\nabla_2 G_j\left(x_{k+1}, y_{k+1}\right)\right] = D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right), \quad \mathbb{E}\left[\frac{1}{b}\sum_{j\in J}\nabla_2 G_j\left(x_k, y_k\right)\right] = D_y\left(x_k, y_k, z_k\right).
$$

The final inequality arises due to $\mathbb{E}\|X - \mathbb{E}[X]\|^2 \leq \mathbb{E}[X^2]$. Additionally, utilizing Assumption 3.4, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left\|v_{k+1}^y - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\
\leq&(1-p)\mathbb{E}\left[\left\|v_k^y - D_y(x_k, y_k, z_k)\right\|^2\right] + \frac{(1-p)}{b}(L_1^g)^2\left(\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right] + \beta_k^2\mathbb{E}\left[\|v_k^y\|^2\right]\right) \\
\leq&(1-p)\mathbb{E}\left[\left\|v_k^y - D_y(x_k, y_k, z_k)\right\|^2\right] + \frac{(1-p)}{b}(L_1^g)^2\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right] \\
&+\frac{2(1-p)}{b}(L_1^g)^2\beta_k^2\mathbb{E}\left[\left\|v_k^y - D_y(x_k, y_k, z_k)\right\|^2\right] + \frac{2(1-p)}{b}(L_1^g)^2\beta_k^2\left(L^f\right)^2\mathbb{E}\left[\left\|y_k - y^*\left(x_k\right)\right\|^2\right].
\end{aligned}
$$

Proof of (2) and (3). Similarly, from the definition of v_{k+1}^x, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|v_{k+1}^x - D_x\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\
=&(1-p)\mathbb{E}\left[\left\|v_k^x + \frac{1}{b}\sum_{i\in I}\left(\nabla_1 F_i\left(x_{k+1}, y_{k+1}\right) - \nabla_1 F_i\left(x_k, y_k\right)\right)\right.\right. \\
&\left.\left.-\frac{1}{b}\sum_{j\in J}\left(\nabla_{12}^2 G_j\left(x_{k+1}, y_{k+1}\right) z_{k+1-}\nabla_{12}^2 G_j\left(x_k, y_k\right) z_k\right) - D_x\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right].
\end{aligned}
$$

Based on the fact that

$$\mathbb{E}\left[\frac{1}{b}\sum_{i\in I}\nabla_1 F_i\left(x_{k+1},y_{k+1}\right)-\frac{1}{b}\sum_{j\in J}\nabla_{12}^2 G_j\left(x_{k+1},y_{k+1}\right)z_{k+1}\right]=D_x(x_{k+1},y_{k+1},z_{k+1}),$$

$$\mathbb{E}\left[\frac{1}{b}\sum_{i\in I}\nabla_1 F_i\left(x_k,y_k\right)-\frac{1}{b}\sum_{j\in J}\nabla_{12}^2 G_j\left(x_k,y_k\right)z_k\right]=D_x(x_k,y_k,z_k),$$

we deduce

$$\begin{aligned}
&\mathbb{E}\left[\left\|v_{k+1}^x-D_x\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^2\right]\\
=&(1-p)\mathbb{E}\left[\|v_k^x-D_x(x_k,y_k,z_k)\|^2\right]+2(1-p)\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I}\left[\nabla_1 F_i\left(x_{k+1},y_{k+1}\right)-\nabla_1 F_i\left(x_k,y_k\right)\right]\right\|^2\right]\\
&+2(1-p)\mathbb{E}\left[\left\|\frac{1}{b}\sum_{j\in J}[\nabla_{12}^2 G_j\left(x_{k+1},y_{k+1}\right)z_{k+1}-\nabla_{12}^2 G_j\left(x_k,y_k\right)z_k]\right\|^2\right]\\
=&(1-p)\mathbb{E}\left[\|v_k^x-D_x(x_k,y_k,z_k)\|^2\right]+2(1-p)\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in I}\left[\nabla_1 F_i\left(x_{k+1},y_{k+1}\right)-\nabla_1 F_i\left(x_k,y_k\right)\right]\right\|^2\right]\\
&+4(1-p)\mathbb{E}\left[\left\|\frac{1}{b}\sum_{j\in J}\left[\nabla_{12}^2 G_j\left(x_{k+1},y_{k+1}\right)z_{k+1}-\nabla_{12}^2 G_j\left(x_k,y_k\right)z_{k+1}\right]\right\|^2\right]\\
&+4(1-p)\mathbb{E}\left[\left\|\frac{1}{b}\sum_{j\in J}\left[\nabla_{12}^2 G_j\left(x_k,y_k\right)z_{k+1}-\nabla_{12}^2 G_j\left(x_k,y_k\right)z_k\right]\right\|^2\right].
\end{aligned}$$

Under Assumption 3.4, it further implies that

$$\begin{aligned}
&\mathbb{E}\left[\left\|v_{k+1}^x-D_x\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^2\right]\\
=&(1-p)\mathbb{E}\left[\|v_k^x-D_x(x_k,y_k,z_k)\|^2\right]+\frac{2(1-p)}{b}(L^f)^2\left(\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right]+\beta_k^2\mathbb{E}\left[\|v_k^y\|^2\right]\right)\\
&+\frac{4(1-p)}{b}R^2(L_2^g)^2\left(\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right]+\beta_k^2\mathbb{E}\left[\|v_k^y\|^2\right]\right)+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2\mathbb{E}\left[\|v_k^z\|^2\right]\\
=&(1-p)\mathbb{E}\left[\|v_k^x-D_x(x_k,y_k,z_k)\|^2\right]+\left(\frac{2(1-p)}{b}(L^f)^2+\frac{4(1-p)}{b}R^2(L_2^g)^2\right)\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right]\\
&+\left(\frac{2(1-p)}{b}(L^f)^2+\frac{4(1-p)}{b}R^2(L_2^g)^2\right)\beta_k^2\mathbb{E}\left[\|v_k^y\|^2\right]+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2\mathbb{E}\left[\|v_k^z\|^2\right]\\
\leq&(1-p)\mathbb{E}\left[\|v_k^x-D_x(x_k,y_k,z_k)\|^2\right]+\left(\frac{2(1-p)}{b}(L^f)^2+\frac{4(1-p)}{b}R^2(L_2^g)^2\right)\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right]\\
&+\left(\frac{2(1-p)}{b}(L^f)^2+\frac{4(1-p)}{b}R^2(L_2^g)^2\right)\beta_k^2\mathbb{E}\left[\|v_k^y-D_y(x_k,y_k,z_k)\|^2\right]\\
&+\left(\frac{2(1-p)}{b}(L^f)^2+\frac{4(1-p)}{b}R^2(L_2^g)^2\right)\beta_k^2\left(L^f\right)^2\mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right]\\
&+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2\mathbb{E}\left[\|v_k^z-D_z(x_k,y_k,z_k)\|^2\right]+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2L_z^2\mathbb{E}\left[\|z_k-z^*\left(x_k\right)\|^2\right]\\
&+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2L_z^2\mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right].
\end{aligned}$$

The proof of (3) is analogous, and hence we omit the details here. □

Theorem F.5. ***(Restatement of Theorem 3.7)***

Fix an iteration $K > 1$ and assume that Assumption 3.1 to 3.2 and 3.4 hold. Choose minibatch size $b < (n + m)$ and the probability $p \in (0, 1]$. Then there exist positive constants c, c_β, and c_γ, such that if

$$\alpha_k \leq \frac{c}{1 + \sqrt{\frac{1-p}{pb}}}, \quad \beta_k = c_\beta \alpha_k, \quad \gamma_k = c_\gamma \alpha_k,$$

the iterates in SPABA satisfy

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] = \mathcal{O}\left(\frac{1 + \sqrt{\frac{1-p}{pb}}}{K}\right).$$

Proof. We consider the Lyapunov function

$$\begin{aligned} L_k =& H_k + \mathbb{E}\left[\left\|y_k - y^*\left(x_k\right)\right\|^2\right] + \mathbb{E}\left[\left\|z_k - z^*\left(x_k\right)\right\|^2\right] \\ &+ \frac{\alpha_k}{p}\left(\mathbb{E}\left[\left\|v_k^x - D_x(x_k, y_k, z_k)\right\|^2\right] + \mathbb{E}\left[\left\|v_k^y - D_y(x_k, y_k, z_k)\right\|^2\right] + \mathbb{E}\left[\left\|v_k^z - D_z(x_k, y_k, z_k)\right\|^2\right]\right) \end{aligned}$$

$$\begin{aligned} & L_{k+1} - L_k \\ \leq & -\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] \\ & + (\alpha_k - \alpha_k)\,\mathbb{E}\left[\left\|D_x(x_k, y_k, z_k) - v_k^x\right\|^2\right] \\ & + \left\{\frac{L^H \alpha_k^2}{2} - \frac{\alpha_k}{2} + \frac{3(\mu + L_1^g)L_{y^*}^2 \alpha_k^2}{\mu L_1^g \beta_k} + \frac{3(\mu + L_1^g)L_{z^*}^2 \alpha_k^2}{\mu L_1^g \gamma_k} + \frac{(1-p)}{bp}(L_1^g)^2 \alpha_k^3 \right. \\ & \quad \left. + \left(4(L^f)^2 + 8R^2(L_2^g)^2\right)\frac{(1-p)}{pb}\alpha_k^3\right\} \mathbb{E}\left[\left\|v_k^x\right\|^2\right] \\ & + \left\{3\alpha_k\left(\left(L^f\right)^2 + (L_2^g R)^2\right) - \frac{\mu L_1^g \gamma_k}{2(\mu + L_1^g)} + \frac{2(1-p)}{bp}(L_1^g)^2 \alpha_k \beta_k^2 \left(L^f\right)^2 \right. \\ & \quad \left. + \left(4(L^f)^2 + 8R^2(L_2^g)^2\right)\frac{(1-p)}{bp}\alpha_k \beta_k^2 \left(L^f\right)^2 + \frac{8(1-p)}{bp}(L_2^g)^2 \alpha_k \gamma_k^2 L_z^2\right\} \mathbb{E}\left[\left\|y_k - y^*\left(x_k\right)\right\|^2\right] \\ & + \left\{3\alpha_k\left(L_1^g\right)^2 - \frac{\mu L_1^g \gamma_k}{2(\mu + L_1^g)} + \frac{8(1-p)}{bp}(L_2^g)^2 \alpha_k \gamma_k^2 L_z^2\right\} \mathbb{E}\left[\left\|z_k - z^*\left(x_k\right)\right\|^2\right] \\ & + \left\{4\frac{\mu + L_1^g}{\mu L_1^g}\beta_k^2 - \alpha_k + \frac{2(1-p)}{bp}(L_1^g)^2 \alpha_k \beta_k^2 + \left(4(L^f)^2 + 8R^2(L_2^g)^2\right)\frac{(1-p)}{bp}\alpha_k \beta_k^2 \right. \\ & \quad \left. + \left(2(L^f)^2 + 4R^2(L_2^g)^2\right)\frac{(1-p)}{bp}\alpha_k \beta_k^2\right\} \mathbb{E}\left[\|v_k^y - D_y(x_k, y_k, z_k)\|^2\right] \\ & + \left\{8\frac{\mu + L_1^g}{\mu L_1^g}\gamma_k^2 + \frac{4(1-p)}{bp}(L_2^g)^2 \alpha_k \gamma_k^2 - \alpha_k\right\} \mathbb{E}\left[\|v_k^z - D_z(x_k, y_k, z_k)\|^2\right] \end{aligned}$$

We choose the step sizes to be

$$\alpha_k = \min\left\{\frac{1}{4L^H}, \frac{c_\alpha}{\sqrt{\frac{1-p}{bp}}}\right\}, \quad \beta_k = c_\beta \alpha_k, \quad \gamma_k = c_\gamma \alpha_k.$$

Furthermore, by analyzing the coefficients of each term in the aforementioned inequalities, we can determine the range of values for c_α, c_β and c_γ.

Analysis of the Coefficient for $\mathbb{E}\left[\|v_k^x\|^2\right]$

By assuming

$$\alpha_k \leq \min\left\{\frac{1}{4L^H}, \frac{\mu L_1^g}{48(\mu+L_1^g)L_{y^*}}\beta_k\right\}, \quad c_\alpha^2 \leq \frac{1}{2\Delta_1},$$

we can deduce

$$\begin{aligned}
&\frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2} + \frac{3(\mu+L_1^g)L_{y^*}^2\alpha_k^2}{\mu L_1^g\beta_k} + \frac{3(\mu+L_1^g)L_{z^*}^2\alpha_k^2}{\mu L_1^g\gamma_k} + \frac{(1-p)}{bp}(L_1^g)^2\alpha_k^3 + \left(4(L^f)^2+8R^2(L_2^g)^2\right)\frac{(1-p)}{pb}\alpha_k^3 \\
=&\frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2} + \frac{3(\mu+L_1^g)L_{y^*}^2\alpha_k^2}{\mu L_1^g\beta_k} + \frac{(1-p)}{pb}\alpha_k^3\left((L_1^g)^2+4(L^f)^2+8R^2(L_2^g)^2\right) \\
\triangleq&\frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2} + \frac{3(\mu+L_1^g)L_{y^*}^2\alpha_k^2}{\mu L_1^g\beta_k} + \frac{(1-p)}{pb}\alpha_k^3\Delta_1^2 \\
\leq&0.
\end{aligned}$$

Analysis of the Coefficient for $\mathbb{E}\left[\|y_k - y^*(x_k)\|^2\right]$

By assuming

$$\alpha_k \leq \frac{\mu L_1^g\gamma_k}{12(\mu+L_1^g)\left(\left(L^f\right)^2+\left(L_2^gR\right)^2\right)}, \quad c_\alpha^2 \leq \frac{\mu L_1^g c_\gamma}{8(\mu+L_1^g)\Delta_2},$$

we can deduce

$$\begin{aligned}
&-\frac{\mu L_1^g\gamma_k}{2(\mu+L_1^g)} + 3\alpha_k\left(\left(L^f\right)^2+\left(L_2^gR\right)^2\right) \\
&+\frac{(1-p)}{bp}\alpha_k^3\left(\left(2(L_1^g)^2c_{\beta_k}^2\left(L^f\right)^2\right)+\left(4(L^f)^2+8R^2(L_2^g)^2\right)\left(L^f\right)^2c_{\beta_k}^2+8(L_2^g)^2L_z^2\right) \\
\triangleq&-\frac{\mu L_1^g\gamma_k}{2(\mu+L_1^g)} + 3\alpha_k\left(\left(L^f\right)^2+\left(L_2^gR\right)^2\right) + \frac{(1-p)}{bp}\alpha_k^3\Delta_2^2 \\
&\leq 0.
\end{aligned}$$

Analysis of the Coefficient for $\mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right]$

By assuming

$$\alpha_k \leq \frac{\mu L_1^g\gamma_k}{12(\mu+L_1^g)\left(L_1^g\right)^2}, \quad c_\alpha^2 c_\gamma \leq \frac{\mu L_1^g}{32(\mu+L_1^g)L_z^2(L_2^g)^2,}$$

we can deduce

$$3\alpha_k\left(L_1^g\right)^2 - \frac{\mu L_1^g\gamma_k}{2(\mu+L_1^g)} + \frac{8(1-p)}{bp}(L_2^g)^2\alpha_k\gamma_k^2L_z^2 \leq 0.$$

Analysis of the Coefficient for $\mathbb{E}\left[\|v_k^y - D_y(x_k,y_k,z_k)\|^2\right]$

By assuming

$$\beta_k^2 \leq \frac{\mu L_1^g}{8(\mu+L_1^g)}\alpha_k, \quad c_\alpha^2 \leq \frac{1}{4\Delta_3},$$

we can deduce

$$\begin{aligned}
&4\frac{\mu+L_1^g}{\mu L_1^g}\beta_k^2 - \alpha_k + \frac{2(1-p)}{bp}(L_1^g)^2\alpha_k\beta_k^2 + \left(4(L^f)^2+8R^2(L_2^g)^2\right)\frac{(1-p)}{bp}\alpha_k\beta_k^2 \\
&\quad+\left(2(L^f)^2+4R^2(L_2^g)^2\right)\frac{(1-p)}{bp}\alpha_k\beta_k^2 \\
&=4\frac{\mu+L_1^g}{\mu L_1^g}\beta_k^2 - \alpha_k + \frac{(1-p)}{bp}\alpha_k^3\Delta_3 \leq 0,
\end{aligned}$$

where

$$\Delta_3 = \left(2(L_1^g)^2 + 6(L^f)^2 + 12R^2(L_2^g)^2\right) c_\beta^2.$$

Analysis of the Coefficient for $\mathbb{E}\left[\|v_k^z - D_z(x_k, y_k, z_k)\|^2\right]$

By assuming

$$\gamma_k^2 \leq \frac{\mu L_1^g}{16(\mu + L_1^g)}\alpha_k, \quad c_\alpha^2 c_\gamma \leq \frac{1}{16(L_2^g)^2},$$

we can deduce

$$8\frac{\mu + L_1^g}{\mu L_1^g}\gamma_k^2 + \frac{4(1-p)}{bp}(L_2^g)^2\alpha_k\gamma_k^2 - \alpha_k \leq 0.$$

Thus, we have obtained the recursive inequality for this theorem

Summing, taking the average, and rearranging, we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla H(x_k)\|^2\right] \leq \frac{L_0}{K\alpha_k}.$$

From the above analysis, the step size α_k should satisfy

$$\alpha_k = \min\left\{\frac{1}{c'_\alpha}, \frac{c_\alpha}{\sqrt{\frac{1-p}{bp}}}\right\}$$

then we have

$$\alpha_k \leq \frac{c}{1 + \sqrt{\frac{1-p}{pb}}},$$

for some constants c. Therefore, we ultimately arrive at the conclusion that

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla H(x_k)\|^2\right] = \mathcal{O}\left(\frac{1 + \sqrt{\frac{1-p}{pb}}}{K}\right)$$

□

Corollary F.6. *Suppose that Assumption 3.1 to Assumption 3.4 hold. If we take $p = b/(n+m+b)$, and $b \leq \sqrt{n+m}$, then the sample complexity is $\mathcal{O}((n+m)^{1/2}\epsilon^{-1})$.*

Proof. In each iteration, it uses $p(n+m) + (1-p)b$ samples on expectation. Let $p = \frac{b}{n+m+b}$ and $b \leq (n+m)^{1/2}$. Thus, the total sample complexity is

$$K(p(n+m) + (1-p)b) = \mathcal{O}\left(\left(1 + \frac{\sqrt{n+m}}{b}\right)\frac{2(n+m)b}{n+m+b}\epsilon^{-1}\right) = \mathcal{O}\left((n+m)^{1/2}\epsilon^{-1}\right).$$

□

G. Proof of Theorem 3.9

Under the expected form setting, in the algorithm, we set $n = m = \tau'$, which represents the mini-batch size.

Lemma G.1. *Under the Assumption 3.1, 3.2, 3.3 and 3.4, We have the following inequalities established:*

$$
\begin{aligned}
(1)\quad \mathbb{E}\left[\left\|v_{k+1}^{y}-D_{y}\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^{2}\right] \leq & (1-p)\mathbb{E}\left[\|v_{k}^{y}-D_{y}(x_{k},y_{k},z_{k})\|^{2}\right]+\frac{(1-p)}{b}(L_{1}^{g})^{2}\alpha_{k}^{2}\mathbb{E}\left[\|v_{k}^{x}\|^{2}\right] \\
& +\frac{2(1-p)}{b}(L_{1}^{g})^{2}\beta_{k}^{2}\mathbb{E}\left[\|v_{k}^{y}-D_{y}(x_{k},y_{k},z_{k})\|^{2}\right] \\
& +\frac{2(1-p)}{b}(L_{1}^{g})^{2}\beta_{k}^{2}\left(L^{f}\right)^{2}\mathbb{E}\left[\|y_{k}-y^{*}\left(x_{k}\right)\|^{2}\right] \\
& +\frac{p\sigma_{g,1}^{2}}{\tau'}.
\end{aligned}
$$

$$
\begin{aligned}
(2)\quad \mathbb{E}\left[\left\|v_{k+1}^{x}-D_{x}\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^{2}\right] \leq & (1-p)\mathbb{E}\left[\|v_{k}^{x}-D_{x}(x_{k},y_{k},z_{k})\|^{2}\right] \\
& +\left(2(L^{f})^{2}+4R^{2}(L_{2}^{g})^{2}\right)\frac{(1-p)}{b}\alpha_{k}^{2}\mathbb{E}\left[\|v_{k}^{x}\|^{2}\right] \\
& +\left(2(L^{f})^{2}+4R^{2}(L_{2}^{g})^{2}\right)\frac{(1-p)}{b}\beta_{k}^{2}\mathbb{E}\left[\|v_{k}^{y}-D_{y}(x_{k},y_{k},z_{k})\|^{2}\right] \\
& +\left(2(L^{f})^{2}+4R^{2}(L_{2}^{g})^{2}\right)\frac{(1-p)}{b}\beta_{k}^{2}\left(L^{f}\right)^{2}\mathbb{E}\left[\|y_{k}-y^{*}\left(x_{k}\right)\|^{2}\right] \\
& +\frac{4(1-p)}{b}(L_{2}^{g})^{2}\gamma_{k}^{2}\mathbb{E}\left[\|v_{k}^{z}-D_{z}(x_{k},y_{k},z_{k})\|^{2}\right] \\
& +\frac{4(1-p)}{b}(L_{2}^{g})^{2}\gamma_{k}^{2}L_{z}^{2}\mathbb{E}\left[\|z_{k}-z^{*}\left(x_{k}\right)\|^{2}\right] \\
& +\frac{4(1-p)}{b}(L_{2}^{g})^{2}\gamma_{k}^{2}L_{z}^{2}\mathbb{E}\left[\|y_{k}-y^{*}\left(x_{k}\right)\|^{2}\right] \\
& +\frac{2p\sigma_{f}^{2}}{\tau'}+\frac{2p\sigma_{g,2}^{2}}{\tau'}.
\end{aligned}
$$

$$
\begin{aligned}
(3)\quad \mathbb{E}\left[\left\|v_{k+1}^{z}-D_{z}\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^{2}\right] \leq & (1-p+\frac{4(1-p)}{b}(L_{2}^{g})^{2}\gamma_{k}^{2})\mathbb{E}\left[\|v_{k}^{z}-D_{z}(x_{k},y_{k},z_{k})\|^{2}\right] \\
& +\left(2(L^{f})^{2}+4R^{2}(L_{2}^{g})^{2}\right)\frac{(1-p)}{b}\alpha_{k}^{2}\mathbb{E}\left[\|v_{k}^{x}\|^{2}\right] \\
& +\left(2(L^{f})^{2}+4R^{2}(L_{2}^{g})^{2}\right)\frac{(1-p)}{b}\beta_{k}^{2}\mathbb{E}\left[\|v_{k}^{y}-D_{y}(x_{k},y_{k},z_{k})\|^{2}\right] \\
& +\left(2(L^{f})^{2}+4R^{2}(L_{2}^{g})^{2}\right)\frac{(1-p)}{b}\beta_{k}^{2}\left(L^{f}\right)^{2}\mathbb{E}\left[\|y_{k}-y^{*}\left(x_{k}\right)\|^{2}\right] \\
& +\frac{4(1-p)}{b}(L_{2}^{g})^{2}\gamma_{k}^{2}L_{z}^{2}\mathbb{E}\left[\|z_{k}-z^{*}\left(x_{k}\right)\|^{2}\right] \\
& +\frac{4(1-p)}{b}(L_{2}^{g})^{2}\gamma_{k}^{2}L_{z}^{2}\mathbb{E}\left[\|y_{k}-y^{*}\left(x_{k}\right)\|^{2}\right] \\
& +\frac{2p\sigma_{f}^{2}}{\tau'}+\frac{2p\sigma_{g,2}^{2}}{\tau'}.
\end{aligned}
$$

Proof. **Proof of (1).**

By the definition of v_{k+1}^{y}, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|v_{k+1}^{y}-D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\
=&p\mathbb{E}\left[\left\|\frac{1}{\tau'}\sum_{j\in[\tau']}\nabla_2 G\left(x_{k+1}, y_{k+1};\zeta_j\right)-D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\
&+(1-p)\mathbb{E}\left[\left\|v_k^{y}+\frac{1}{b}\sum_{\zeta_j\in J}\left[\nabla_2 G\left(x_{k+1}, y_{k+1};\zeta_j\right)-\nabla_2 G\left(x_k, y_k;\zeta_j\right)\right]-D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right]
\end{aligned}
$$

For the first term, based on Assumption 3.3, we have

$$
p\mathbb{E}\left[\left\|\frac{1}{\tau'}\sum_{j\in[\tau']}\nabla_2 G\left(x_{k+1}, y_{k+1};\zeta_j\right)-D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right]\leq\frac{p\sigma_{g,1}^2}{\tau'}.
$$

For the second term, analogous to the proof of Lemma F.4, we have

$$
\begin{aligned}
&(1-p)\mathbb{E}\left[\left\|v_k^{y}+\frac{1}{b}\sum_{\zeta_j\in J}\left[\nabla_2 G\left(x_{k+1}, y_{k+1};\zeta_j\right)-\nabla_2 G\left(x_k, y_k;\zeta_j\right)\right]-D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\
\leq&(1-p)\mathbb{E}\left[\left\|v_k^{y}-D_y(x_k, y_k, z_k)\right\|^2\right]+\frac{(1-p)}{b}(L_1^g)^2\alpha_k^2\mathbb{E}\left[\left\|v_k^{x}\right\|^2\right] \\
&+\frac{2(1-p)}{b}(L_1^g)^2\beta_k^2\mathbb{E}\left[\left\|v_k^{y}-D_y(x_k, y_k, z_k)\right\|^2\right] \\
&+\frac{2(1-p)}{b}(L_1^g)^2\beta_k^2\left(L^f\right)^2\mathbb{E}\left[\left\|y_k-y^*\left(x_k\right)\right\|^2\right].
\end{aligned}
$$

In summary, (1) is proved.

Proof of (2) and (3). Similarly, from the definition of v_{k+1}^{x}, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|v_{k+1}^{x}-D_x\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\
=&p\mathbb{E}\left[\left\|\frac{1}{\tau'}\sum_{i\in[\tau']}\nabla_1 F\left(x_{k+1}, y_{k+1};\xi_i\right)-\frac{1}{\tau'}\sum_{j\in[\tau']}\nabla_{12}^2 G\left(x_{k+1}, y_{k+1};\zeta_j\right)-D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\
&+(1-p)\mathbb{E}\left[\left\|v_k^{x}+\frac{1}{b}\sum_{i\in I}\left(\nabla_1 F_i\left(x_{k+1}, y_{k+1}\right)-\nabla_1 F_i\left(x_k, y_k\right)\right)\right.\right. \\
&\left.\left.-\frac{1}{b}\sum_{j\in J}\left(\nabla_{12}^2 G_j\left(x_{k+1}, y_{k+1}\right)z_{k+1}-\nabla_{12}^2 G_j\left(x_k, y_k\right)z_k\right)-D_x\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right].
\end{aligned}
$$

For the first term, based on Assumption 3.3, we have

$$
p\mathbb{E}\left[\left\|\frac{1}{\tau'}\sum_{i\in[\tau']}\nabla_1 F\left(x_{k+1}, y_{k+1};\xi_i\right)-\frac{1}{\tau'}\sum_{j\in[\tau']}\nabla_{12}^2 G\left(x_{k+1}, y_{k+1};\zeta_j\right)-D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right]\leq\frac{2p\sigma_f^2}{\tau'}+\frac{2p\sigma_{g,2}^2}{\tau'}.
$$

For the second term, analogous to the proof of Lemma F.4, we have

$$
\begin{aligned}
&(1-p)\mathbb{E}\Bigg[\Bigg\| v_k^x+\frac{1}{b}\sum_{i\in I}\left(\nabla_1 F_i\left(x_{k+1},y_{k+1}\right)-\nabla_1 F_i\left(x_k,y_k\right)\right)\\
&\quad -\frac{1}{b}\sum_{j\in J}\left(\nabla_{12}^2 G_j\left(x_{k+1},y_{k+1}\right)z_{k+1-}\nabla_{12}^2 G_j\left(x_k,y_k\right)z_k\right)-D_x\left(x_{k+1},y_{k+1},z_{k+1}\right)\Bigg\|^2\Bigg]\\
\leq&(1-p)\mathbb{E}\left[\|v_k^x-D_x(x_k,y_k,z_k)\|^2\right]\\
&+\left(2(L^f)^2+4R^2(L_2^g)^2\right)\frac{(1-p)}{b}\alpha_k^2\mathbb{E}\left[\|v_k^x\|^2\right]\\
&+\left(2(L^f)^2+4R^2(L_2^g)^2\right)\frac{(1-p)}{b}\beta_k^2\mathbb{E}\left[\|v_k^y-D_y(x_k,y_k,z_k)\|^2\right]\\
&+\left(2(L^f)^2+4R^2(L_2^g)^2\right)\frac{(1-p)}{b}\beta_k^2\left(L^f\right)^2\mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right]\\
&+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2\mathbb{E}\left[\|v_k^z-D_z(x_k,y_k,z_k)\|^2\right]\\
&+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2L_z^2\mathbb{E}\left[\|z_k-z^*\left(x_k\right)\|^2\right]\\
&+\frac{4(1-p)}{b}(L_2^g)^2\gamma_k^2L_z^2\mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right].
\end{aligned}
$$

Therefore, (2) is proven. The proof of (3) is analogous, and hence we omit the details here. □

Theorem G.2. ***(Restatement of Theorem 3.9)*** *Fix an iteration $K>1$ and assume that Assumption 3.1 to Assumption 3.4 hold. Choose minibatch size τ' and $b<\tau'$, the probability $p\in(0,1]$. Then there exist positive constants c, c_β, and c_γ, such that if*

$$
\alpha_k\leq\frac{c}{1+\sqrt{\frac{1-p}{pb}}},\quad \beta_k=c_\beta\alpha_k,\quad \gamma_k=c_\gamma\alpha_k,
$$

the iterates in SPABA satisfy

$$
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla H\left(x_k\right)\|^2\right]=\mathcal{O}\left(\frac{1+\sqrt{\frac{1-p}{pb}}}{K}+\frac{1}{Kp\tau'}+\frac{\sigma}{\tau'}\right).
$$

Proof. We consider the Lyapunov function

$$
\begin{aligned}
L_k=&H_k+\mathbb{E}\left[\|y_k-y^*\left(x_k\right)\|^2\right]+\mathbb{E}\left[\|z_k-z^*\left(x_k\right)\|^2\right]\\
&+\frac{\alpha_k}{p}\left(\mathbb{E}\left[\|v_k^x-D_x(x_k,y_k,z_k)\|^2\right]+\mathbb{E}\left[\|v_k^y-D_y(x_k,y_k,z_k)\|^2\right]+\mathbb{E}\left[\|v_k^z-D_z(x_k,y_k,z_k)\|^2\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
& L_{k+1} - L_k \\
\leq & -\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] \\
& + (\alpha_k - \alpha_k)\,\mathbb{E}\left[\left\|D_x(x_k, y_k, z_k) - v_k^x\right\|^2\right] \\
& + \left\{\frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2} + \frac{3(\mu + L_1^g)L_{y^*}^2\alpha_k^2}{\mu L_1^g \beta_k} + \frac{3(\mu + L_1^g)L_{z^*}^2\alpha_k^2}{\mu L_1^g \gamma_k} + \frac{(1-p)}{bp}(L_1^g)^2\alpha_k^3\right. \\
& \quad \left. + \left(4(L^f)^2 + 8R^2(L_2^g)^2\right)\frac{(1-p)}{pb}\alpha_k^3\right\}\mathbb{E}\left[\left\|v_k^x\right\|^2\right] \\
& + \left\{3\alpha_k\left(\left(L^f\right)^2 + (L_2^g R)^2\right) - \frac{\mu L_1^g \gamma_k}{2(\mu + L_1^g)} + \frac{2(1-p)}{bp}(L_1^g)^2\alpha_k\beta_k^2\left(L^f\right)^2\right. \\
& \quad \left. + \left(4(L^f)^2 + 8R^2(L_2^g)^2\right)\frac{(1-p)}{bp}\alpha_k\beta_k^2\left(L^f\right)^2 + \frac{8(1-p)}{bp}(L_2^g)^2\alpha_k\gamma_k^2 L_z^2\right\}\mathbb{E}\left[\left\|y_k - y^*\left(x_k\right)\right\|^2\right] \\
& + \left\{3\alpha_k\left(L_1^g\right)^2 - \frac{\mu L_1^g \gamma_k}{2(\mu + L_1^g)} + \frac{8(1-p)}{bp}(L_2^g)^2\alpha_k\gamma_k^2 L_z^2\right\}\mathbb{E}\left[\left\|z_k - z^*\left(x_k\right)\right\|^2\right] \\
& + \left\{4\frac{\mu + L_1^g}{\mu L_1^g}\beta_k^2 - \alpha_k + \frac{2(1-p)}{bp}(L_1^g)^2\alpha_k\beta_k^2 + \left(4(L^f)^2 + 8R^2(L_2^g)^2\right)\frac{(1-p)}{bp}\alpha_k\beta_k^2\right. \\
& \quad \left. + \left(2(L^f)^2 + 4R^2(L_2^g)^2\right)\frac{(1-p)}{bp}\alpha_k\beta_k^2\right\}\mathbb{E}\left[\left\|v_k^y - D_y(x_k, y_k, z_k)\right\|^2\right] \\
& + \left\{8\frac{\mu + L_1^g}{\mu L_1^g}\gamma_k^2 + \frac{4(1-p)}{bp}(L_2^g)^2\alpha_k\gamma_k^2 - \alpha_k\right\}\mathbb{E}\left[\left\|v_k^z - D_z(x_k, y_k, z_k)\right\|^2\right] \\
& + \frac{\sigma_{g,1}^2\alpha_k}{\tau'} + \frac{4\sigma_f^2\alpha_k}{\tau'} + \frac{4\sigma_{g,2}\alpha_k}{\tau'}.
\end{aligned}
$$

Following the proof process of Theorem 3.7, we have obtained the recursive inequality for this theorem

$$
\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] \leq L_k - L_{k+1} + \frac{\sigma_{g,1}^2\alpha_k}{\tau'} + \frac{4\sigma_f^2\alpha_k}{\tau'} + \frac{4\sigma_{g,2}^2\alpha_k}{\tau'} \leq L_k - L_{k+1} + \frac{\sigma}{\tau'}\alpha_k,
$$

where

$$
\sigma = \sigma_{g,1}^2 + 4\sigma_f^2 + 4\sigma_{g,2}^2, \quad \alpha_k \leq \frac{c}{1 + \sqrt{\frac{1-p}{pb}}}.
$$

Summing, taking the average, and rearranging, we obtain

$$
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] \leq \frac{2L_0}{K\alpha_k} + \frac{2\sigma}{\tau'} = \frac{2L_0'}{K\alpha_k} + \frac{2}{Kp\tau'} + \frac{2\sigma}{\tau'},
$$

where the last equation is based on the fact that

$$
L_0 = H(x_0) + \mathbb{E}[\|y_0 - y^*(y_0)\|^2] + \mathbb{E}[\|z_0 - z^*(y_0)\|^2] + \frac{\alpha_k\sigma}{p\tau'} \triangleq L_0' + \frac{\alpha_k\sigma}{p\tau'}.
$$

Therefore, we ultimately arrive at the conclusion that

$$
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] = \mathcal{O}\left(\frac{1 + \sqrt{\frac{1-p}{pb}}}{K} + \frac{1}{Kp\tau'} + \frac{\sigma}{\tau'}\right).
$$

□

Corollary G.3. *Suppose that Assumption 3.1 to Assumption 3.4 hold. If we take $p = b/(n+m+b)$, $\tau' = \mathcal{O}(\epsilon^{-1})$ and $b \leq \sqrt{\tau'}$, then the sample complexity is $\mathcal{O}(\epsilon^{-1.5})$.*

Proof. In each iteration, it uses $p(n+m) + (1-p)b$ samples on expectation. Let $p = \frac{b}{n+m+b}$, $\tau' = \mathcal{O}(\epsilon^{-1})$ and $b \leq \sqrt{\tau'}$. Thus, the total sample complexity is

$$K(p\tau' + (1-p)b) = \mathcal{O}\left(\epsilon^{-1}\left(1 + \frac{\sqrt{\tau'}}{b} + \frac{\tau'+b}{\tau' b}\right)\frac{2\tau' b}{\tau'+b}\right) = \mathcal{O}\left(\sqrt{\tau'}\epsilon^{-1}\right) = \mathcal{O}\left(\epsilon^{-1.5}\right).$$

□

H. Proof of Theorem 3.11

Lemma H.1. *Suppose Assumptions 3.1, 3.2 and 3.3 hold, then we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|D_{k+1}^{x} - D_{k}^{x}\right\|^2\right] \leq & \tilde{\Delta}\alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + \tilde{\Delta}\beta_k^2 \mathbb{E}[\|v_k^y\|^2] + 4\left(L_1^g\right)^2 \gamma_k^2 \mathbb{E}[\|v_k^z\|^2], \\
\mathbb{E}\left[\left\|D_{k+1}^{z} - D_{k}^{z}\right\|^2\right] \leq & \tilde{\Delta}\alpha_k^2 \mathbb{E}[\|v_k^x\|^2] + \tilde{\Delta}\beta_k^2 \mathbb{E}[\|v_k^y\|^2] + 4\left(L_1^g\right)^2 \gamma_k^2 \mathbb{E}[\|v_k^z\|^2],
\end{aligned}$$

where $\tilde{\Delta} = (2\left(L^f\right)^2 + 4R^2\left(L_2^g\right)^2$.

Proof. Taking the expectation conditionally on x_k, y_k, z_k yields

$$\begin{aligned}
E_k\left[\left\|D_{k+1}^{x} - D_k^x\right\|^2\right] = & E_k\left[\|\nabla_1 F\left(x_{k+1}, y_{k+1}; \xi\right) - \nabla_1 F\left(x_k, y_k; \xi\right)\right. \\
& \left. - \nabla_{12}^2 G\left(x_{k+1}, y_{k+1}; \zeta\right) z_{k+1} - \nabla_{12}^2 G\left(x_k, y_k; \zeta\right) z_k\right\|^2\Big] \\
\leq & 2E_k\left[\|\nabla_1 F\left(x_{k+1}, y_{k+1}; \xi\right) - \nabla_1 F\left(x_k, y_k; \xi\right)\|^2\right] \\
& + 4E_k\left[\left\|\left(\nabla_{12}^2 G\left(x_{k+1}, y_{k+1}; \zeta\right) - \nabla_{12}^2 G\left(x_k, y_k; \zeta\right)\right) z_{k+1}\right\|^2\right] \\
& + 4E_k\left[\left\|\nabla_{12}^2 G\left(x_k, y_k; \zeta\right)\left(z_{k+1} - z_k\right)\right\|^2\right],
\end{aligned}$$

where the inequality is derived using the Cauchy-Schwarz inequality. For the term $\|z_{k+1} - z_k\|^2$, based on the definition of z_{k+1} and the contractivity of projection, we have

$$\begin{aligned}
\|z_{k+1} - z_k\|^2 = & \left\|\mathrm{Proj}_{\mathbb{B}(R)}\left(z_k - \gamma_k v_k^z\right) - z_k\right\|^2 = \left\|\mathrm{Proj}_{\mathbb{B}(R)}\left(z_k - \gamma_k v_k^z\right) - \mathrm{Proj}_{\mathbb{B}(R)}(z_k)\right\|^2 \\
& \leq \|z_k - \gamma_k v_k^z - z_k\|^2 = \gamma_k^2 \|v_k^z\|^2.
\end{aligned}$$

Thus, substituting into the above equation, we obtain

$$\begin{aligned}
E_k\left[\left\|D_{k+1}^{x} - D_k^x\right\|^2\right] \leq & 2\left(L^f\right)^2\left(\alpha_k^2 E_k[\|v_k^x\|^2] + \beta_k^2 E_k[\|v_k^y\|^2]\right) \\
& + 4R^2\left(L_2^g\right)^2\left(\alpha_k^2 E_k[\|v_k^x\|^2] + \beta_k^2 E_k[\|v_k^y\|^2]\right) + 4\left(L_1^g\right)^2 \gamma_k^2 E_k[\|v_k^z\|^2] \\
= & \left(2\left(L^f\right)^2 + 4R^2\left(L_2^g\right)^2\right)\alpha_k^2 E_k[\|v_k^x\|^2] \\
& + \left(2\left(L^f\right)^2 + 4R^2\left(L_2^g\right)^2\right)\beta_k^2 E_k[\|v_k^y\|^2] + 4\left(L_1^g\right)^2 \gamma_k^2 E_k[\|v_k^z\|^2],
\end{aligned}$$

where the inequality results from the boundedness generated by projecting z_k, as well as Assumptions 3.1, 3.2 and 3.3. Finally, by taking the total expectation, the lemma is proven.

Similarly, we can derive the inequality concerning $\mathbb{E}\left[\left\|D_{k+1}^{z} - D_k^z\right\|^2\right]$. □

Lemma H.2. *Under the Assumption 3.1, 3.2, 3.3 and 3.4, We have the following inequalities established:*

$$\begin{aligned}(1)\quad \mathbb{E}\left[\left\|v_{k+1}^y - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \leq & \left(\left(1-\rho_k^y\right)^2 + 4\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\beta_k^2\right)\mathbb{E}\left[\left\|v_k^y - D_y\left(x_k, y_k, z_k\right)\right\|^2\right] \\ & + 2\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\alpha_k^2\mathbb{E}\left[\left\|v_k^x\right\|^2\right] \\ & + 4\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\beta_k^2\mathbb{E}\left[\left\|D_y(x_k, y_k, z_k)\right\|^2\right] + 2\left(\rho_k^y\right)^2\sigma_{g,1}^2,\end{aligned}$$

$$\begin{aligned}(2)\quad & \mathbb{E}\left[\left\|v_{k+1}^x - D_x\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\ \leq & \left(1-\rho_k^x\right)^2\mathbb{E}\left[\left\|v_k^x - D_x\left(x_k, y_k, z_k\right)\right\|^2\right] \\ & + 4\left(1-\rho_k^x\right)^2\tilde{\Delta}\beta_k^2\mathbb{E}\left[\left\|v_k^y - D_y(x_k, y_k, z_k)\right\|^2\right] + 16\left(1-\rho_k^x\right)^2\left(L_1^g\right)^2\gamma_k^2\mathbb{E}\left[\left\|v_k^z - D_z(x_k, y_k, z_k)\right\|^2\right] \\ & + 2\left(1-\rho_k^x\right)^2\tilde{\Delta}\alpha_k^2\mathbb{E}\left[\left\|v_k^x\right\|^2\right] + 4\left(1-\rho_k^x\right)^2\tilde{\Delta}\beta_k^2\mathbb{E}\left[\left\|D_y(x_k, y_k, z_k)\right\|^2\right] \\ & + 16\left(1-\rho_k^x\right)^2\left(L_1^g\right)^2\gamma_k^2\mathbb{E}\left[\left\|D_z(x_k, y_k, z_k)\right\|^2\right] + 4\left(\rho_k^x\right)^2\sigma_{g,2}^2\mathbb{E}\left[\left\|z_{k+1} - z^*\left(x_{k+1}\right)\right\|^2\right] \\ & + 2\left(\rho_k^x\right)^2\left(2\sigma_{g,2}^2R^2 + \sigma_f^2\right),\end{aligned}$$

$$\begin{aligned}(3)\quad & \mathbb{E}\left[\left\|v_{k+1}^z - D_z\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\ \leq & \left(\left(1-\rho_k^z\right)^2 + 16\left(1-\rho_k^z\right)^2\left(L_1^g\right)^2\gamma_k^2\right)\mathbb{E}\left[\left\|v_k^z - D_z\left(x_k, y_k, z_k\right)\right\|^2\right] \\ & + 4\left(1-\rho_k^z\right)^2\tilde{\Delta}\beta_k^2\mathbb{E}\left[\left\|v_k^y - D_y(x_k, y_k, z_k)\right\|^2\right] \\ & + 2\left(1-\rho_k^z\right)^2\tilde{\Delta}\alpha_k^2\mathbb{E}\left[\left\|v_k^x\right\|^2\right] + 4\left(1-\rho_k^z\right)^2\tilde{\Delta}\beta_k^2\mathbb{E}\left[\left\|D_y(x_k, y_k, z_k)\right\|^2\right] \\ & + 16\left(1-\rho_k^z\right)^2\left(L_1^g\right)^2\gamma_k^2\mathbb{E}\left[\left\|D_z(x_k, y_k, z_k)\right\|^2\right] + 4\left(\rho_k^z\right)^2\sigma_{g,2}^2\mathbb{E}\left[\left\|z_{k+1} - z^*\left(x_{k+1}\right)\right\|^2\right] \\ & + 2\left(\rho_k^z\right)^2\left(2\sigma_{g,2}^2R^2 + \sigma_f^2\right).\end{aligned}$$

Proof. **proof of (1)**

By the definition of v_{k+1}^y, we have

$$\begin{aligned}\left\|v_{k+1}^y - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2 = & \left\|D_{k+1}^y + \left(1-\rho_k^y\right)\left(v_k^y - D_k^y\right) - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2 \\ = & \left\|\left(1-\rho_k^y\right)\left(v_k^y - D_y\left(x_k, y_k, z_k\right)\right) + \rho_k^y\left(D_{k+1}^y - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right)\right. \\ & \left. + \left(1-\rho_k^y\right)\left(D_{k+1}^y - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right) - D_k^y + D_y\left(x_k, y_k, z_k\right)\right)\right\|^2,\end{aligned}$$

Taking the expectation conditionally on x_k, y_k, z_k, and utilizing that D_{k+1}^y and D_k^y are unbiased estimates of $D_y(x_{k+1}, y_{k+1}, z_{k+1})$ and $D_y(x_k, y_k, z_k)$ respectively, yields

$$\begin{aligned}& E_k\left[\left\|v_{k+1}^y - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\ \leq & \left(1-\rho_k^y\right)^2 E_k\left[\left\|v_k^y - D_y\left(x_k, y_k, z_k\right)\right\|^2\right] + 2\left(\rho_k^y\right)^2 E_k\left[\left\|D_{k+1}^y - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\ & + 2\left(1-\rho_k^y\right)^2 E_k\left[\left\|D_{k+1}^y - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right) - D_k^y + D_y\left(x_k, y_k, z_k\right)\right\|^2\right] \\ \leq & \left(1-\rho_k^y\right)^2 E_k\left[\left\|v_k^y - D_y\left(x_k, y_k, z_k\right)\right\|^2\right] + 2\left(\rho_k^y\right)^2 E_k\left[\left\|D_{k+1}^y - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \\ & + 2\left(1-\rho_k^y\right)^2 E_k\left[\left\|D_{k+1}^y - D_k^y\right\|^2\right],\end{aligned}$$

For the second term, by setting $|\mathcal{S}_2| = 1$ in Lemma I.3, we obtain

$$E_k\left[\left\|D_{k+1}^y - D_y\left(x_{k+1}, y_{k+1}, z_{k+1}\right)\right\|^2\right] \leq \sigma_{g,1}^2,$$

For the third term, from Assumption 3.4, we obtain

$$E_k\left[\left\|D^y_{k+1}-D^y_k\right\|^2\right]=E_k\left[\left\|\nabla_2 G(x_{k+1},y_{k+1};\zeta)-\nabla_2 G(x_k,y_k;\zeta)\right\|^2\right]\le\left(L_1^g\right)^2\left(\alpha_k^2 E_k\left[\left\|v_k^x\right\|^2\right]+\beta_k^2 E_k\left[\left\|v_k^y\right\|^2\right]\right),$$

Taking the total expectation, ultimately we can derive

$$\begin{aligned}
\mathbb{E}\left[\left\|v^y_{k+1}-D_y\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^2\right]\le&\left(1-\rho_k^y\right)^2\mathbb{E}\left[\left\|v_k^y-D_y\left(x_k,y_k,z_k\right)\right\|^2\right]+2\left(\rho_k^y\right)^2\sigma_{g,1}^2\\
&+2\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\alpha_k^2\mathbb{E}\left[\left\|v_k^x\right\|^2\right]+2\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\beta_k^2\mathbb{E}\left[\left\|v_k^y\right\|^2\right]\\
\le&\left(\left(1-\rho_k^y\right)^2+4\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\beta_k^2\right)\mathbb{E}\left[\left\|v_k^y-D_y\left(x_k,y_k,z_k\right)\right\|^2\right]\\
&+2\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\alpha_k^2\mathbb{E}\left[\left\|v_k^x\right\|^2\right]\\
&+4\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\beta_k^2\mathbb{E}\left[\left\|D_y(x_k,y_k,z_k)\right\|^2\right]\\
&+2\left(\rho_k^y\right)^2\sigma_{g,1}^2.
\end{aligned}$$

Thus, the (1) is proven.

proof of (2) and (3)

Based on the definition of v^x_{k+1} and the fact that D^x_{k+1} and D^x_k are unbiased estimates of $D_x(x_{k+1},y_{k+1},z_{k+1})$ and $D_x(x_k,y_k,z_k)$ respectively, we have

$$\begin{aligned}
&E_k\left[\left\|v^x_{k+1}-D_x\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^2\right]\\
=&E_k\left\|D^x_{k+1}+\left(1-\rho_k^x\right)\left(v_k^x-D_k^x\right)-D_x\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^2\\
=&E_k\left[\left\|\left(1-\rho_k^x\right)\left(v_k^x-D_x\left(x_k,y_k,z_k\right)\right)+\rho_k^x\left(D^x_{k+1}-D_x\left(x_{k+1},y_{k+1},z_{k+1}\right)\right)\right.\right.\\
&\left.\left.+\left(1-\rho_k^x\right)\left(D^x_{k+1}-D_x\left(x_{k+1},y_{k+1},z_{k+1}\right)-D_k^x+D_x\left(x_k,y_k,z_k\right)\right)\right\|^2\right]\\
=&\left(1-\rho_k^x\right)^2E_k\left[\left\|v_k^x-D_x\left(x_k,y_k,z_k\right)\right\|^2\right]+2\left(\rho_k^x\right)^2E_k\left[\left\|D^x_{k+1}-D_x\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^2\right]\\
&+2\left(1-\rho_k^x\right)^2E_k\left[\left\|D^x_{k+1}-D^x_k\right\|^2\right],
\end{aligned}$$

where the second term can be bounded using Lemma I.3 (with $|\mathcal{S}_1|$ and $|\mathcal{S}_2|$ set to 1), and the third term can be bounded using Lemma H.1, thus further leading to

$$\begin{aligned}
&E_k\left[\left\|v^x_{k+1}-D_x\left(x_{k+1},y_{k+1},z_{k+1}\right)\right\|^2\right]\\
\le&\left(1-\rho_k^x\right)^2E_k\left[\left\|v_k^x-D_x\left(x_k,y_k,z_k\right)\right\|^2\right]+2\left(\rho_k^x\right)^2\left(2\sigma_{g,2}^2R^2+\sigma_f^2\right)\\
&+4\left(\rho_k^x\right)^2\sigma_{g,2}^2E_k\left[\left\|z_{k+1}-z^*\left(x_{k+1}\right)\right\|^2\right]+2\left(1-\rho_k^x\right)^2\tilde{\Delta}\alpha_k^2E_k\left[\left\|v_k^x\right\|^2\right]\\
&+2\left(1-\rho_k^x\right)^2\tilde{\Delta}\beta_k^2E_k\left[\left\|v_k^y\right\|^2\right]+8\left(1-\rho_k^x\right)^2\left(L_1^g\right)^2\gamma_k^2E_k\left[\left\|v_k^z\right\|^2\right]\\
\le&\left(1-\rho_k^x\right)^2E_k\left[\left\|v_k^x-D_x\left(x_k,y_k,z_k\right)\right\|^2\right]+2\left(\rho_k^x\right)^2\left(2\sigma_{g,2}^2R^2+\sigma_f^2\right)+4\left(\rho_k^x\right)^2\sigma_{g,2}^2E_k\left[\left\|z_{k+1}-z^*\left(x_{k+1}\right)\right\|^2\right]\\
&+4\left(1-\rho_k^x\right)^2\tilde{\Delta}\beta_k^2\mathbb{E}\left[\left\|v_k^y-D_y(x_k,y_k,z_k)\right\|^2\right]+4\left(1-\rho_k^x\right)^2\tilde{\Delta}\beta_k^2E_k\left[\left\|D_y(x_k,y_k,z_k)\right\|^2\right]\\
&+16\left(1-\rho_k^x\right)^2\left(L_1^g\right)^2\gamma_k^2E_k\left[\left\|v_k^z-D_z(x_k,y_k,z_k)\right\|^2\right]+16\left(1-\rho_k^x\right)^2\left(L_1^g\right)^2\gamma_k^2E_k\left[\left\|D_z(x_k,y_k,z_k)\right\|^2\right]\\
&+2\left(1-\rho_k^x\right)^2\tilde{\Delta}\alpha_k^2E_k\left[\left\|v_k^x\right\|^2\right].
\end{aligned}$$

By taking the total expectation and rearranging the above expression, the (2) is proven.

We can similarly prove (3). □

Theorem H.3. ***(Restatement of Theorem 3.11)***

Fix an iteration $K > 1$ and assume that Assumption 3.1 to Assumption 3.4 hold. Then there exist positive constants η, c_β, c_γ, c_x, c_y and c_z such that if

$$\alpha_k = \frac{1}{(\eta+k)^{1/3}}, \quad \beta_k = c_\beta \alpha_k, \quad \gamma_k = c_\gamma \alpha_k; \quad \rho_k^x = c_x \alpha_k^2, \quad \rho_k^y = c_y \alpha_k^2, \quad \rho_k^z = c_z \alpha_k^2,$$

the iterates in SRMBA satisfy

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla H\left(x_k\right)\|^2\right] = \mathcal{O}\left(\frac{1}{K^{2/3}} + \frac{\log(K-1)\sigma}{K^{2/3}}\right).$$

Proof. We consider the Lyapunov function

$$\begin{aligned} L_k =& \mathbb{E}\left[H\left(x_k\right)\right] + A\mathbb{E}\left[\|y_k - y^*\left(x_k\right)\|^2\right] + B\mathbb{E}\left[\|z_k - z^*\left(x_k\right)\|^2\right] \\ &+ \frac{1}{C_{k-1}}\mathbb{E}[\|v_k^y - D_y(x_k,y_k,z_k)\|^2] + \frac{1}{D_{k-1}}\mathbb{E}[\|v_k^x - D_x(x_k,y_k,z_k)\|^2] + \frac{1}{F_{k-1}}\mathbb{E}[\|v_k^z - D_z(x_k,y_k,z_k)\|^2]. \end{aligned}$$

Using Lemma D.6 Lemma D.7, Lemma E.2, Lemma E.1 and Lemma E.5, we get

$$\begin{aligned} L_{k+1} - L_k =& \mathbb{E}\left[H\left(x_{k+1}\right)\right] - E\left[H\left(x_k\right)\right] + A\left(\mathbb{E}\left[\|y_{k+1} - y^*\left(x_{k+1}\right)\|^2\right] - \mathbb{E}\left[\|y_k - y^*\left(x_k\right)\|^2\right]\right) \\ &+ B\left(\mathbb{E}\left[\|z_{k+1} - z^*\left(x_{k+1}\right)\|^2\right] - \mathbb{E}\left[\|z_k - z^*\left(x_k\right)\|^2\right]\right) \\ &+ \frac{1}{C_k}\mathbb{E}\left[\left\|v_{k+1}^y - D_y(x_{k+1},y_{k+1},z_{k+1})\right\|^2\right] - \frac{1}{C_{k-1}}\mathbb{E}\left[\|v_k^y - D_y(x_k,y_k,z_k)\|^2\right] \\ &+ \frac{1}{D_k}\mathbb{E}\left[\left\|v_{k+1}^x - D_x(x_{k+1},y_{k+1},z_{k+1})\right\|^2\right] - \frac{1}{D_{k-1}}\mathbb{E}\left[\|v_k^x - D_x(x_k,y_k,z_k)\|^2\right] \\ &+ \frac{1}{F_k}\mathbb{E}\left[\left\|v_{k+1}^z - D_z(x_{k+1},y_{k+1},z_{k+1})\right\|^2\right] - \frac{1}{F_{k-1}}\mathbb{E}\left[\|v_k^z - D_z(x_k,y_k,z_k)\|^2\right] \end{aligned}$$

By incorporating Lemmas D.7, F.3, and H.2 into the above inequality, we can derive

$$
\begin{aligned}
L_{k+1} - L_k \leq & -\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] \\
& + \left(3\left(\left(L^f\right)^2 + \left(L_2^g R\right)^2\right)\alpha_k - \frac{\mu L_1^g \beta_k}{2(\mu + L_1^g)} A\right)\mathbb{E}\left[\left\|y_k - y^*\left(x_k\right)\right\|^2\right] \\
& + \left(\left(B + \frac{4\left(\rho_k^x\right)^2\sigma_{g,2}^2}{D_k} + \frac{4\left(\rho_k^z\right)^2\sigma_{g,2}^2}{F_k}\right)\left(1 - \frac{\mu L_1^g \gamma_k}{2(\mu + L_1^g)}\right) - B + 3\left(L_1^g\right)^2\alpha_k\right)\mathbb{E}\left[\left\|z_k - z^*\left(x_k\right)\right\|^2\right] \\
& + \left(\alpha_k + \frac{\left(1-\rho_k^x\right)^2}{D_k} - \frac{1}{D_{k-1}}\right)\mathbb{E}\left[\left\|v_k^x - D_x(x_k, y_k, z_k)\right\|^2\right] \\
& + \left(\frac{4\left(1-\rho_k^x\right)^2\tilde{\Delta}\beta_k^2}{D_k} - \frac{1}{\mu + L_1^g}\beta_k A + \frac{4\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\beta_k^2}{C_k} + \frac{4\left(1-\rho_k^z\right)^2\tilde{\Delta}\beta_k^2}{F_k}\right)\mathbb{E}\left[\left\|D_y(x_k, y_k, z_k)\right\|^2\right] \\
& + \left(4\frac{\mu + L_1^g}{\mu L_1^g}\beta_k^2 A + \frac{\left(1-\rho_k^y\right)^2 + 4\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\beta_k^2}{C_k} - \frac{1}{C_{k-1}} + \frac{4\left(1-\rho_k^x\right)^2\tilde{\Delta}\beta_k^2}{D_k} + \frac{4\left(1-\rho_k^z\right)^2\tilde{\Delta}\beta_k^2}{F_k}\right)\times \\
& \quad \mathbb{E}\left[\left\|v_k^y - D_y(x_k, y_k, z_k)\right\|^2\right] \\
& + \left(\frac{16\left(1-\rho_k^x\right)^2\left(L_1^g\right)^2\gamma_k^2}{D_k} - \frac{1}{\mu + L_1^g}\gamma_k\left(B + \frac{4\left(\rho_k^x\right)^2\sigma_{g,2}^2}{D_k} + \frac{4\left(\rho_k^z\right)^2\sigma_{g,2}^2}{F_k}\right) + \frac{16\left(1-\rho_k^z\right)^2\left(L_1^g\right)^2\gamma_k^2}{F_k}\right)\times \\
& \quad \mathbb{E}\left[\left\|D_z(x_k, y_k, z_k)\right\|^2\right] \\
& + \Bigg(4\frac{\mu + L_1^g}{\mu L_1^g}\gamma_k^2\left(B + \frac{4\left(\rho_k^x\right)^2\sigma_{g,2}^2}{D_k} + \frac{4\left(\rho_k^z\right)^2\sigma_{g,2}^2}{F_k}\right) + \frac{16\left(1-\rho_k^x\right)^2\left(L_1^g\right)^2\gamma_k^2}{D_k} \\
& \quad + \frac{\left(1-\rho_k^z\right)^2 + 16\left(1-\rho_k^z\right)^2\left(L_1^g\right)^2\gamma_k^2}{F_k} - \frac{1}{F_{k-1}}\Bigg)\mathbb{E}\left[\left\|v_k^z - D_z(x_k, y_k, z_k)\right\|^2\right] \\
& + \Bigg(\frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2} + \frac{3(\mu + L_1^g)L_{y^*}^2\alpha_k^2}{\mu L_1^g \beta_k}A + \frac{3(\mu + L_1^g)L_{z^*}^2\alpha_k^2}{\mu L_1^g \gamma_k}\left(B + \frac{4\left(\rho_k^x\right)^2\sigma_{g,2}^2}{D_k} + \frac{4\left(\rho_k^z\right)^2\sigma_{g,2}^2}{F_k}\right) \\
& \quad + \frac{2\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\alpha_k^2}{C_k} + \frac{2\left(1-\rho_k^x\right)^2\tilde{\Delta}\alpha_k^2}{D_k} + \frac{2\left(1-\rho_k^z\right)^2\tilde{\Delta}\alpha_k^2}{F_k}\Bigg)\mathbb{E}\left[\left\|v_k^x\right\|^2\right] \\
& + \frac{2\left(\rho_k^y\right)^2\sigma_{g,1}^2}{C_k} + \frac{2\left(\rho_k^x\right)^2\left(2\sigma_{g,2}^2R^2 + \sigma_f^2\right)}{D_k} + \frac{2\left(\rho_k^z\right)^2\left(2\sigma_{g,2}^2R^2 + \sigma_f^2\right)}{F_k},
\end{aligned}
$$

We select the coefficients of the Lyapunov function and the step sizes of the algorithm as follows

$$
\begin{gathered}
\alpha_k = \frac{1}{(\eta + k)^{1/3}}, \quad \beta_k = c_\beta\alpha_k, \quad \gamma_k = c_\gamma\alpha_k; \quad \rho_k^x = c_x\alpha_k^2, \quad \rho_k^y = c_y\alpha_k^2, \quad \rho_k^z = c_z\alpha_k^2; \\
A = B = 1, \quad C_k = \phi_1\alpha_k, \qquad D_k = \phi_2\alpha_k, \qquad F_k = \phi_3\alpha_k.
\end{gathered}
$$

Analysis of the Coefficient for $\mathbb{E}\left[\left\|y_k - y^*\left(x_k\right)\right\|^2\right]$

Due to the assumption that

$$
\alpha_k \leq \frac{\mu L_1^g}{6(\mu + L_1^g)\left(\left(L^f\right)^2 + \left(L_2^g R\right)^2\right)}\beta_k,
$$

it follows that

$$
3\left(\left(L^f\right)^2 + \left(L_2^g R\right)^2\right)\alpha_k - \frac{\mu L_1^g \beta_k}{2(\mu + L_1^g)}A \leq 0.
$$

Analysis of the Coefficient for $\mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right]$

By assuming

$$c_x^2 \leq \frac{\mu L_1^g c_\gamma}{16(\mu + L_1^g)c_2}, \quad c_z^2 \leq \frac{\mu L_1^g c_\gamma c_3}{32(\mu + L_1^g)}, \quad \alpha_k \leq \frac{\mu L_1^g}{48(\mu + L_1^g)(L_1^g)^2},$$

we have

$$\begin{aligned}
&\left(B + \frac{4\left(\rho_k^x\right)^2 \sigma_{g,2}^2}{D_k} + \frac{4\left(\rho_k^z\right)^2 \sigma_{g,2}^2}{F_k}\right)\left(1 - \frac{\mu L_1^g \gamma_k}{2(\mu + L_1^g)}\right) - B + 3\left(L_1^g\right)^2 \alpha_k \\
&\leq -\frac{\mu L_1^g \gamma_k}{2(\mu + L_1^g)} + \frac{4\left(\rho_k^x\right)^2 \sigma_{g,2}^2}{D_k} + \frac{4\left(\rho_k^z\right)^2 \sigma_{g,2}^2}{F_k} + 3\left(L_1^g\right)^2 \alpha_k \\
&\leq 0.
\end{aligned}$$

Analysis of the Coefficient for $\mathbb{E}\left[\|v_k^x - D_x(x_k, y_k, z_k)\|^2\right]$

Based on the definition of α_k and the choice of D_k

$$\begin{aligned}
\frac{\left(1-\rho_k^x\right)^2}{D_k} - \frac{1}{D_{k-1}} \leq \frac{1}{D_k} - \frac{1}{D_{k-1}} - \frac{\rho_k^x}{D_k} &= \frac{1}{\phi_2}\left(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k-1}} - c_x \alpha_k\right) \\
&= \frac{1}{\phi_2}\left((\eta + k)^{1/3} - (\eta + k - 1)^{1/3} - c_x \alpha_k\right) \\
&\leq \frac{1}{\phi_2}\left(\frac{2^{2/3}}{3(\eta + k)^{2/3}} - c_x \alpha_k\right) \\
&\leq \frac{1}{\phi_2}\left(\frac{2^{2/3}}{3}\alpha_k^2 - c_x \alpha_k\right) \\
&\leq -\alpha_k,
\end{aligned}$$

where the second inequality follows from $(x+y)^{1/3} - x^{1/3} \leq y/(3x^{2/3})$ and $\eta \geq 2$, the third inequality is based on the definition of α_k, and the final inequality results from our choice of $\alpha_k \leq 1/L^f$ and $c_x = \phi_2 + 1/L^f$. Therefore, the coefficient of this term is

$$\alpha_k + \frac{\left(1-\rho_k^x\right)^2}{D_k} - \frac{1}{D_{k-1}} \leq 0.$$

Analysis of the Coefficient for $\mathbb{E}\left[\|D_y(x_k, y_k, z_k)\|^2\right]$

Due to the assumption that

$$\beta_k \leq \min\left\{\frac{\phi_2}{8\tilde{\Delta}(\mu + L_1^g)}, \quad \frac{\phi_1}{16(L_1^g)^2(\mu + L_1^g)}, \quad \frac{\phi_3}{32\tilde{\Delta}(\mu + L_1^g)}\right\}\alpha_k,$$

it follows that

$$\frac{4\left(1-\rho_k^x\right)^2 \tilde{\Delta}\beta_k^2}{D_k} - \frac{1}{\mu + L_1^g}\beta_k A + \frac{4\left(1-\rho_k^y\right)^2 \left(L_1^g\right)^2 \beta_k^2}{C_k} + \frac{4\left(1-\rho_k^z\right)^2 \tilde{\Delta}\beta_k^2}{F_k} \leq 0.$$

Analysis of the Coefficient for $\mathbb{E}\left[\|v_k^y - D_y(x_k, y_k, z_k)\|^2\right]$

Similarly, when $\alpha_k \leq 1/L^f$ and $c_y = \phi_1 + 1/L^f$, we have

$$\frac{\left(1-\rho_k^y\right)^2}{C_k} - \frac{1}{C_{k-1}} \leq -\alpha_k.$$

Furthermore, by assuming

$$\beta_k^2 \leq \min\left\{\frac{\mu L_1^g}{8(\mu+L_1^g)}\alpha_k, \quad \frac{\phi_1}{16(L_1^g)^2}\alpha_k^2, \quad \frac{\phi_2}{32\tilde{\Delta}}\alpha_k^2, \quad \frac{\phi_3}{64\tilde{\Delta}}\alpha_k^2\right\},$$

we have

$$4\frac{\mu+L_1^g}{\mu L_1^g}\beta_k^2 A + \frac{\left(1-\rho_k^y\right)^2 + 4\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\beta_k^2}{C_k} - \frac{1}{C_{k-1}} + \frac{4\left(1-\rho_k^x\right)^2\tilde{\Delta}\beta_k^2}{D_k} + \frac{4\left(1-\rho_k^z\right)^2\tilde{\Delta}\beta_k^2}{F_k} \leq 0.$$

Analysis of the Coefficient for $\mathbb{E}\left[\left\|D_z(x_k,y_k,z_k)\right\|^2\right]$

Due to the assumption that

$$\gamma_k \leq \min\left\{\frac{\phi_2}{32(L_1^g)^2(\mu+L_1^g)}, \quad \frac{\phi_3}{64(L_1^g)^2(\mu+L_1^g)}\right\}\alpha_k,$$

we have

$$\begin{aligned}
&\frac{16\left(1-\rho_k^x\right)^2\left(L_1^g\right)^2\gamma_k^2}{D_k} - \frac{1}{\mu+L_1^g}\gamma_k\left(B + \frac{4\left(\rho_k^x\right)^2\sigma_{g,2}^2}{D_k} + \frac{4\left(\rho_k^z\right)^2\sigma_{g,2}^2}{F_k}\right) + \frac{16\left(1-\rho_k^z\right)^2\left(L_1^g\right)^2\gamma_k^2}{F_k}\\
&\leq \frac{16\left(1-\rho_k^x\right)^2\left(L_1^g\right)^2\gamma_k^2}{D_k} - \frac{1}{\mu+L_1^g}\gamma_k B + \frac{16\left(1-\rho_k^z\right)^2\left(L_1^g\right)^2\gamma_k^2}{F_k}\\
&\leq 0.
\end{aligned}$$

Analysis of the Coefficient for $\mathbb{E}\left[\|v_k^z - D_z(x_k,y_k,z_k)\|^2\right]$

Similar to the analysis of $\mathbb{E}\left[\left\|v_k^y - D_y(x_k,y_k,z_k)\right\|^2\right]$, by assuming

$$\begin{aligned}
\gamma_k^2 &\leq \min\left\{\frac{\phi_3}{32(L_1^g)^2}\alpha_k^2, \quad \frac{\phi_2}{64(L_1^g)^2}\alpha_k^2, \quad \frac{\mu L_1^g}{64(\mu+L_1^g)}\alpha_k\right\},\\
\alpha_k &\leq \min\left\{\frac{\phi_2}{8\sigma_{g,2}^2c_x^2}, \quad \frac{\phi_3}{8\sigma_{g,2}^2c_z^2}\right\},
\end{aligned}$$

we have

$$\begin{aligned}
&4\frac{\mu+L_1^g}{\mu L_1^g}\gamma_k^2\left(B + \frac{4\left(\rho_k^x\right)^2\sigma_{g,2}^2}{D_k} + \frac{4\left(\rho_k^z\right)^2\sigma_{g,2}^2}{F_k}\right) + \frac{16\left(1-\rho_k^x\right)^2\left(L_1^g\right)^2\gamma_k^2}{D_k}\\
&\quad + \frac{\left(1-\rho_k^z\right)^2 + 16\left(1-\rho_k^z\right)^2\left(L_1^g\right)^2\gamma_k^2}{F_k} - \frac{1}{F_{k-1}} \leq 0.
\end{aligned}$$

Analysis of the Coefficient for $\mathbb{E}\left[\left\|v_k^x\right\|^2\right]$

By assuming that the step sizes and the parameters of the Lyapunov function satisfy

$$\begin{aligned}
\alpha_k \leq \min\left\{\frac{\mu L_1^g}{24(\mu+L_1^g)L_{y^*}^2}\beta_k, \quad \frac{\mu L_1^g}{48(\mu+L_1^g)L_{y^*}^2}\gamma_k, \quad \frac{1}{4L^H}\right\},\\
\phi_1 \geq 12(L_1^g)^2, \quad \phi_2 \geq 12\tilde{\Delta}, \quad \phi_3 \geq 12\tilde{\Delta},
\end{aligned}$$

we have

$$\begin{aligned}
&\frac{L^H\alpha_k^2}{2} - \frac{\alpha_k}{2} + \frac{3(\mu+L_1^g)L_{y^*}^2\alpha_k^2}{\mu L_1^g\beta_k}A + \frac{3(\mu+L_1^g)L_{z^*}^2\alpha_k^2}{\mu L_1^g\gamma_k}\left(B + \frac{4\left(\rho_k^x\right)^2\sigma_{g,2}^2}{D_k} + \frac{4\left(\rho_k^z\right)^2\sigma_{g,2}^2}{F_k}\right)\\
&\quad + \frac{2\left(1-\rho_k^y\right)^2\left(L_1^g\right)^2\alpha_k^2}{C_k} + \frac{2\left(1-\rho_k^x\right)^2\tilde{\Delta}\alpha_k^2}{D_k} + \frac{2\left(1-\rho_k^z\right)^2\tilde{\Delta}\alpha_k^2}{F_k} \leq 0.
\end{aligned}$$

Combining the above analysis of the coefficients of each term, we can obtain the simplified inequality as follows

$$L_{k+1}-L_k \leq -\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right]+\frac{2\left(\rho_k^y\right)^2\sigma_{g,1}^2}{C_k}+\frac{2\left(\rho_k^x\right)^2\left(2\sigma_{g,2}^2R^2+\sigma_f^2\right)}{D_k}+\frac{2\left(\rho_k^z\right)^2\left(2\sigma_{g,2}^2R^2+\sigma_f^2\right)}{F_k},$$

Summing, taking the average, and rearranging, we obtain

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] \\
\leq&\frac{L_0}{K}+\frac{1}{K}\sum_{k=0}^{K-1}\left(\frac{2\left(\rho_k^y\right)^2\sigma_{g,1}^2}{C_k}+\frac{2\left(\rho_k^x\right)^2\left(2\sigma_{g,2}^2R^2+\sigma_f^2\right)}{D_k}+\frac{2\left(\rho_k^z\right)^2\left(2\sigma_{g,2}^2R^2+\sigma_f^2\right)}{F_k}\right) \\
&=\frac{L_0}{K}+\frac{1}{K}\sum_{k=0}^{K-1}\alpha_k^3\left(\frac{2\left(c_y\right)^2\sigma_{g,1}^2}{\phi_1}+\frac{2\left(c_x\right)^2\left(2\sigma_{g,2}^2R^2+\sigma_f^2\right)}{\phi_2}+\frac{2\left(c_z\right)^2\left(2\sigma_{g,2}^2R^2+\sigma_f^2\right)}{\phi_3}\right) \\
&\leq\frac{L_0}{K}+\frac{\log(K-1)}{K}\left(\frac{2\left(c_y\right)^2\sigma_{g,1}^2}{\phi_1}+\frac{2\left(c_x\right)^2\left(2\sigma_{g,2}^2R^2+\sigma_f^2\right)}{\phi_2}+\frac{2\left(c_z\right)^2\left(2\sigma_{g,2}^2R^2+\sigma_f^2\right)}{\phi_3}\right).
\end{aligned}$$

Based on $\alpha_k \geq \alpha_K$, we can finally obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right]=\mathcal{O}\left(\frac{1}{K^{2/3}}+\frac{\log(K-1)}{K^{2/3}}\right).$$

□

I. MA-SOBA-q: Vanilla minibatch SGD + Standard Momentum

For the expectation form setting, we introduce SRMBA, which employs mini-batch stochastic estimation in its estimation module and selects Moving-average in the acceleration module that reference the update direction from the previous iteration.

To illustrate further, at each iteration k, we draw two random set $\mathcal{S}_1$ and $\mathcal{S}_2$ with a fixed mini-batch size of S, for the functions f and g respectively, to perform a stochastic estimation of $D_\bullet$. γ_k, β_k and α_k are the step sizes and ρ_k is the moving average parameter. A trade-off between the step size and batch size has been made, with more detailed descriptions to be provided in Theorem I.1. Furthermore, we introduce historical information v_{k-1}^x and u_{k-1}, and employ the moving average technique for acceleration, specifically by forming a convex combination of v_{k-1}^x and D_{k-1}^x.

Theorem I.1. ***(Expection form problem (1))***

Fix an iteration $K>1$ and assume that Assumption 3.1 to 3.3 hold. The mini-batch size S is chosen to be K^q Then there exist positive constants c_α, c_β, c_γ and c_ρ such that if

$$\begin{aligned}
\alpha_k = c_\alpha K^{-p}, &\quad \beta_k = c_\beta K^{-p}, \\
\gamma_k = c_\gamma K^{-p}, &\quad \rho_k = c_\rho K^{-p},
\end{aligned}$$

the iterates in MA-SOBA-q satisfy

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right]=\mathcal{O}\left(\frac{1}{K^{1-p}}+\frac{1}{K^{p+q}}\right).$$

Remark I.2. In Theorem I.1, we discuss the trade-off between step sizes and mini-batch sizes, their exponents need to satisfy $q+2p=1$, ensuring that the sampling complexity is $\mathcal{O}(\epsilon^{-2})$.

Algorithm 5 MA-SOBA-q

```
Input: Initializations $(x_{-1}, y_{-1}, z_{-1})$, $(x_0, y_0, z_0)$, and $v_{-1}^x$, number of total iterations $K$, step size $\{\alpha_k, \beta_k, \gamma_k\}$, momentum parameter$\rho_k$;
for $k = 0$ to $K-1$ do
    Sample $\mathcal{S}_1$ for $f$ and $\mathcal{S}_2$ for $g$;
    $v_k^x = (1-\rho_{k-1})v_{k-1}^x + \rho_{k-1} D_{k-1}^x$;
    $x_{k+1} = x_k - \alpha_k v_k^x$;
    $D_k^x = \nabla_1 F(x_k, y_k; \mathcal{S}_1) - \nabla_{12}^2 G(x_k, y_k; \mathcal{S}_2) z_k$;
    $v_k^y = \nabla_2 G(x_k, y_k; \mathcal{S}_2)$;
    $y_{k+1} = y_k - \beta_k v_k^y$;
    $v_k^z = \nabla_{22}^2 G(x_k, y_k; \mathcal{S}_2) z_k - \nabla_2 F(x_k, y_k; \mathcal{S}_1)$;
    $z_{k+1} = z_k - \gamma_k v_k^z$.
end for
```

Lemma I.3. *Under the Assumption 3.1 to 3.3, we have*

$$
\begin{aligned}
\mathbb{E}\left[\|D_y(x_k, y_k, z_k) - D_k^y\|^2\right] &\le \frac{\sigma_{g,1}^2}{|\mathcal{S}_2|}, \\
\mathbb{E}\left[\|D_z(x_k, y_k, z_k) - D_k^z\|^2\right] &\le \frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}(R^2 + \mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right]) + \frac{\sigma_f^2}{|\mathcal{S}_1|}, \\
\mathbb{E}\left[\|D_x(x_k, y_k, z_k) - D_k^x\|^2\right] &\le \frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}(R^2 + \mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right]) + \frac{\sigma_f^2}{|\mathcal{S}_1|}.
\end{aligned}
$$

Proof. Based on the definition of D_k^y and Assumption 3.3, we have

$$
\mathbb{E}\left[\|D_y(x_k, y_k, z_k) - D_k^y\|^2\right] = \mathbb{E}[\|D_y(x_k, y_k, z_k) - \frac{1}{|\mathcal{S}_2|}\sum_{\zeta \in \mathcal{S}_2} \nabla_2 G_j(x_k, y_k; \zeta)\|^2] \le \frac{\sigma_{g,1}^2}{|\mathcal{S}_2|}.
$$

$$
\begin{aligned}
\mathbb{E}\left[\|D_z(x_k, y_k, z_k) - D_k^z\|^2\right] &= \mathbb{E}\left[\left\|\nabla_{22}^2 g(x_k, y_k) z_k - \nabla_2 f(x_k, y_k) - \nabla_{22}^2 G(x_k, y_k; \mathcal{S}_2) z_k + \nabla_2 F(x_k, y_k; \mathcal{S}_1)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\nabla_{22}^2 g(x_k, y_k) z_k - \nabla_{22}^2 G(x_k, y_k; \mathcal{S}_2) z_k\right\|^2\right] + \mathbb{E}\left[\|\nabla_2 f(x_k, y_k) - \nabla_2 F(x_k, y_k; \mathcal{S}_1)\|^2\right] \\
&\le \mathbb{E}\left[\left\|\nabla_{22}^2 g(x_k, y_k) - \nabla_{22}^2 G(x_k, y_k; \mathcal{S}_2)\right\|^2\right] (\mathbb{E}[\|z_k - z^*(x_k)\|^2] + \mathbb{E}[\|z^*(x_k)\|^2]) \\
&\quad + \mathbb{E}\left[\|\nabla_2 f(x_k, y_k) - \nabla_2 F(x_k, y_k; \mathcal{S}_1)\|^2\right] \\
&\le \frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}(R^2 + \mathbb{E}\left[\|z_k - z^*(x_k)\|^2\right]) + \frac{\sigma_f^2}{|\mathcal{S}_1|},
\end{aligned}
$$

where the second equation holds because mini-batch estimation is unbiased. The last inequation is due to Assumption 3.3 and Lemma D.3. Similarly, we can obtain inequalities regarding D_k^x. □

Theorem I.4. ***(Restatement of Theorem I.1)***

Fix an iteration $K > 1$ and assume that Assumption 3.1 to 3.3 hold. Let $|\mathcal{S}_1| = |\mathcal{S}_2| = K^q$. The step sizes α_k, β_k, γ_k, and ρ_k have the same order of $\Theta(K^{-p})$, $p > 0$, and satisfy

$$
\begin{aligned}
\alpha_k &\le \min\{\frac{1}{2L^H}, \frac{1}{16L_{y^*}^2}\beta_k, \frac{1}{64(L^H)^2}\rho_k, 2\rho_k\}, \quad \beta_k \le \min\{\frac{4}{\mu}, \frac{\mu}{16\Delta}\beta_k\}, \\
\rho_k &\le \min\{\frac{\mu^2}{24\Delta}\beta_k, \frac{\mu^2}{24(L_1^g)^2}\gamma_k, 1\}, \quad \rho_k^2 \le \frac{\mu^2}{8}\gamma_k.
\end{aligned}
$$

Then the iterates in SRMBA satisfy

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right]=\mathcal{O}(\frac{1}{K^{1-p}}+\frac{1}{K^{p+q}}).$$

Proof. Consider the Lyapunov function in the form of

$$L_k=\mathbb{E}\left[H\left(x_k\right)\right]+A\mathbb{E}\left[\left\|y_k-y^*\left(x_k\right)\right\|^2\right]+B\mathbb{E}\left[\left\|z_k-z^*\left(x_k\right)\right\|^2\right]+C\mathbb{E}[\|\nabla H(x_k)-v_k^x\|^2]. \tag{26}$$

In Lemma E.2, we provide the descent lemma for the second and third terms in (26), and Lemma D.7 provides the descent lemma for the first term. Related to the last term, refer to Lemma E.1. Therefore, we have

$$\begin{aligned}
L_{k+1}-L_k =&\mathbb{E}\left[H\left(x_{k+1}\right)\right]-E\left[H\left(x_k\right)\right]+A\left(\mathbb{E}\left[\left\|y_{k+1}-y^*\left(x_{k+1}\right)\right\|^2\right]-\mathbb{E}\left[\left\|y_k-y^*\left(x_k\right)\right\|^2\right]\right)\\
&+B\left(\mathbb{E}\left[\left\|z_{k+1}-z^*\left(x_{k+1}\right)\right\|^2\right]-\mathbb{E}\left[\left\|z_k-z^*\left(x_k\right)\right\|^2\right]\right)\\
&+C\left(\mathbb{E}[\|\nabla H(x_{k+1})-v_{k+1}^x\|^2]-\mathbb{E}[\|\nabla H(x_k)-v_k^x\|^2]\right)\\
&\leq-\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right]\\
&+\left(\frac{L^H\alpha_k^2}{2}-\frac{\alpha_k}{2}+A\frac{2L_{y^*}^2\alpha_k^2}{\beta_k\mu}+B\frac{3L_{z^*}^2\alpha_k^2}{\gamma_k\mu}+C\frac{2\left(L^H\right)^2\alpha_k^2}{\rho_k}\right)\mathbb{E}\left[\|v_k^x\|^2\right]\\
&+\left(-A\beta_k\mu+8\Delta B\gamma_k+6C\Delta\rho_k\right)\mathbb{E}\left[\|y_k-y^*(x_k)\|^2\right]\\
&+\left(-B\gamma_k\mu+2B\gamma_k^2\frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}+6C\left(L_1^g\right)^2\rho_k+\rho_k^2C\frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}\right)\mathbb{E}\left[\left\|z_k-z^*\left(x_k\right)\right\|^2\right]\\
&+\left(\frac{\alpha_k}{2}-C\rho_k\right)\mathbb{E}\left[\left\|\nabla H\left(x_k\right)-v_k^x\right\|^2\right]\\
&+2A\beta_k^2\frac{\sigma_{g,1}^2}{|\mathcal{S}_2|}+2B\gamma_k^2\left(\frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}R^2+\frac{2\sigma_f^2}{|\mathcal{S}_1|}\right)+2C\rho_k^2\left(\frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}R^2+\frac{2\sigma_f^2}{|\mathcal{S}_1|}\right),
\end{aligned}$$

where the inequality holds by utilizing Lemma D.6 and Lemma I.3. Furthermore, we have

$$L_{k+1}-L_k\leq-\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right]+2A\beta_k^2\frac{\sigma_{g,1}^2}{|\mathcal{S}_2|}+2B\gamma_k^2\left(\frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}R^2+\frac{2\sigma_f^2}{|\mathcal{S}_1|}\right)+2C\rho_k^2\left(\frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}R^2+\frac{2\sigma_f^2}{|\mathcal{S}_1|}\right), \tag{27}$$

if the following system of inequalities holds

$$\begin{cases}
\frac{L^H\alpha_k^2}{2}-\frac{\alpha_k}{2}+A\frac{2L_{y^*}^2\alpha_k^2}{\beta_k\mu}+B\frac{3L_{z^*}^2\alpha_k^2}{\gamma_k\mu}+C\frac{2\left(L^H\right)^2\alpha_k^2}{\rho_k}\leq 0,\\
-A\beta_k\mu+8\Delta B\gamma_k+6C\Delta\rho_k\leq 0,\\
-B\gamma_k\mu+2B\gamma_k^2\frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}+6C\left(L_1^g\right)^2\rho_k+\rho_k^2C\frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}\leq 0,\\
\frac{\alpha_k}{2}-C\rho_k\leq 0.
\end{cases} \tag{28}$$

We choose the coefficients of the Lyapunov function to be $A=B=\mu$, $C=1$. In fact, it is only necessary to require that

α_k, β_k, γ_k , and ρ_k have the same order of $\Theta(K^{-p})$, $p > 0$, and satisfy

$$\begin{aligned}
\alpha_k \le & \min\{\frac{1}{2L^H}, \frac{1}{16L_{y^*}^2}\beta_k, \frac{1}{64(L^H)^2}\rho_k, 2\rho_k\}, \\
\beta_k \le & \min\{\frac{4}{\mu}, \frac{\mu}{16\Delta}\beta_k\}, \\
\rho_k \le & \min\{\frac{\mu^2}{24\Delta}\beta_k, \frac{\mu^2}{24(L_1^g)^2}\gamma_k, 1\}, \\
\rho_k^2 \le & \frac{\mu^2}{8}\gamma_k,
\end{aligned}$$

then (28) holds. Rearranging (27), we have

$$\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] \le L_k - L_{k+1} + 2\mu\beta_k^2\frac{\sigma_{g,1}^2}{|\mathcal{S}_2|} + 2\mu\gamma_k^2\left(\frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}R^2 + \frac{2\sigma_f^2}{|\mathcal{S}_1|}\right) + 2\rho_k^2\left(\frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}R^2 + \frac{2\sigma_f^2}{|\mathcal{S}_1|}\right)$$

Summing and telescoping yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] \le \frac{2L_0}{K} + \frac{1}{K}\sum_{k=0}^{K-1}\left(2\mu\beta_k^2\frac{\sigma_{g,1}^2}{|\mathcal{S}_2|} + 2\mu\gamma_k^2\left(\frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}R^2 + \frac{2\sigma_f^2}{|\mathcal{S}_1|}\right) + 2\rho_k^2\left(\frac{2\sigma_{g,2}^2}{|\mathcal{S}_2|}R^2 + \frac{2\sigma_f^2}{|\mathcal{S}_1|}\right)\right),$$

let $|\mathcal{S}_1| = |\mathcal{S}_2| = K^q$, then we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla H\left(x_k\right)\right\|^2\right] = \mathcal{O}(\frac{1}{K^{1-p}} + \frac{1}{K^{p+q}}).$$

□