Distributed Bilevel Optimization with Communication Compression

Yutong He [*1] **Jie Hu** [*1] **Xinmeng Huang** [2] **Songtao Lu** [3] **Bin Wang** [4] **Kun Yuan** [1 5 6]

Abstract

Stochastic bilevel optimization tackles challenges involving nested optimization structures. Its fast-growing scale nowadays necessitates efficient distributed algorithms. In conventional distributed bilevel methods, each worker must transmit full-dimensional stochastic gradients to the server every iteration, leading to significant communication overhead and thus hindering efficiency and scalability. To resolve this issue, we introduce the *first* family of distributed bilevel algorithms with communication compression. The primary challenge in algorithmic development is mitigating bias in hypergradient estimation caused by the nested structure. We first propose C-SOBA, a simple yet effective approach with unbiased compression and provable linear speedup convergence. However, it relies on strong assumptions on bounded gradients. To address this limitation, we explore the use of moving average, error feedback, and multi-step compression in bilevel optimization, resulting in a series of advanced algorithms with relaxed assumptions and improved convergence properties. Numerical experiments show that our compressed bilevel algorithms can achieve $10\times$ reduction in communication overhead without severe performance degradation.

1. Introduction

Large-scale optimization and learning have emerged as indispensable tools in numerous applications. Solving such large and intricate problems poses a formidable challenge, usually demanding hours or days to complete. Consequently, it is imperative to expedite large-scale optimization and learning with distributed algorithms. In distributed learning, multiple workers collaborate to solve a global problem through inter-worker communications. In most current implementations (Smola & Narayanamurthy, 2010; Li et al., 2014; Strom, 2015; Gibiansky, 2017), each worker transmits full-dimensional gradients to a central server for updating model parameters. Since the size of full-dimensional gradients is massive, communicating them per iteration incurs substantial overhead, which impedes algorithmic efficiency and scalability (Seide et al., 2014; Chilimbi et al., 2014).

To mitigate this issue, communication compression (Alistarh et al., 2017; Bernstein et al., 2018; Stich et al., 2018; Richtárik et al., 2021; Huang et al., 2022) has been developed to reduce overhead. Instead of transmitting full gradient/model tensors, these strategies communicate compressed tensors with substantially smaller sizes per iteration. Two prevalent approaches of compression are quantization and sparsification. Quantization (Alistarh et al., 2017; Horvath et al., 2019; Seide et al., 2014) involves mapping input tensors from a large, potentially infinite, set to a smaller set of discrete values, such as 1-bit quantization (Seide et al., 2014) or natural compression (Horvath et al., 2019). In contrast, sparsification (Wangni et al., 2018; Stich et al., 2018; Safaryan et al., 2021) entails dropping a certain number of entries to obtain sparse tensors for communication, such as rand-K or top-K compressor (Stich et al., 2018). Both approaches have demonstrated strong empirical performance in communication savings.

Communication compression has widespread application in single-level stochastic optimization. However, many machine learning tasks, including adversarial learning (Madry et al., 2018), meta-learning (Bertinetto et al., 2019), hyperparameter optimization (Franceschi et al., 2018), reinforcement learning (Hong et al., 2023), neural architecture search (Liu et al., 2018), and imitation learning (Arora et al., 2020) involve upper- and lower-level optimization formulations that go beyond the conventional single-level paradigm. Addressing such nested problems has prompted substantial attention towards stochastic bilevel optimization (Ghadimi & Wang, 2018; Ji et al., 2021). While tremendous efforts have been made (Yang et al., 2022; Chen et al., 2022; Tarzanagh et al., 2022; Yang et al., 2023b) to solve distributed stochastic bilevel optimization, *no existing algorithms, to the best of our knowledge, have been developed under communication compression*. To fill this gap, this paper provides the

[*]Equal contribution [1]Peking University [2]University of Pennsylvania [3]IBM Research [4]Zhejiang University [5]National Engineering Labratory for Big Data Analytics and Applications [6]AI for Science Institute, Beijing, China. Correspondence to: Kun Yuan <kunyuan@pku.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

first comprehensive study on distributed stochastic bilevel optimization with communication compression.

1.1. Distributed Bilevel Optimization

We consider distributed stochastic bilevel problems with the following nested upper- and lower-level structure:

$$\min_{x\in\mathbb{R}^{d_x}} \quad \Phi(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(x, y^\star(x)), \tag{1a}$$

$$\text{s.t.} \quad y^\star(x) = \operatorname*{argmin}_{y\in\mathbb{R}^{d_y}} \; g(x,y) \triangleq \frac{1}{n}\sum_{i=1}^{n} g_i(x,y). \tag{1b}$$

Here, n denotes the number of workers, with each worker i privately owns its upper-level cost function $f_i : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$, lower-level cost function $g_i : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$, and local data distribution $\mathcal{D}_{f_i}, \mathcal{D}_{g_i}$ such that

$$f_i(x,y) \triangleq \mathbb{E}_{\phi\sim\mathcal{D}_{f_i}}[F(x,y;\phi)],$$
$$g_i(x,y) \triangleq \mathbb{E}_{\xi\sim\mathcal{D}_{g_i}}[G(x,y;\xi)].$$

The objective for all workers is to find a global solution to bilevel problem (1). Typical applications of problem (1) can be found in (Yang et al., 2021; Tarzanagh et al., 2022).

1.2. Challenges in Compressed Bilevel Optimization

Conceptually, if each worker i could access the accurate oracle function $f_i^\star(x) \triangleq f_i(x, y^\star(x))$ without any sampling noise, a straightforward framework to solve (1) under compressed communication (with compressors $\{\mathcal{C}_i\}_{i=1}^n$) is

$$x^{k+1} = x^k - \frac{1}{n}\sum_{i=1}^{n} \mathcal{C}_i\big(\nabla f_i^\star(x^k)\big), \tag{2}$$

where each worker transmits the compressed hypergradient to the server to update model parameters. However, update (2) demands an accurate estimate of the hypergradient $\nabla f_i^\star(x)$, which can be written as (Ghadimi & Wang, 2018)

$$\nabla f_i^\star(x) = \nabla_x f_i(x, y^\star(x)) - \Big(\nabla^2_{xy} g(x, y^\star(x)) \cdot \big[\nabla^2_{yy} g(x, y^\star(x))\big]^{-1} \cdot \nabla_y f_i(x, y^\star(x))\Big) \tag{3}$$

with

$$\nabla^2_{xy} g(x, y^\star(x)) = \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{xy}\, g_i(x, y^\star(x)), \tag{4a}$$

$$\nabla^2_{yy} g(x, y^\star(x)) = \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{yy}\, g_i(x, y^\star(x)). \tag{4b}$$

It is challenging to precisely evaluate $\nabla f_i^\star(x)$ through (3) in distributed bilevel optimization, particularly under compressed communication, for the following reasons:

- **Unavailable $y^\star(x)$.** The solution $y^\star(x)$ to problem (1b) is not directly accessible. Existing literature (Ghadimi & Wang, 2018; Ji et al., 2021) often introduces iterative loops to approximately solve problem (1b), leading to expensive computation costs and biased estimates of $y^\star(x)$.
- **Inexact Hessian inversion.** Even provided with the accurate $y^\star(x)$, it is cumbersome to evaluate the global $\nabla^2_{xy} g(x, y^\star(x))$ and $[\nabla^2_{yy} g(x, y^\star(x))]^{-1}$ through (4) as it incurs expensive matrix communication. Recent works (Tarzanagh et al., 2022; Xiao & Ji, 2023) propose to communicate imprecise Hessian/Jacobian-vector products achieved by approximate implicit differentiation, which inevitably introduces bias in estimating $\nabla f_i^\star(x)$.
- **Compression-incurred distortion.** As indicated by (3) and (4), specific Jacobean matrices shall be communicated to tackle sub-problem (1b). One may consider using the compressed proxies $\mathcal{C}_i(\nabla^2_{yy}\, g_i(x, y^\star(x)))$ and $\mathcal{C}_i(\nabla^2_{xy}\, g_i(x, y^\star(x)))$ to replace Jacobians matrices in (4). However, the compression incurs information distortion, which brings additional bias when evaluating $\nabla f_i^\star(x)$.

To summarize, practical bilevel algorithms with communication compression essentially perform

$$x^{k+1} = x^k - \frac{1}{n}\sum_{i=1}^{n} \mathcal{C}_i\big(\nabla f_i^\star(x^k) + \textbf{bias}\big) \tag{5}$$

rather than (2), where the **bias** originates from the nested bilevel structure of (1), as opposed to data sampling or communication compression. This bias term poses substantial challenges in developing distributed bilevel algorithms with communication compression. Most existing single-level compression techniques, including error feedback (Stich et al., 2018; Richtárik et al., 2021) and multi-step compression (Huang et al., 2022; He et al., 2023a), require unbiased estimates of gradients[1] (*i.e.*, $\nabla f_i^\star(x)$) every iteration, and are thus not directly applicable to bilevel problem (1). This calls for the urgent need to develop new algorithms that can effectively mitigate the bias incurred by the nested bilevel structure, as well as new analyses to clarify how this bias impacts the convergence of compressed bilevel algorithms.

1.3. Contributions and Main Results

Contributions. This paper develops the first set of bilevel algorithms with communication compression.

- SOBA (Dagréou et al., 2022) is a newly introduced single-loop bilevel algorithm with lightweight communication and computation. While SOBA still suffers from biased hypergradient estimates, we surprisingly find applying unbiased compression directly to SOBA yields a simple yet effective compressed bilevel algorithm, which is denoted as distributed **SOBA** with communication **C**ompression, or **C-SOBA** for brevity. Under the strong assumption of bounded gradients, we establish its convergence as well as computational and communication complexities[2].

[1] While error feedback and multi-step compression accommodate biased compressors, they need accurate or unbiased gradients.

[2] Throughout the paper, the computational and communication complexities are referred to in an asymptotic sense, see Table 1.

Table 1. Comparison between distributed bilevel algorithms with communication compression. For simplicity, we unify the compression variance and heterogeneity bounds in both upper and lower levels. Notation n is the number of workers, ϵ is the target precision such that $\mathbb{E}\left[\|\nabla\Phi(\hat{x})\|_2^2\right] \leq \epsilon$, ω is compression-related parameter (see Assumption 2.4), σ^2 is the variance of stochastic oracles, b^2 bounds the gradient dissimilarity. We also list the best-known single-level compression algorithm in the bottom line for reference.

Algorithms	**#A. Comp.**†	**#A. Comm.**$^{\diamond}$	**Single Loop**	**Mechanism**‡	**Heter. Asp.***	**Implement**$^{\triangleleft}$
C-SOBA (Alg. 1 green)	$\frac{(1+\omega)\sigma^2+\omega b^2}{n\epsilon^2}$	$\frac{\omega b^2}{n\epsilon^2} + \frac{1+\omega/n}{\epsilon}^{\triangleright}$	✓	—	BG + BGD	☺
CM-SOBA (Alg. 1 pink)	$\frac{(1+\omega)\sigma^2+\omega b^2}{n\epsilon^2}$	$\frac{\omega b^2}{n\epsilon^2} + \frac{1+\omega/n}{\epsilon}$	✓	MA	BGD	☺
+MSC (Alg. 4)	$\frac{\sigma^2}{n\epsilon^2}$	$\frac{1+\omega}{\epsilon}$	✘	MA + MSC	BGD	☺
EF-SOBA (Alg. 2)	$\frac{(1+\omega)^7\sigma^2}{n\epsilon^2}$	$\frac{1+\omega+\omega^5/n}{\epsilon}$	✓	EF + MA	None	☹
+MSC (Alg. 5)	$\frac{\sigma^2}{n\epsilon^2}$	$\frac{1+\omega}{\epsilon}$	✘	EF + MSC	None	☹
Single-level EF21-SGDM (Fatkhullin et al., 2023)	$\frac{\sigma^2}{n\epsilon^2}$	$\frac{1+\omega}{\epsilon}$	✓	EF + MA	None	☹

$^{\diamond}$ Asymptotic communication complexity: number of communication rounds when $\sigma \to 0$ (smaller is better).
† Asymptotic computational complexity: number of gradient/Jacobian evaluations per worker when $\epsilon \to 0$ (smaller is better).
‡ Compression mechanisms: "MA", "EF", and "MSC" refer to as moving average, error feedback, and multi-step compression.
* Data heterogeneity assumptions (fewer/milder is better). "BG" and "BGD" denote bounded gradients (Assumption 3.2) and bounded gradient dissimilarity (Assumptions 3.1 and 4.1), respectively. "BG" is much more restrictive than "BGD".
$^{\triangleleft}$ Easy to implement or not. ☺ indicates "easy to implement" and ☹ indicates "relatively harder to implement" due to the EF mechanism.
$^{\triangleright}$ With gradient upper bound B_x (Assumption 3.2), a more precise complexity is $\frac{\omega b^2}{n\epsilon^2} + \frac{\sqrt[3]{\omega(n+\omega)^2 b^2 B_x^4}}{n\epsilon^{4/3}} + \frac{(1+\omega/n)(1+B_x)}{\epsilon}$.

- While commonly adopted in literature, the bounded-gradient assumption of C-SOBA is restrictive. To address this limitation, we leverage **M**oving average to enhance the theoretical performance of C-SOBA, proposing the refined **CM-SOBA** method. CM-SOBA converges under the more relaxed assumption of bounded heterogeneity with improved complexities, compared to C-SOBA.
- The convergence of CM-SOBA still relies on the magnitude of data heterogeneity. When local data distributions $\mathcal{D}_{f_i}$ and $\mathcal{D}_{g_i}$ differ drastically across workers, the performance of C-SOBA and CM-SOBA substantially degrade. To mitigate this issue, we further incorporate **E**rror **F**eedback into CM-SOBA, leading to the **EF-SOBA** algorithm. EF-SOBA does not rely on any assumptions regarding data heterogeneity, making it suitable for applications with severe data heterogeneity.
- Finally, owing to the bias in (5) incurred by the nested structure, the established communication and computation complexities of the aforementioned compressed bilevel algorithms are less favorable compared to the best-known single-level compressed algorithms (Huang et al., 2022; Fatkhullin et al., 2023). Consequently, we utilize multi-step compression to enhance the convergence of C-SOBA and EF-SOBA, attaining the same complexities as the best-known single-level compressed algorithms.

Results in Table 1. All established algorithms, along with their assumptions and complexities are listed in Table 1. It is noteworthy that the utilization of more advanced mechanisms relaxes assumptions and substantially improves complexities, albeit introducing increased intricacy in algorithmic structures and implementations. Furthermore, our algorithms can achieve the same theoretical complexities as the best-known single-level compression algorithm (Fatkhullin et al., 2023), demonstrating its efficacy in overcoming the bias incurred by the nested structure in bilevel problems.

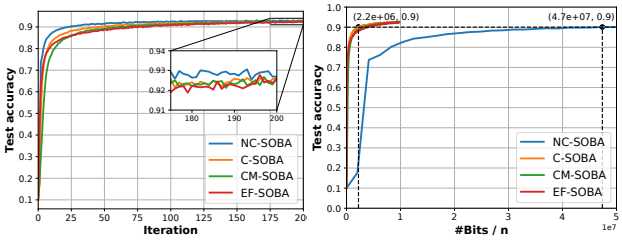

Figure 1. Hyper-representation on MNIST under homogeneous data distributions. *NC-SOBA indicates non-compressed SOBA.*

Experiments. Our numerical experiments demonstrate that the proposed algorithms can achieve $10\times$ reduction in communicated bits, compared to non-compressed distributed bilevel algorithms, see Fig. 1 and Sec. 7.

Analysis. Our analysis also contributes new insights. They furnish convergence guarantees even when utilizing biased gradient estimates in compressors as shown in (5). Additionally, they elucidate how upper- and lower-level compression exerts distinct influences on convergence, enlightening the compressor selection for upper- and lower-level problems.

1.4. Related Work

Bilevel optimization. A key challenge in bilevel optimization lies in accurately estimating the hypergradient $\nabla\Phi(x)$. Various algorithms have emerged to tackle this challenge, employing techniques such as approximate implicit differentiation (Domke, 2012; Ghadimi & Wang, 2018; Grazzi et al.,

2020; Ji et al., 2021), iterative differentiation (Franceschi et al., 2018; Maclaurin et al., 2015; Domke, 2012; Grazzi et al., 2020; Ji et al., 2021), and Neumann series (Chen et al., 2021; Hong et al., 2023). However, these methods introduce additional inner loops that lead to increased computational overhead and deteriorated computation complexity. A recent work (Dagréou et al., 2022) introduces SOBA, a novel single-loop framework, to enable simultaneous updates of the lower- and upper-level variables. Recent efforts have been made to develop distributed bilevel algorithms within the federated learning setup (Tarzanagh et al., 2022; Yang et al., 2021; Huang et al., 2023a) and decentralized scenarios (Yang et al., 2022; Chen et al., 2022; 2023a; Lu et al., 2022a; Gao et al., 2023). However, bilevel optimization with communication compression has not been studied in existing literature to our knowledge.

Communication compression. Communication compression shows notable success in single-level distributed optimization (Alistarh et al., 2017; Bernstein et al., 2018; Stich et al., 2018). Two main approaches are quantization and sparsification. Quantization strategies include Sign-SGD (Seide et al., 2014; Bernstein et al., 2018), TurnGrad (Wen et al., 2017), and natural compression (Horvath et al., 2019). On the other hand, classical sparsification strategies involve rand-K and top-K (Stich, 2019; Wangni et al., 2018). Compression introduces information distortion, which slows down convergence and incurs more communication rounds to achieve desired solutions. Various advanced techniques such as error feedback (Richtárik et al., 2021; Stich et al., 2018), multi-step compression (Huang et al., 2022; He et al., 2023a), and momentum (Fatkhullin et al., 2023; Huang et al., 2023b) are developed to effectively mitigate the impact of compression-incurred errors. Furthermore, the optimal convergence for single-level distributed optimization with communication compression is established in (Huang et al., 2022; He et al., 2023a). However, none of these results have been established for bilevel stochastic optimization.

2. Preliminaries

Notations. For a second-order continuously differentiable function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$, we denote $\nabla_x f(x,y)$ and $\nabla_y f(x,y)$ as the partial gradients in terms of x and y, respectively. Correspondingly, $\nabla^2_{xy} f(x,y) \in \mathbb{R}^{d_x \times d_y}$ and $\nabla^2_{yy} f(x,y) \in \mathbb{R}^{d_y \times d_y}$ represent its Jacobian matrices. The full gradient of f is represented as $\nabla f(x,y) \triangleq \left(\nabla_x f(x,y)^\top, \nabla_y f(x,y)^\top\right)^\top$.

Basic assumptions. We now introduce some basic assumptions needed throughout theoretical analysis.

Assumption 2.1 (CONTINUITY)**.** For any i $(1 \le i \le n)$,

- function f_i is C_f-Lipschitz continuous with respect to y;
- functions ∇f_i, ∇g_i, $\nabla^2_{xy} g_i$, $\nabla^2_{yy} g_i$ are Lipschitz continuous with constants L_f, L_g, $L_{g_{xy}}$, $L_{g_{yy}}$, respectively.

Assumption 2.2 (STRONG CONVEXITY)**.** For any i $(1 \le i \le n)$, g_i is μ_g-strongly convex with respect to y.

In the above assumptions, we allow data heterogeneity to exist across different workers, *i.e.*, every (f_i, g_i) differs from each other. Furthermore, we do not assume the Lipschitz continuity of f_i with respect to x, which relaxes the assumptions used in prior works (Lu et al., 2022b; Yang et al., 2023a; Huang et al., 2023a; Chen et al., 2022; Lu et al., 2022a; Yang et al., 2022).

Assumption 2.3 (STOCHASTIC NOISE)**.** There exists $\sigma \ge 0$ such that for any i $(1 \le i \le n)$, and any $x \in \mathbb{R}^{d_x}$, $y \in \mathbb{R}^{d_y}$,

- the gradient oracles satisfy:

$$\begin{aligned}
&\mathbb{E}_{\phi\sim\mathcal{D}_{f_i}}[\nabla F(x,y;\phi)] = \nabla f_i(x,y),\\
&\mathbb{E}_{\xi\sim\mathcal{D}_{g_i}}[\nabla_y G(x,y;\xi)] = \nabla_y g_i(x,y),\\
&\mathbb{E}_{\phi\sim\mathcal{D}_{f_i}}\left[\|\nabla F(x,y;\phi) - \nabla f_i(x,y)\|_2^2\right] \le \sigma^2,\\
&\mathbb{E}_{\xi\sim\mathcal{D}_{g_i}}\left[\|\nabla_y G(x,y;\xi) - \nabla_y g_i(x,y)\|_2^2\right] \le \sigma^2;
\end{aligned}$$

- the Jacobian oracles satisfy:

$$\begin{aligned}
&\mathbb{E}_{\xi\sim\mathcal{D}_{g_i}}\left[\nabla^2_{xy} G(x,y;\xi)\right] = \nabla^2_{xy} g_i(x,y),\\
&\mathbb{E}_{\xi\sim\mathcal{D}_{g_i}}\left[\nabla^2_{yy} G(x,y;\xi)\right] = \nabla^2_{yy} g_i(x,y),\\
&\mathbb{E}_{\xi\sim\mathcal{D}_{g_i}}\left[\left\|\nabla^2_{xy} G(x,y;\xi) - \nabla^2_{xy} g_i(x,y)\right\|_2^2\right] \le \sigma^2,\\
&\mathbb{E}_{\xi\sim\mathcal{D}_{g_i}}\left[\left\|\nabla^2_{yy} G(x,y;\xi) - \nabla^2_{yy} g_i(x,y)\right\|_2^2\right] \le \sigma^2.
\end{aligned}$$

The following notion is for compressors.

Assumption 2.4 (UNBIASED COMPRESSION)**.** A compressor $\mathcal{C}(\cdot) : \mathbb{R}^{d_c} \to \mathbb{R}^{d_c}$ is ω-unbiased $(\omega \ge 0)$, if for any input $x \in \mathbb{R}^{d_c}$, we have

$$\mathbb{E}[\mathcal{C}(x)] = x, \quad \text{and} \quad \mathbb{E}\left[\|\mathcal{C}(x) - x\|_2^2\right] \le \omega \|x\|_2^2.$$

Different compressors yield different values for ω. Generally speaking, a large ω indicates more aggressive compression and, consequently, induces more information distortion. Below, we also assume the conditional independence among all local compressors, *i.e.*, the outputs of local compressors are mutually independent, conditioned on the inputs.

3. Compressed SOBA

SOBA (Dagréou et al., 2022) is a single-loop bilevel algorithm with lightweight communication and computational costs, originally devised for single-node optimization. In this section, we extend SOBA to address distributed bilevel optimization (1) and then incorporate communication compression, resulting in our first compressed bilevel algorithm.

Non-compressed SOBA. To address the Hessian-inversion issue when evaluating the hypergradient $\nabla\Phi(x)$ for problem (1), SOBA introduces $z^\star \triangleq -[\nabla^2_{yy} g(x, y^\star(x))]^{-1} \cdot$

$\nabla_y f(x, y^\star(x))$, which can be viewed as the solution to the following distributed optimization problem:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{2}z^\top\nabla^2_{yy}g_i(x,y^\star(x))\,z+z^\top\nabla_y f_i(x,y^\star(x))\right\}.$$

By simultaneously solving the lower-level problem, estimating the Hessian-inverse-vector product, and minimizing the upper-level problem, we achieve distributed recursions:

$$x^{k+1} = x^k - \frac{\alpha}{n}\sum_{i=1}^{n}\left(\nabla^2_{xy}g_i(x^k,y^k)z^k + \nabla_x f_i(x^k,y^k)\right),$$
$$y^{k+1} = y^k - \frac{\beta}{n}\sum_{i=1}^{n}\nabla_y g_i(x^k,y^k),$$
$$z^{k+1} = z^k - \frac{\gamma}{n}\sum_{i=1}^{n}\left(\nabla^2_{yy}g_i(x^k,y^k)z^k + \nabla_y f_i(x^k,y^k)\right),$$

where a central server collects local information to update global variables. We call this algorithm non-compressed SOBA (NC-SOBA). NC-SOBA accommodates stochastic variants by introducing noisy gradient/Jacobian oracles.

Compressed SOBA. When each worker compresses its information before communicating with the central server, we obtain **C**ompressed **SOBA**, or C-SOBA for short. To detail the algorithm, we let each worker i independently sample data $\phi_i^k \sim \mathcal{D}_{f_i}$ and $\xi_i^k \sim \mathcal{D}_{g_i}$, and calculate

$$D^k_{x,i} \triangleq \nabla^2_{xy}G(x^k,y^k;\xi_i^k)z^k + \nabla_x F(x^k,y^k;\phi_i^k), \quad (6a)$$
$$D^k_{y,i} \triangleq \nabla_y G(x^k,y^k;\xi_i^k), \quad (6b)$$
$$D^k_{z,i} \triangleq \nabla^2_{yy}G(x^k,y^k;\xi_i^k)z^k + \nabla_y F(x^k,y^k;\phi_i^k). \quad (6c)$$

Next, worker i transmits $\mathcal{C}_i^u(D^k_{x,i})$, $\mathcal{C}_i^\ell(D^k_{y,i})$, and $\mathcal{C}_i^\ell(D^k_{z,i})$ to the central server where $\mathcal{C}_i^u$ and $\mathcal{C}_i^\ell$ are the upper-level ω_u- and lower-level ω_ℓ-unbiased compressors utilized by worker i. The implementation of C-SOBA is listed in Algorithm 1 where the update of x^{k+1} follows the green line. A Clip operation is conducted before updating z to boost the algorithmic performance, in which

$$\mathrm{Clip}(\tilde{z}^{k+1};\rho) \triangleq \min\left\{1, \rho/\|\tilde{z}^{k+1}\|_2\right\}\cdot\tilde{z}^{k+1}.$$

Intuition behind clipping operation. The variable z is intended to estimate $z^\star(x)$, which, under Assumptions 2.1 and 2.2, should not exceed a magnitude of C_f/μ_g (refer to Lemma B.2). However, during gradient descent steps, as in our algorithms, the update of z may surpass this limit. Therefore, it is natural to opt for the nearest neighbor of z with a magnitude no greater than C_f/μ_g instead. A rough estimation $\rho \geq C_f/\mu_g$ suffices as an upper bound. Alternatively, we can directly confine z to the domain $\mathcal{B}(0,\rho)$, the closed ball of dimension d_y centered at 0 with a radius of ρ, using projected gradient descent. Both approaches result in the projection operation $z^{k+1} = \mathcal{P}_{\mathcal{B}(0,\rho)}(\tilde{z}^{k+1})$, equivalent to the clipping operation $z^{k+1} = \mathrm{Clip}(\tilde{z}^{k+1},\rho)$. Here, $\mathcal{P}_\Omega(\cdot)$ denotes projection onto a closed convex set Ω.

Algorithm 1 C-SOBA and CM-SOBA

Input: α, β, γ, ρ, x^0, y^0, $z^0(\|z^0\|_2 \leq \rho)$, h_x^0;
for $k = 0, 1, \cdots, K-1$ **do**
  on each worker:
    Compute $D^k_{x,i}$, $D^k_{y,i}$, $D^k_{z,i}$ as in (6);
    Send $\mathcal{C}_i^u(D^k_{x,i}), \mathcal{C}_i^\ell(D^k_{y,i}), \mathcal{C}_i^\ell(D^k_{z,i})$ to the server;
  on server:
    $x^{k+1} = x^k - (\alpha/n)\sum_{i=1}^n \mathcal{C}_i^u(D^k_{x,i})$ (C-SOBA);
    $h_x^{k+1} = (1-\theta)h_x^k + (\theta/n)\sum_{i=1}^n \mathcal{C}_i^u(D^k_{x,i})$;
    $x^{k+1} = x^k - \alpha\cdot h_x^k$ (CM-SOBA);
    $y^{k+1} = y^k - (\beta/n)\sum_{i=1}^n \mathcal{C}_i^\ell(D^k_{y,i})$;
    $\tilde{z}^{k+1} = z^k - (\gamma/n)\sum_{i=1}^n \mathcal{C}_i^\ell(D^k_{z,i})$;
    $z^{k+1} = \mathrm{Clip}(\tilde{z}^{k+1};\rho)$;
    Broadcast $x^{k+1}, y^{k+1}, z^{k+1}$ to all workers;
end for

The justification for this operation in theory is straightforward; z^{k+1} is always closer than (or at an equal distance with) $\tilde{z}^{k+1}$ to $z^\star(x^{k+1})$.

In particular, assuming $\rho \geq C_f/\mu_g$, the non-expansiveness property of projection operators allows us to deduce:

$$\begin{aligned}&\left\|z^{k+1} - z^\star\left(x^{k+1}\right)\right\|_2\\ =&\left\|\mathcal{P}_{\mathcal{B}(0,\rho)}\left(\tilde{z}^{k+1}\right) - \mathcal{P}_{\mathcal{B}(0,\rho)}\left(z^\star\left(x^{k+1}\right)\right)\right\|_2\\ \leq&\left\|\tilde{z}^{k+1} - z^\star\left(x^{k+1}\right)\right\|_2.\end{aligned}$$

Convergence. To establish the convergence for C-SOBA, we need more assumptions beyond those discussed in Sec. 2.

Assumption 3.1 (Bounded Heterogeneity)**.** There exist constants $b_f \geq 0$, $b_g \geq 0$, such that for any $x \in \mathbb{R}^{d_x}$, $y \in \mathbb{R}^{d_y}$, it holds that

$$\begin{aligned}\|\nabla f_i(x,y) - \nabla f(x,y)\|_2^2 &\leq b_f^2,\\ \|\nabla_y g_i(x,y) - \nabla_y g(x,y)\|_2^2 &\leq b_g^2,\\ \left\|\nabla^2_{xy} g_i(x,y) - \nabla^2_{xy} g(x,y)\right\|_2^2 &\leq b_g^2,\\ \left\|\nabla^2_{yy} g_i(x,y) - \nabla^2_{yy} g(x,y)\right\|_2^2 &\leq b_g^2.\end{aligned} \quad (7)$$

For conciseness, we present results with the notation $b^2 \triangleq \max\{b_f^2, b_g^2\}$ in the main text and defer the detailed counterparts associated with b_f^2 and b_g^2 to Appendix B.

Assumption 3.2 (Bounded Gradients)**.** There exists constant $B_x \geq 0$, such that for any (x^k, y^k, z^k) generated

by C-SOBA(Alg. 1), we have

$$\left\|\nabla^2_{xy} g(x^k, y^k) z^k + \nabla_x f(x^k, y^k)\right\|_2^2 \leq B_x^2. \tag{8}$$

It is noteworthy that the above assumption is milder than the L_{f_x}-Lipschitz continuity of f with respect to x (Lu et al., 2022b; Yang et al., 2023a; Huang et al., 2023a; Chen et al., 2022; Lu et al., 2022a; Yang et al., 2022), which implies $B_x \triangleq \sqrt{2} L_g C_f / \mu_g + \sqrt{2} L_{f_x}$.

Assumption 3.3 (2ND-ORDER SMOOTHNESS)**.** Jacobian matrices $\nabla^2_{xy} g$, $\nabla^2_{yy} g$ are $L_{g_{xy}}$- and $L_{g_{yy}}$-smooth, $\nabla^2 f$ is L_{ff}-Lipschitz continuous.

Theorem 3.4. *Under Assumptions 2.1–2.4 and 3.1–3.3, if we set the hyperparameters as in Appendix B.1, C-SOBA converges as*

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla \Phi(x^k)\right\|_2^2\right] \\
=&\mathcal{O}\Bigg(\frac{\sqrt{(1+\omega_\ell+\omega_u)\Delta}\sigma + \sqrt{(\omega_\ell+\omega_u)\Delta} b}{\sqrt{nK}} \\
&+ \frac{\Delta^{\frac{3}{4}}((1+\omega_\ell)\sigma^2 + \omega_\ell b^2)^{\frac{1}{4}}\sqrt{1+\omega_u/n} B_x}{n^{1/4} K^{\frac{3}{4}}} \\
&+ \frac{\Delta^{\frac{3}{4}}\left((1+\omega_\ell)\sigma^2 + \omega_\ell b^2\right)^{\frac{1}{4}}\left((1+\omega_u)\sigma^2 + \omega_u b^2\right)^{\frac{1}{2}}}{(nK)^{3/4}} \\
&+ \frac{\sqrt{(1+\omega_u)(1+\omega_\ell/n)}\Delta\sigma + \sqrt{\omega_u(1+\omega_\ell/n)} b}{\sqrt{n}K} \\
&+ \frac{\sqrt{(1+\omega_\ell/n)(1+\omega_u/n)}\Delta B_x}{K} \\
&+ \frac{(1+\omega_\ell/n+\omega_u/n)\Delta}{K}\Bigg),
\end{aligned} \tag{9}$$

where $\Delta \triangleq \max\{\Phi(x^0), \|y^0 - y^\star(x^0)\|_2^2, \|z^0 - z^\star(x^0)\|_2^2\}$.

Asymptotic complexities. C-SOBA achieves asymptotic linear speedup with a rate of $\mathcal{O}(1/\sqrt{nK})$ as the number of iterations $K \to \infty$. This corresponds to an asymptotic sampling/computational complexity of $\mathcal{O}(1/(n\epsilon^2))$ as $\epsilon \to 0$. Furthermore, C-SOBA asymptotically requires $\mathcal{O}(\omega/(n\epsilon^2))$ communication rounds to achieve an ϵ-accurate solution when $\epsilon \to 0$, see more details in Table 1.

Consistency with non-compression methods. The leading term in (9) reduces to $\mathcal{O}(\sqrt{\Delta}\sigma/\sqrt{nK})$ when $\omega_\ell = \omega_u = 0$, which is consistent with non-compression algorithms.

A recommended choice of ω_u and ω_l. According to (He et al., 2023b), an ω-unbiased compressor $\mathcal{C}(x)$ with input dimension d will transmit at least $\mathcal{O}(d/(1+\omega))$ bits per communication round. If $\omega_u + \omega_\ell$ is bounded away from 0, C-SOBA will asymptotically transmit

$$\mathcal{O}\Bigg(\underbrace{\frac{(\omega_\ell+\omega_u)\Delta(\sigma^2+b^2)}{n\epsilon^2}}_{\text{asym. comm. rounds}} \cdot \underbrace{\left(\frac{d_x}{1+\omega_u} + \frac{d_y}{1+\omega_\ell}\right)}_{\text{bits per round}}\Bigg)$$

bits to achieve an ϵ-accurate solution. To minimize the communicated bits, a recommended choice for ω_ℓ and ω_u satisfies

$$(1+\omega_u)/(1+\omega_\ell) = \Theta\left(\sqrt{d_x/d_y}\right). \tag{10}$$

When using rand-K compressors, we can set different values of K for lower- and upper-level compression to achieve the recommended relation of ω_ℓ and ω_u that satisfy (10). We refer the readers to the ablation experiments on different choices in Appendix D.4.

4. CM-SOBA Algorithm

Although simple and effective, C-SOBA relies on strong assumptions, particularly Assumption 3.2 on bounded gradients. Moreover, the typically large upper bound B_x significantly hampers the convergence performance. These strong assumptions and inferior convergence complexities are attributed to the bias introduced by the nested bilevel structure, as elucidated in (5).

CM-SOBA. To enhance the convergence properties, we introduce a momentum procedure $h_x^{k+1} = (1-\theta)h_x^k + (\theta/n)\sum_{i=1}^n \mathcal{C}_i^u(D_{x,i}^k)$ to tweak the descent direction of x, *i.e.*, $x^{k+1} = x^k - \alpha h_x^k$. We refer to this new algorithm as **C**ompressed **SOBA** with **M**omentum, abbreviated as CM-SOBA. The implementation is outlined in Algorithm 1, where the update of x^{k+1} is indicated by the pink color. CM-SOBA is inspired by the momentum-based algorithm (Chen et al., 2023b) which eliminates the dependence on Assumption 3.2 in the single-node scenario.

Convergence. With an additional momentum step, CM-SOBA converges with more relaxed assumptions. In particular, it removes the strong assumption on bounded gradients.

Assumption 4.1 (POINT-WISE BOUNDED HETEROGENEITY)**.** There exist constants $b_f \geq 0$, $b_g \geq 0$, such that for any $x \in \mathbb{R}^{d_x}$, $y = y^\star(x)$, (7) holds.

Assumption 4.1 is weaker than Assumption 3.1 since it only assumes bounds at $(x, y^\star(x))$, as opposed to arbitrary (x, y). By writing $b^2 \triangleq \max\{b_f^2, b_g^2\}$, we have the follows.

Theorem 4.2. *Under Assumptions 2.1–2.4 and 4.1, if we set the hyperparameters as in Appendix B.2, CM-SOBA converges as*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla \Phi(x^k)\right\|_2^2\right]$$

$$=\mathcal{O}\left(\frac{\sqrt{(1+\omega_\ell+\omega_u)\Delta}\sigma+\sqrt{(\omega_\ell+\omega_u)\Delta}b}{\sqrt{nK}}+\frac{(1+\omega_\ell/n+\omega_u/n)\Delta}{K}\right), \tag{11}$$

in which $\Delta \triangleq \max\{\Phi(x^0), \|h_x^0-\nabla\Phi(x^0)\|_2^2, \|y^0-y^\star(x^0)\|_2^2, \|z^0-z^\star(x^0)\|_2^2\}$.

Improved complexities. CM-SOBA achieves an asymptotic linear speedup rate under more relaxed assumptions, compared to C-SOBA. Furthermore, by eliminating the influence of the gradient upper bound B_x on the convergence, we observe from (11) that CM-SOBA enjoys a faster rate, especially when B_x is large.

5. EF-SOBA Algorithm

Despite the simplicity, C-SOBA and CM-SOBA rely on the restrictive Assumptions 3.1 and 4.1 concerning bounded data heterogeneity. When local data distributions $\mathcal{D}_{f_i}$ and $\mathcal{D}_{g_i}$ differ drastically across workers, the bounded-data-heterogeneity assumption can be violated, significantly deteriorating the convergence performance of C-SOBA and CM-SOBA. This section is devoted to developing compressed bilevel algorithms that are robust to data heterogeneity.

Error feedback to upper-level compressors. When updating x in the upper-level optimization, each worker i in C-SOBA or CM-SOBA transmits $\mathcal{C}_i^u(D_{x,i}^k)$ to the central server at each iteration k. Since $D_{x,i}^k$ does not approach zero due to the sampling randomness and data heterogeneity, $\mathcal{C}_i^u(D_{x,i}^k)$ does not converge to $D_{x,i}^k$ either. This reveals that compression-incurred distortion persists even when $k\to\infty$, explaining why C-SOBA or CM-SOBA necessitates Assumption 3.1 or 4.1 to bound compression distortions.

Inspired by (Fatkhullin et al., 2023), we employ error feedback to alleviate the impact of data heterogeneity when solving the upper-level optimization. Consider recursions

$$h_{x,i}^{k+1} = (1-\theta)h_{x,i}^k + \theta\cdot D_{x,i}^k \tag{12a}$$
$$m_{x,i}^{k+1} = m_{x,i}^k + \delta_u\cdot\mathcal{C}_i^u(h_{x,i}^{k+1}-m_{x,i}^k), \tag{12b}$$
$$\hat{h}_x^{k+1} = \hat{h}_x^k + \frac{\delta_u}{n}\sum_{i=1}^n \mathcal{C}_i^u(h_{x,i}^{k+1}-m_{x,i}^k), \tag{12c}$$
$$x^{k+1} = x^k - \alpha\cdot\hat{h}_x^k, \tag{12d}$$

in which δ_u is a positive scaling coefficient, $m_{x,i}^k$ is an auxiliary variable to track $h_{x,i}^k$, and $\hat{h}_x^k = (1/n)\sum_{i=1}^n m_{x,i}^k$ holds for any $k\geq 0$. In each iteration k, it is the difference $h_{x,i}^k - m_{x,i}^k$ that is compressed and transmitted instead of $h_{x,i}^k$ itself. When $m_{x,i}^k$ converges to a fixed point as $k\to\infty$, we have $m_{x,i}^k\to h_{x,i}^k$ from (12b) and hence $\hat{h}_x^k\to(1/n)\sum_{i=1}^n h_{x,i}^k$. In other words, error feedback (12) removes the compression-incurred distortion asymptotically, making x update along the exact momentum direction even when data heterogeneity exists.

Error feedback to lower-level compressors. One may naturally wonder whether the same error feedback technique (12) can be used for the lower-level compressors, *i.e.*,

$$m_{y,i}^{k+1} = m_{y,i}^k + \delta_\ell\cdot\mathcal{C}_i^\ell(D_{y,i}^k - m_{y,i}^k), \tag{13a}$$
$$m_{z,i}^{k+1} = m_{z,i}^k + \delta_\ell\cdot\mathcal{C}_i^\ell(D_{z,i}^k - m_{z,i}^k), \tag{13b}$$
$$m_{y}^{k+1} = m_{y}^k + \frac{\delta_\ell}{n}\sum_{i=1}^n\mathcal{C}_i^\ell(D_{y,i}^k - m_{y,i}^k), \tag{13c}$$
$$m_{z}^{k+1} = m_{z}^k + \frac{\delta_\ell}{n}\sum_{i=1}^n\mathcal{C}_i^\ell(D_{z,i}^k - m_{z,i}^k), \tag{13d}$$

and let y and z update along the direction m_y and m_z. The answer, however, is *negative*. Since the hypergradient $D_{x,i}$ used in x-update relies heavily on the accurate values of y and z as shown in (6a), a more refined error feedback is needed for lower-level compressors. As illustrated in Appendix A, y and z must be updated following an unbiased estimate of their gradient direction. However, m_y and m_z provide biased estimates under the presence of δ_u. To address this issue, we propose using

$$\hat{D}_y^k = m_y^k + \frac{1}{n}\sum_{i=1}^n\mathcal{C}_i^\ell(D_{y,i}^k - m_{y,i}^k) \tag{14a}$$
$$\hat{D}_z^k = m_z^k + \frac{1}{n}\sum_{i=1}^n\mathcal{C}_i^\ell(D_{z,i}^k) - m_{z,i}^k). \tag{14b}$$

to update y and z for the unbiasedness (*i.e.*, $\mathbb{E}[\hat{D}_y^k] = \mathbb{E}[D_y^k] = \nabla_y G(x^k,y^k)$ and the similar applies to z). The resulting algorithm, termed as compressed **SOBA** with **E**rror **F**eedback, or EF-SOBA for short, is listed in Algorithm 2.

Theorem 5.1. *Under Assumptions 2.1–2.4, if hyperparameters are set as in Appendix B.3, EF-SOBA converges as*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right] \tag{15}$$
$$=\mathcal{O}\left(\frac{(1+\omega_u)^2(1+\omega_\ell)^{3/2}\sqrt{\Delta}\sigma}{\sqrt{nK}}+\frac{\omega_u^{1/3}(1+\omega_u)^{1/3}\Delta^{2/3}\sigma^{2/3}}{K^{2/3}}+\frac{(1+\omega_u)\Delta}{K}+\frac{(1+\omega_u)^2\sqrt{\omega_\ell(1+\omega_\ell)}\Delta}{\sqrt{n}K}+\frac{(1+\omega_u)^2\omega_\ell^3\Delta}{nK}\right)$$

where Δ represents algorithmic initialization constants detailed in Appendix B.3.

According to the above theorem, EF-SOBA achieves linear speedup convergence without relying on Assumption

Algorithm 2 EF-SOBA

Input: α, β, γ, θ, ρ, δ_u, δ_ℓ, x^0, y^0, $z^0(\|z^0\|_2 \leq \rho)$, $\{m^0_{x,i}\}$, $\{m^0_{y,i}\}$, $\{m^0_{z,i}\}$, $\{h^0_{x,i}\}$, $\hat{h}^0_x = \frac{1}{n}\sum_{i=1}^n m^0_{x,i}$, $m^0_y = \frac{1}{n}\sum_{i=1}^n m^0_{y,i}$, $m^0_z = \frac{1}{n}\sum_{i=1}^n m^0_{z,i}$;
for $k = 0, 1, \cdots, K-1$ **do**
  on worker:
  Compute $D^k_{x,i}$, $D^k_{y,i}$, $D^k_{z,i}$ as in (6);
  Update $h^{k+1}_{x,i}$ and $m^{k+1}_{x,i}$ as in (12a) and (12b);
  Update $m^{k+1}_{y,i}$ and $m^{k+1}_{z,i}$ as in (13a) and (13b);
  Send $\mathcal{C}^u_i(h^{k+1}_{x,i} - m^k_{x,i}), \mathcal{C}^\ell_i(D^k_{y,i} - m^k_{y,i}), \mathcal{C}^\ell_i(D^k_{z,i} - m^k_{z,i})$ to the server;
  on server:
  Update $\hat{D}^k_y$ and $\hat{D}^k_z$ as in (14a) and (14b);
  $x^{k+1} = x^k - \alpha \cdot \hat{h}^k_x$; $\quad y^{k+1} = y^k - \beta \cdot \hat{D}^k_y$;
  $\tilde{z}^{k+1} = z^k - \gamma \cdot \hat{D}^k_z$; $\quad z^{k+1} = \mathrm{Clip}(\tilde{z}^{k+1}, \rho)$;
  Update $\hat{h}^{k+1}_x$, m^{k+1}_y, m^{k+1}_z as in (12c), (13c), (13d);
  Broadcast $x^{k+1}, y^{k+1}, z^{k+1}$ to all workers;
end for

3.1 or 4.1. Furthermore, the convergence of EF-SOBA is unaffected by data heterogeneity b^2.

6. Convergence Acceleration

While C-SOBA, CM-SOBA, and EF-SOBA can converge, their computational and communication complexities are inferior to those of the best-known single-level compression algorithms, such as EF21-SGDM (Loizou & Richtárik, 2020) and NEOLITHIC (Huang et al., 2022). In Appendix C.2, we leverage **M**ulti-**S**tep **C**ompression (Huang et al., 2022) to expedite CM-SOBA and EF-SOBA, leading to CM-SOBA-MSC and EF-SOBA-MSC, respectively, to achieve the same complexities as EF21-SGDM and NEOLITHIC, see Table 1. EF-SOBA-MSC converges as follows:

Theorem 6.1. *Under Assumptions 2.1–2.4, with hyperparameters in Appendix C.2, EF-SOBA-MSC converges as*

$$\mathcal{O}\left(\frac{\sqrt{\Delta}\sigma}{\sqrt{nT}} + \frac{(1+\omega_\ell+\omega_u)\Delta\tilde{\Theta}(1)}{T}\right), \tag{16}$$

where T denotes the total number of iterations (i.e., #outer-loop recursion $\times$ #inner rounds in MSC) of EF-SOBA-MSC, and $\tilde{\Theta}$ hides logarithm terms independent of T.

Similarly, CM-SOBA-MSC can converge at the same rate as given in (16). For details on algorithmic development and convergence properties, please refer to Appendix C.

7. Experiments

In this section, we evaluate the performance of the proposed compressed bilevel algorithms on two problems: hyper-representation and hyperparameter optimization. To demonstrate the impact of data heterogeneity on the compared algorithms, we follow the approach in (Hsu et al., 2019) by partitioning the dataset using a Dirichlet distribution with a concentration parameter $\alpha = 0.1$. We use the unbiased compressor, scaled rand-K, to compress communication, and K is specified differently in various experiments.

Hyper-representation. Hyper-representation can be formulated as bilevel optimization in which the upper-level problem optimizes the intermediate representation parameter to obtain better feature representation on validation data, while the lower-level optimizes the weights of the downstream tasks on training data. We conduct the experiments on MNIST dataset with MLP and CIFAR-10 dataset with CNN. The detailed problem formulation and experimental setup can be found in Appendix D.1.

Figure 1 compares C-SOBA, CM-SOBA, and EF-SOBA with *non-compressed* distributed SOBA (NC-SOBA) under homogeneous data distributions. It is observed that compressed algorithms can achieve a $10\times$ reduction in communicated bits without substantial performance degradation. Figure 2 illustrates the performance under heterogeneous data distributions. It is observed that C-SOBA and CM-SOBA deteriorate in this scenario. However, error feedback significantly benefits convergence, and its convergence (in terms of iterations) and test accuracy of EF-SOBA are close to those of NC-SOBA. This is consistent with the theory that EF-SOBA is more robust to data heterogeneity. To justify its efficiency on various datasets and models, additional results of CIFAR-10 with CNN are provided in Appendix D.3. Furthermore, for more detailed discussions regarding the adjustment of the momentum parameter, please consult Appendix D.3.

Hyperparameter optimization. Hyperparameter optimization aims to enhance the performance of learning models by optimizing hyperparameters. In our experiments on the MNIST dataset using an MLP model, the left two figures in Fig. 3 depict the test accuracy performance under homogeneous data distributions. It is observed that all compressed bilevel algorithms perform on par with the non-compressed algorithm but achieve a $10\times$ reduction in communicated bits. The right two figures illustrate the test accuracy performance under heterogeneous data distributions, further corroborating the superiority of EF-SOBA in resisting data heterogeneity. The problem formulation, experimental setup, and additional numerical results can be found in Appendix D.2.

Runtime comparison. Understanding the importance of evaluating the practical trade-offs between computation and communication, we performed a runtime comparison to complement our theoretical findings with empirical evidence. The results of our hyper-representation experiment on the MNIST dataset with the MLP backbone are presented

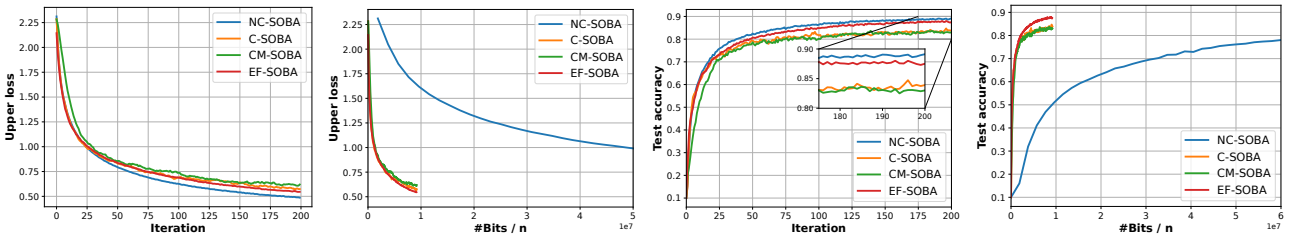

Figure 2. Hyper-representation on MNIST under heterogeneous data distributions.

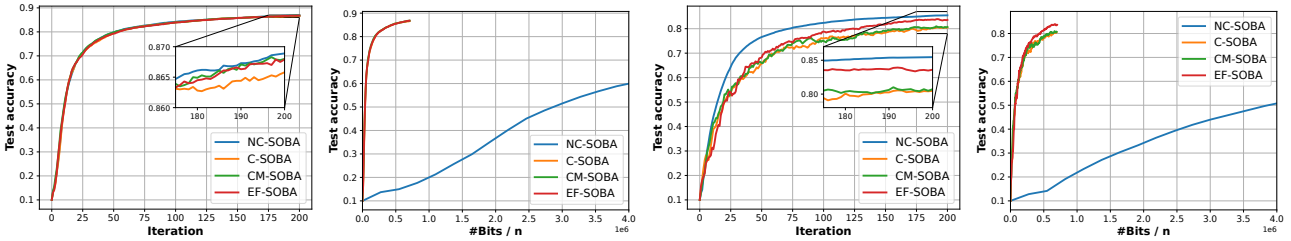

Figure 3. Hyperparameter optimization on MNIST under homogeneous (left two figures) and heterogeneous (right two figures) data distributions.

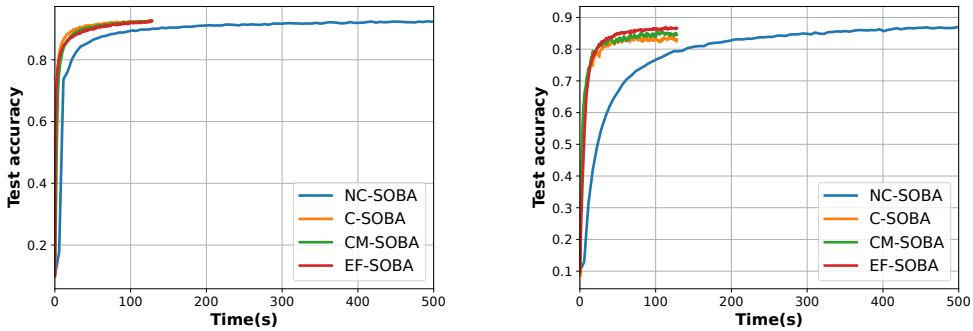

Figure 4. Running time comparison under homogeneous(left) and heterogeneous(right) data distributions.

in Fig. 4, conducted under both homogeneous and heterogeneous data distributions. It is evident that the convergence with respect to running time closely aligns with those observed for communication bits. Additionally, we observed that the computation time across all compared algorithms closely matches. This result highlights the computational efficiency of our proposed compression techniques, which introduce minimal computational overhead. Our experiments unequivocally demonstrate that communication time is the dominant factor affecting total runtime in a distributed scenario.

8. Conclusion and Limitation

This paper introduces the first set of distributed bilevel algorithms with communication compression and establishes their convergence guarantees. In experiments, these algorithms achieve a $10\times$ reduction in communication overhead.

However, our developed algorithms are only compatible with unbiased compressors, excluding biased but contractive compressors such as Top-K. In future work, we will explore bilevel algorithms with contractive compressors and investigate their convergence properties.

Impact Statement

This paper centers on the theoretical analysis of machine learning algorithm convergence. We do not foresee any significant societal consequences arising from our work, none of which we believe need to be explicitly emphasized.

Acknowlegement

The work of Kun Yuan is supported by Natural Science Foundation of China under Grants 12301392 and 92370121.

References

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, 2017.

Arora, S., Du, S., Kakade, S., Luo, Y., and Saunshi, N. Provable representation learning for imitation learning via bi-level optimization. In *International Conference on Machine Learning*, pp. 367–376. PMLR, 2020.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. Signsgd: compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, 2018.

Bertinetto, L., Henriques, J., Torr, P., and Vedaldi, A. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations (ICLR), 2019*. International Conference on Learning Representations, 2019.

Chen, T., Sun, Y., and Yin, W. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34:25294–25307, 2021.

Chen, X., Huang, M., and Ma, S. Decentralized bilevel optimization. *arXiv preprint arXiv:2206.05670*, 2022.

Chen, X., Huang, M., Ma, S., and Balasubramanian, K. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *International Conference on Machine Learning*, pp. 4641–4671. PMLR, 2023a.

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

Chilimbi, T., Suzue, Y., Apacible, J., and Kalyanaraman, K. Project adam: Building an efficient and scalable deep learning training system. In *11th USENIX symposium on operating systems design and implementation (OSDI 14)*, pp. 571–582, 2014.

Dagréou, M., Ablin, P., Vaiter, S., and Moreau, T. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *Advances in Neural Information Processing Systems*, 35:26698–26710, 2022.

Domke, J. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pp. 318–326. PMLR, 2012.

Fatkhullin, I., Tyurin, A., and Richtárik, P. Momentum provably improves error feedback! *arXiv preprint arXiv:2305.15155*, 2023.

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

Gao, H., Gu, B., and Thai, M. T. On the convergence of distributed stochastic bilevel optimization algorithms over a network. In *International Conference on Artificial Intelligence and Statistics*, pp. 9238–9281. PMLR, 2023.

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

Gibiansky, A. Bringing hpc techniques to deep learning. *Baidu Research, Tech. Rep.*, 2017.

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

He, Y., Huang, X., Chen, Y., Yin, W., and Yuan, K. Lower bounds and accelerated algorithms in distributed stochastic optimization with communication compression. *arXiv preprint arXiv:2305.07612*, 2023a.

He, Y., Huang, X., and Yuan, K. Unbiased compression saves communication in distributed optimization: When and how much? *arXiv preprint arXiv:2305.16297*, 2023b.

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.

Horvath, S., Ho, C.-Y., Horvath, L., Sahu, A. N., Canini, M., and Richtárik, P. Natural compression for distributed deep learning. *ArXiv*, abs/1905.10988, 2019.

Hsu, T. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *CoRR*, abs/1909.06335, 2019. URL http://arxiv.org/abs/1909.06335.

Huang, M., Zhang, D., and Ji, K. Achieving linear speedup in non-iid federated bilevel learning. *arXiv preprint arXiv:2302.05412*, 2023a.

Huang, X., Chen, Y., Yin, W., and Yuan, K. Lower bounds and nearly optimal algorithms in distributed learning with communication compression. *Advances in Neural Information Processing Systems*, 35:18955–18969, 2022.

Huang, X., Li, P., and Li, X. Stochastic controlled averaging for federated learning with communication compression. *arXiv preprint arXiv:2308.08165*, 2023b.

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

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

Li, M., Andersen, D. G., Park, J. W., Smola, A. J., Ahmed, A., Josifovski, V., Long, J., Shekita, E. J., and Su, B.-Y. Scaling distributed machine learning with the parameter server. In *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pp. 583–598, 2014.

Liu, H., Simonyan, K., and Yang, Y. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

Loizou, N. and Richtárik, P. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.

Lu, S., Cui, X., Squillante, M. S., Kingsbury, B., and Horesh, L. Decentralized bilevel optimization for personalized client learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5543–5547. IEEE, 2022a.

Lu, S., Zeng, S., Cui, X., Squillante, M., Horesh, L., Kingsbury, B., Liu, J., and Hong, M. A stochastic linearized augmented lagrangian method for decentralized bilevel optimization. *Advances in Neural Information Processing Systems*, 35:30638–30650, 2022b.

Maclaurin, D., Duvenaud, D., and Adams, R. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pp. 2113–2122. PMLR, 2015.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Pedregosa, F. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pp. 737–746. PMLR, 2016.

Richtárik, P., Sokolov, I., and Fatkhullin, I. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *ArXiv*, abs/2106.05203, 2021.

Safaryan, M. H., Shulgin, E., and Richtárik, P. Uncertainty principle for communication compression in distributed and federated learning and the search for an optimal compressor. *Information and Inference: A Journal of the IMA*, 2021.

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *INTERSPEECH*, 2014.

Smola, A. and Narayanamurthy, S. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710, 2010.

Stich, S. U. Local sgd converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2019.

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems*, 2018.

Strom, N. Scalable distributed dnn training using commodity gpu cloud computing. *Interspeech 2015*, 2015.

Tarzanagh, D. A., Li, M., Thrampoulidis, C., and Oymak, S. Fednest: Federated bilevel, minimax, and compositional optimization. In *International Conference on Machine Learning*, pp. 21146–21179. PMLR, 2022.

Wangni, J., Wang, J., Liu, J., and Zhang, T. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, 2018.

Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y., and Li, H. H. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, 2017.

Xiao, P. and Ji, K. Communication-efficient federated hypergradient computation via aggregated iterative differentiation. *arXiv preprint arXiv:2302.04969*, 2023.

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

Yang, S., Zhang, X., and Wang, M. Decentralized gossip-based stochastic bilevel optimization over communication networks. *Advances in Neural Information Processing Systems*, 35:238–252, 2022.

Yang, Y., Xiao, P., and Ji, K. Achieving $\mathcal{O}(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. *arXiv preprint arXiv:2312.03807*, 2023a.

Yang, Y., Xiao, P., and Ji, K. Simfbo: Towards simple, flexible and communication-efficient federated bilevel learning. *arXiv preprint arXiv:2305.19442*, 2023b.

A. Necessity of Unbiased Lower Level Gradients

In this section, we demonstrate why the lower-level variables, y and z are supposed to be updated following an unbiased estimate of their gradient direction.

First of all, it's noteworthy that the lower level in a bilevel optimization problem is not equivalent to a single level one. Recall the final goal of our bilevel algorithm is to optimize $\Phi(x)$, and y, z are actually auxiliary variables to help improve the estimate precision of the hypergradient $\nabla\Phi(x)$. Consequently, by following Lemma B.4, y^k, z^k are expected to be good estimation of $y^\star(x^k)$ and $z^\star(x^k)$, respectively. Thus, in single-loop algorithms, where lower/upper level variables are updated alternatively, the lower level cannot be regarded as a single-level optimization problem since x^k varies in each iteration.

Even if we assume the stability of upper-level solution x^k, upper-level algorithms, *e.g.*, EF21-SGDM which we use in EF-SOBA, cannot work well in the strongly-convex lower-level scenario. When we modify the non-convex objectives into a strongly-convex one, it's natural to believe that the algorithm can guarantee at least the same convergence rate as before, since the assumption has become stronger. However, it's worth noting that the convergence metric also varies in different settings. Specifically, the metric for a non-convex objective $f(x)$ is usually $\|\nabla f(\bar{x}^K)\|_2^2$ with $\bar{x}^K = \frac{1}{K+1}\sum_{k=0}^K x^k$, while that for strongly-convex $f(x)$ should be $\|x^K - x^\star\|_2^2$ or $f(x^K) - f(x^\star)$, where $x^\star$ is the minimum of $f(x)$. In the typical SOBA structure, $\|y^k - y^\star(x^k)\|_2^2$ is the actually concerned metric, which negates applying EF21-SGDM for lower-level design.

Next, we go through some technical issues to demonstrate the role of unbiased gradient estimates in this *different* optimization problem. Consider the following recursion for optimizing a strongly-convex objective $f(x)$:

$$x^{k+1} = x^k - \alpha g^k,$$

where g^k is a stochastic estimator of $\nabla f(x^k)$. When trying to establish descent inequalities with respect to $\mathbb{E}\left[\left\|x^k - x^\star\right\|_2^2\right]$, we consider

$$\mathbb{E}\left[\left\|x^{k+1} - x^\star\right\|_2^2\right] = \mathbb{E}\left[\left\|x^k - \alpha g^k - x^\star\right\|_2^2\right] = \mathbb{E}\left[\left\|x^k - x^\star\right\|_2^2\right] - 2\alpha\mathbb{E}\left[\left\langle x^k - x^\star, g^k\right\rangle\right] + \alpha^2\mathbb{E}\left[\left\|g^k\right\|_2^2\right]. \tag{17}$$

If g^k is an unbiased estimate of $\nabla f(x^k)$, $\mathbb{E}\left[\left\langle x^k - x^\star, g^k\right\rangle\right] = \mathbb{E}\left[\left\langle x^k - x^\star, \nabla f(x^k)\right\rangle\right]$ can be directly used in proving descent inequalities, leaving $\alpha^2\mathbb{E}\left[\left\|g^k\right\|_2^2\right]$ the only term including noise. Otherwise, if g^k is a biased one, an additional term $2\alpha\mathbb{E}\left[\left\langle x^k - x^\star, g^k - \nabla f(x^k)\right\rangle\right]$ which is upper bounded by $\alpha^{2-t}\mathbb{E}\left[\left\|x^k - x^\star\right\|_2^2\right] + \alpha^t\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|_2^2\right]$ will either include bigger noise (if $t < 2$) or deteriorate contraction of $\mathbb{E}\left[\left\|x^k - x^\star\right\|_2^2\right]$ (if $t \geq 2$).

Remark. Based on the above discussions, it is possible that non-convex algorithms like EF21-SGDM without unbiasedness can be applied to the lower level by using $\bar{y}^k$, $\bar{z}^k$ for hypergradient estimation. Exploration of this type of algorithms is left to future work.

B. Convergence Analysis

In this section, we provide proofs for the theoretical results in Sections 3, 4, 5 and 6. Throughout this section, we have the following notations.

Notation.

- $y_\star^k \triangleq y^\star(x^k)$, $z_\star^k \triangleq z^\star(x^k)$;
- $\mathcal{F}^k$: the σ-field of random variables already generated at the beginning of iteration k;
- $\mathbb{E}_k[\cdot] \triangleq \mathbb{E}[\cdot \mid \mathcal{F}^k]$;
- $F_i^k \triangleq f_i(x^k, y^k)$, $F^k \triangleq f(x^k, y^k)$, $F_{\star,i}^k \triangleq f(x^k, y_\star^k)$, $F_\star^k \triangleq f(x^k, y_\star^k)$, $G_i^k \triangleq g_i(x^k, y^k)$, $G^k \triangleq g(x^k, y^k)$, $G_{\star,i}^k \triangleq g_i(x^k, y_\star^k)$, $G_\star^k \triangleq g(x^k, y_\star^k)$;

- $\mathcal{X}_+^k \triangleq \mathbb{E}\left[\left\|x^{k+1}-x^k\right\|_2^2\right]$, $\mathcal{Y}_+^k \triangleq \mathbb{E}\left[\left\|y^{k+1}-y^k\right\|_2^2\right]$, $\mathcal{Z}_+^k \triangleq \mathbb{E}\left[\left\|z^{k+1}-z^k\right\|_2^2\right]$, $\mathcal{Y}^k \triangleq \mathbb{E}\left[\left\|y^k-y_\star^k\right\|_2^2\right]$, $\mathcal{Z}^k \triangleq \mathbb{E}\left[\left\|z^k-z_\star^k\right\|_2^2\right]$;
- $D_x^k \triangleq \frac{1}{n}\sum_{i=1}^n D_{x,i}^k$, $D_y^k \triangleq \frac{1}{n}\sum_{i=1}^n D_{y,i}^k$, $D_z^k \triangleq \frac{1}{n}\sum_{i=1}^n D_{z,i}^k$, $\hat{D}_x^k \triangleq \frac{1}{n}\sum_{i=1}^n \mathcal{C}_i^u(D_{x,i}^k)$;
- $\hat{D}_{y,i}^k$ and $\hat{D}_{z,i}^k$ denote compressed local update estimations, which depend on different lower-level compression mechanisms:

$$\hat{D}_{y,i}^k \triangleq \begin{cases} \mathcal{C}_i^\ell(D_{y,i}^k), & \text{in Alg. 1 and 4,} \\ m_{y,i}^k + \mathcal{C}_i^\ell(D_{y,i}^k - m_{y,i}^k), & \text{in Alg. 2 and 5,} \end{cases} \quad \hat{D}_{z,i}^k \triangleq \begin{cases} \mathcal{C}_i^\ell(D_{z,i}^k), & \text{in Alg. 1 and 4,} \\ m_{z,i}^k + \mathcal{C}_i^\ell(D_{z,i}^k - m_{z,i}^k), & \text{in Alg. 2 and 5.} \end{cases}$$

The following lemmas are frequently used in our analysis.

Lemma B.1 ((Chen et al., 2023b), Lemma B.1)**.** *Under Assumptions 2.1, 2.2, if $\beta < \frac{2}{L_g+\mu_g}$, the following inequality holds:*

$$\|y^k - \beta\nabla_y G^k - y_\star^k\|_2 \le (1-\beta\mu_g)\|y^k - y_\star^k\|_2.$$

Lemma B.2 ((Chen et al., 2023b), Lemma B.2)**.** *Under Assumptions 2.1, 2.2, there exist positive constants $L_{\nabla\Phi}$, $L_{y^\star}$, $L_{z^\star}$, such that $\nabla\Phi(x)$, $y^\star(x)$, $z^\star(x)$ are $L_{\nabla\Phi}$, $L_{y^\star}$, $L_{z^\star}$-Lipschitz continuous, respectively. Moreover, we have $\|z^\star(x)\|_2 \le \frac{C_f}{\mu_g}$ for all $x \in \mathbb{R}^{d_x}$.*

Notation. For convenience, we define the following constants:

$$L_1^2 \triangleq L_f^2 + L_{g_{yy}}^2\rho^2, \quad L_2^2 \triangleq L_f^2 + L_{g_{xy}}^2\rho^2, \quad L_3^2 \triangleq L_g^2(1+3L_{y^\star}^2), \quad L_4^2 \triangleq L_1^2(1+3L_{y^\star}^2) + 3L_g^2L_{z^\star}^2,$$
$$L_5^2 \triangleq L_2^2(1+3L_{y^\star}^2) + 3L_g^2L_{z^\star}^2, \quad \sigma_1^2 \triangleq \sigma^2(1+\rho^2), \quad \tilde{\sigma}^2 \triangleq \sigma^2/R, \quad \tilde{\sigma}_1^2 \triangleq \sigma_1^2/R, \quad \omega_1 \triangleq 1+6\omega_\ell(1+\omega_\ell),$$
$$\omega_2 \triangleq 1+36\omega_u(1+\omega_u), \quad \tilde{\omega}_\ell \triangleq \omega_\ell\left(\omega_\ell/(1+\omega_\ell)\right)^R, \quad \tilde{\omega}_u \triangleq \omega_u\left(\omega_u/(1+\omega_u)\right)^R.$$

Lemma B.3 (Variance Bounds)**.** *Under Assumption 2.3, we have the following variance bounds for Alg. 1 and 2:*

$$\begin{aligned} &\mathrm{Var}\left[D_{y,i}^k \mid \mathcal{F}^k\right] \le \sigma^2, \quad \mathrm{Var}\left[D_y^k \mid \mathcal{F}^k\right] \le \sigma^2/n; \\ &\mathrm{Var}\left[D_{x,i}^k \mid \mathcal{F}^k\right] \le \sigma_1^2, \quad \mathrm{Var}\left[D_x^k \mid \mathcal{F}^k\right] \le \sigma_1^2/n; \\ &\mathrm{Var}\left[D_{z,i}^k \mid \mathcal{F}^k\right] \le \sigma_1^2, \quad \mathrm{Var}\left[D_z^k \mid \mathcal{F}^k\right] \le \sigma_1^2/n. \end{aligned}$$

Proof. Note that $\|z^k\|_2 \le \rho$, Lemma B.3 is a direct consequence of Assumption 2.3 and the definition of σ_1. □

Lemma B.4 (Gradient Bias)**.** *Under Assumptions 2.1, 2.2, 2.3, if $\rho \ge C_f/\mu_g$, the following inequality holds for C-SOBA, CM-SOBA (Alg. 1) and EF-SOBA (Alg. 2):*

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k) - \nabla\Phi(x^k)\right\|_2^2\right] \le 3L_2^2\sum_{k=0}^K \mathcal{Y}^k + 3L_g^2\sum_{k=0}^K \mathcal{Z}^k. \tag{18}$$

Proof. By Assumption 2.3, we have $\mathbb{E}_k(D_x^k) = \nabla_{xy}^2 G^k z^k + \nabla_x F^k$. Since $\nabla\Phi(x^k) = \nabla_{xy}^2 G_\star^k z_\star^k + \nabla_x F_\star^k$, we have

$$\begin{aligned} &\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k) - \nabla\Phi(x^k)\right\|_2^2\right] \\ =&\mathbb{E}\left[\left\|(\nabla_x F^k - \nabla_x F_\star^k) + (\nabla_{xy}^2 G^k - \nabla_{xy}^2 G_\star^k)z_\star^k + \nabla_{xy}^2 G^k(z^k - z_\star^k)\right\|_2^2\right] \\ \le&3\mathbb{E}\left[\left\|\nabla_x F^k - \nabla_x F_\star^k\right\|_2^2\right] + 3\mathbb{E}\left[\left\|(\nabla_{xy}^2 G^k - \nabla_{xy}^2 G_\star^k)z_\star^k\right\|_2^2\right] + 3\mathbb{E}\left[\left\|\nabla_{xy}^2 G^k(z^k - z_\star^k)\right\|_2^2\right] \\ \le&3L_f^2\mathbb{E}\left[\left\|y^k - y_\star^k\right\|_2^2\right] + 3L_{g_{xy}}^2\rho^2\mathbb{E}\left[\left\|y^k - y_\star^k\right\|_2^2\right] + 3L_g^2\mathbb{E}\left[\left\|z^k - z_\star^k\right\|_2^2\right], \end{aligned} \tag{19}$$

where the first inequality uses Cauchy-Schwarz inequality and the second inequality uses Assumption 2.1, Lemma B.2 and $\rho \ge C_f/\mu_g$. Summing (19) from $k=0$ to $K-1$ we achieve (18). □

Lemma B.5 (Local Vanilla Compression Error in Lower Level)**.** *Under Assumptions 2.1, 2.2, 2.3, 2.4, and 4.1(or 3.1), the following inequality holds for C-SOBA and CM-SOBA (Alg. 1):*

$$\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\hat{D}_{y,i}^k-D_{y,i}^k\right\|_2^2\right]\leq 2\omega_\ell L_g^2\sum_{k=0}^{K}\mathcal{Y}^k+K\omega_\ell\sigma^2+2K\omega_\ell b_g^2, \tag{20}$$

$$\sum_{k=0}^{k-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\hat{D}_{z,i}^k-D_{z,i}^k\right\|_2^2\right]\leq 6\omega_\ell L_1^2\sum_{k=0}^{K}\mathcal{Y}^k+6\omega_\ell L_g^2\sum_{k=0}^{K}\mathcal{Z}^k+K\omega_\ell\sigma_1^2+4K\omega_\ell b_f^2+\frac{4K\omega_\ell C_f^2}{\mu_g^2}b_g^2. \tag{21}$$

Proof. By definition, we have $\hat{D}_{y,i}^k=\mathcal{C}_i^\ell(D_{y,i}^k)$, thus

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{D}_{y,i}^k-D_{y,i}^k\right\|_2^2\right]\leq&\omega_\ell\mathbb{E}\left[\left\|D_{y,i}^k\right\|_2^2\right]\leq\omega_\ell\mathbb{E}\left[\left\|\nabla_y G_i^k\right\|_2^2\right]+\omega_\ell\sigma^2\\ \leq&2\omega_\ell\mathbb{E}\left[\left\|\nabla_y G_{\star,i}^k\right\|_2^2\right]+2\omega_\ell L_g^2\mathbb{E}\left[\left\|y^k-y_\star^k\right\|_2^2\right]+\omega_\ell\sigma^2,\end{aligned} \tag{22}$$

where the first inequality uses Assumption 2.4, the second inequality uses Lemma B.3, the third inequality uses $\nabla_y G_i^k=\nabla_y G_{\star,i}^k+(\nabla_y G_i^k-\nabla_y G_{\star,i}^k)$, Cauchy-Schwarz inequality and Assumption 2.1. Averaging (22) from $i=1$ to n, we obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\hat{D}_{y,i}^k-D_{y,i}^k\right\|_2^2\right]\leq&\frac{2\omega_\ell}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla_y G_{\star,i}^k-\nabla_y G_\star^k\right\|_2^2\right]+2\omega_\ell L_g^2\mathbb{E}\left[\left\|y^k-y_\star^k\right\|_2^2\right]+\omega_\ell\sigma^2\\ \leq&2\omega_\ell b_g^2+2\omega_\ell L_g^2\mathbb{E}\left[\left\|y^k-y_\star^k\right\|_2^2\right]+\omega_\ell\sigma^2,\end{aligned} \tag{23}$$

where the first inequality uses $\nabla_y G_\star^k=0$, and the second inequality uses Assumption 4.1 (or Assumption 3.1). Summing (23) from $k=0$ to $K-1$ we obtain (20). Similarly, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{D}_{z,i}^k-D_{z,i}^k\right\|_2^2\right]\\ \leq&\omega_\ell\mathbb{E}\left[\left\|\nabla_{yy}^2 G_i^k z^k+\nabla_y F_i^k\right\|_2^2\right]+\omega_\ell\sigma_1^2\\ \leq&6\omega_\ell\mathbb{E}\left[\left\|(\nabla_{yy}^2 G_i^k-\nabla_{yy}^2 G_{\star,i}^k)z^k\right\|_2^2\right]+6\omega_\ell\mathbb{E}\left[\left\|\nabla_{yy}^2 G_{\star,i}^k(z^k-z_\star^k)\right\|_2^2\right]+6\omega_\ell\mathbb{E}\left[\left\|\nabla_y F_i^k-\nabla_y F_{\star,i}^k\right\|_2^2\right]\\ &+4\omega_\ell\mathbb{E}\left[\left\|(\nabla_{yy}^2 G_{\star,i}^k-\nabla_{yy}^2 G_\star^k)z_\star^k\right\|_2^2\right]+4\omega_\ell\mathbb{E}\left[\left\|\nabla_y F_{\star,i}^k-\nabla_y F_\star^k\right\|_2^2\right]+\omega_\ell\sigma_1^2\\ \leq&6\omega_\ell\left(L_f^2+L_{g_{yy}}^2\rho^2\right)\mathbb{E}\left[\left\|y^k-y_\star^k\right\|_2^2\right]+6\omega_\ell L_g^2\mathbb{E}\left[\left\|z^k-z_\star^k\right\|_2^2\right]+\frac{4\omega_\ell C_f^2}{\mu_g^2}\mathbb{E}\left[\left\|\nabla_{yy}^2 G_{\star,i}^k-\nabla_{yy}^2 G_\star^k\right\|_2^2\right]\\ &+4\omega_\ell\mathbb{E}\left[\left\|\nabla_y F_{\star,i}^k-\nabla_y F_\star^k\right\|_2^2\right]+\omega_\ell\sigma_1^2,\end{aligned} \tag{24}$$

where the first inequality uses Assumption 2.4 and Lemma B.3, the second inequality uses $\nabla_{yy}^2 G_\star^k z_\star^k+\nabla_y F_\star^k=0$ and Cauchy-Schwarz inequality, the third inequality uses Assumption 2.1 and Lemma B.2. Averaging (24) from $i=1$ to n, summing from $k=0$ to $K-1$, we achieve (21). □

Lemma B.6 (Global Vanilla Compression Error in Lower Level)**.** *Under Assumptions 2.1, 2.2, 2.3, 2.4 and 4.1(or 3.1), the following inequalities hold for C-SOBA and CM-SOBA (Alg. 1):*

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla_y G^k-\hat{D}_y^k\right\|_2^2\right]\leq\frac{2\omega_\ell L_g^2}{n}\sum_{k=0}^{K}\mathcal{Y}^k+\frac{(1+\omega_\ell)K\sigma^2}{n}+\frac{2\omega_\ell K b_g^2}{n}, \tag{25}$$

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k)-\hat{D}_z^k\right\|_2^2\right]\leq\frac{6\omega_\ell L_1^2}{n}\sum_{k=0}^{K}\mathcal{Y}^k+\frac{6\omega_\ell L_g^2}{n}\sum_{k=0}^{K}\mathcal{Z}^k+\frac{(1+\omega_\ell)K\sigma_1^2}{n}+\frac{4\omega_\ell K b_f^2}{n}+\frac{4\omega_\ell C_f^2 K b_g^2}{n\mu_g^2}. \tag{26}$$

Proof. By Assumption 2.3, we have

$$\mathbb{E}\left[\left\|\nabla_y G^k-\hat{D}_y^k\right\|_2^2\right]=\mathbb{E}\left[\left\|(\hat{D}_y^k-D_y^k)+(D_y^k-\nabla_y G^k)\right\|_2^2\right]=\mathbb{E}\left[\left\|\hat{D}_y^k-D_y^k\right\|_2^2\right]+\mathbb{E}\left[\left\|D_y^k-\nabla_y G^k\right\|_2^2\right].$$

Thus, by Assumption 2.4 and conditional independence of the compressors, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla_y G^k-\hat{D}_y^k\right\|_2^2\right]=&\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n(\hat{D}_{y,i}^k-D_{y,i}^k)\right\|_2^2\right]+\mathbb{E}\left[\left\|D_y^k-\nabla_y G^k\right\|_2^2\right]\\=&\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left[\left\|\hat{D}_{y,i}^k-D_{y,i}^k\right\|_2^2\right]+\mathbb{E}\left[\left\|D_y^k-\nabla_y G^k\right\|_2^2\right]\\\leq&\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left[\left\|\hat{D}_{y,i}^k-D_{y,i}^k\right\|_2^2\right]+\frac{1}{n}\sigma^2,\end{aligned}\tag{27}$$

where the inequality uses Lemma B.3. Summing (27) from $k=0$ to $K-1$ and applying (20), we obtain (25). Similarly, we have

$$\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k)-\hat{D}_z^k\right\|_2^2\right]\leq\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left[\left\|\hat{D}_{z,i}^k-D_{z,i}^k\right\|_2^2\right]+\frac{1}{n}\sigma_1^2.\tag{28}$$

Summing (28) from $k=0$ to $K-1$ and applying (21), we obtain (26). □

B.1. Proof of Theorem 3.4

Before proving Theorem 3.4, we need a few additional lemmas.

Lemma B.7 ((Dagréou et al., 2022), Lemma C.1 and C.2)**.** *Under Assumptions 2.1, 2.2, 3.3, there exists positive constants L_{yx} and L_{zx}, such that $y^\star(x)$ and $z^\star(x)$ are L_{yx} and L_{zx}-smooth, respectively.*

Lemma B.8 (Bounded Second Moment of Vanilla Compressed Update Directions)**.** *Under Assumptions 2.1, 2.2, 2.3, 2.4 and 3.1, the following inequalities hold for C-SOBA (Alg. 1):*

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right]\leq\left(1+\frac{\omega_u}{n}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right]+\frac{(1+\omega_u)K\sigma_1^2}{n}+\frac{2\omega_u K b_f^2}{n}+\frac{2\rho^2\omega_u K b_g^2}{n},\tag{29}$$

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\hat{D}_y^k\right\|_2^2\right]\leq\left(1+\frac{2\omega_\ell}{n}\right)L_g^2\sum_{k=0}^{K}\mathcal{Y}^k+\frac{(1+\omega_\ell)K\sigma^2}{n}+\frac{2\omega_\ell K b_g^2}{n},\tag{30}$$

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\hat{D}_z^k\right\|_2^2\right]\leq\left(3+\frac{6\omega_\ell}{n}\right)L_1^2\sum_{k=0}^{K}\mathcal{Y}^k+\left(3+\frac{6\omega_\ell}{n}\right)L_g^2\sum_{k=0}^{K}\mathcal{Z}^k+\frac{(1+\omega_\ell)K\sigma_1^2}{n}+\frac{4\omega_1 K b_f^2}{n}+\frac{4C_f^2\omega_\ell K b_g^2}{n\mu_g^2}.\tag{31}$$

Proof. Let $\mathcal{F}_x^k$ denote the σ-field of $\{D_{x,i}^k\mid 1\leq i\leq n\}$ and random variables already generated at the beginning of iteration k. By Assumption 2.3 and Lemma B.3, we have

$$\mathbb{E}\left[\left\|D_x^k\right\|_2^2\right]=\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right]+\mathrm{Var}\left[D_x^k\mid\mathcal{F}^k\right]\leq\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right]+\frac{\sigma_1^2}{n}.\tag{32}$$

Similarly,

$$\mathbb{E}\left[\left\|D_{x,i}^k\right\|_2^2\right]\leq\mathbb{E}\left[\left\|\mathbb{E}_k(D_{x,i}^k)\right\|_2^2\right]+\sigma_1^2.\tag{33}$$

Averaging (33) from $i=1$ to n, we obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|D_{x,i}^k\right\|_2^2\right]\leq&\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\mathbb{E}_k(D_{x,i}^k)\right\|_2^2\right]+\sigma_1^2\\=&\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right]+\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\mathbb{E}_k(D_{x,i}^k)-\mathbb{E}_k(D_x^k)\right\|_2^2\right]+\sigma_1^2\\\leq&\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right]+\frac{2}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|(\nabla_{xy}^2G_i^k-\nabla_{xy}^2G^k)z^k\right\|_2^2\right]+\frac{2}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\nabla_x F_i^k-\nabla_x F^k\right\|_2^2\right]+\sigma_1^2\end{aligned}$$

$$\leq \mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right] + 2\rho^2 b_g^2 + 2b_f^2 + \sigma_1^2, \tag{34}$$

where the second inequality uses Cauchy-Schwarz inequality, and the third inequality uses Assumption 3.1. For $\hat{D}_x^k$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right] =& \mathbb{E}\left[\mathbb{E}\left[\|\hat{D}_x^k\|^2 \middle| \mathcal{F}_x^k\right]\right] \\
=& \mathbb{E}\left[\left\|D_x^k\right\|_2^2\right] + \mathbb{E}\left[\mathbb{E}\left[\left\|\hat{D}_x^k - D_x^k\right\|^2 \middle| \mathcal{F}_x^k\right]\right] \\
=& \mathbb{E}\left[\left\|D_x^k\right\|_2^2\right] + \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\mathbb{E}\left[\left\|\hat{D}_{x,i}^k - D_{x,i}^k\right\|^2 \middle| \mathcal{F}_x^k\right]\right] \\
\leq& \mathbb{E}\left[\left\|D_x^k\right\|_2^2\right] + \frac{\omega_u}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|D_{x,i}^k\right\|_2^2\right],
\end{aligned} \tag{35}$$

where the inequality uses Assumption 2.4. Applying (32)(34) to (35) and summing from $k=0$ to $K-1$, we obtain (29). By Assumption 2.1, we have

$$\mathbb{E}\left[\left\|\mathbb{E}_k(D_y^k)\right\|_2^2\right] = \mathbb{E}\left[\left\|\nabla_y G^k - \nabla_y G_\star^k\right\|_2^2\right] \leq L_g^2 \mathbb{E}\left[\left\|y^k - y_\star^k\right\|_2^2\right].$$

Consequently,

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{D}_y^k\right\|_2^2\right] =& \mathbb{E}\left[\left\|\mathbb{E}_k(D_y^k)\right\|_2^2\right] + \mathbb{E}\left[\left\|\hat{D}_y^k - \mathbb{E}_k(D_y^k)\right\|_2^2\right] \\
\leq& L_g^2 \mathbb{E}\left[\left\|y^k - y_\star^k\right\|_2^2\right] + \mathbb{E}\left[\left\|\hat{D}_y^k - \nabla_y G^k\right\|_2^2\right].
\end{aligned} \tag{36}$$

Summing (36) from $k=0$ to $K-1$ and applying Lemma B.6, we obtain (30). Similarly, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k)\right\|_2^2\right] =& \mathbb{E}\left[\left\|(\nabla_{yy}^2 G^k - \nabla_{yy}^2 G_\star^k) z^k + \nabla_{yy}^2 G_\star^k (z^k - z_\star^k) + (\nabla_y F^k - \nabla_y F_\star^k)\right\|_2^2\right] \\
\leq& 3(L_f^2 + L_{g_{yy}}^2 \rho^2) \mathbb{E}\left[\left\|y^k - y_\star^k\right\|_2^2\right] + 3L_g^2 \mathbb{E}\left[\left\|z^k - z_\star^k\right\|_2^2\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{D}_z^k\right\|_2^2\right] =& \mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k)\right\|_2^2\right] + \mathbb{E}\left[\left\|\hat{D}_z^k - \mathbb{E}_k(D_z^k)\right\|_2^2\right] \\
\leq& 3L_1^2 \mathbb{E}\left[\left\|y^k - y_\star^k\right\|_2^2\right] + 3L_g^2 \mathbb{E}\left[\left\|z^k - z_\star^k\right\|_2^2\right].
\end{aligned} \tag{37}$$

Summing (37) from $k=0$ to $K-1$ and applying Lemma B.6, we obtain (31). □

Lemma B.9 (Lower Level Convergence)**.** *Under Assumptions 2.1, 2.2, 2.3, 2.4, 3.1, 3.2, 3.3 and assume* $\beta < \min\left\{\frac{2}{\mu_g + L_g}, \frac{\mu_g}{8(1+4\omega_\ell/n)L_g^2}\right\}$, $\gamma \leq \min\left\{\frac{1}{L_g}, \frac{\mu_g}{12(1+4\omega_\ell/n)L_g^2}\right\}$, $\rho \geq \frac{C_f}{\mu_g}$, *and*

$$\alpha \leq \sqrt{\frac{\mu_g L_{y^\star}^2 \beta}{L_{yx}^2\left(B_x^2(1+\omega_u/n) + (1+\omega_u)\sigma_1^2/n + 2\omega_u b_f^2/n + 2\rho^2\omega_u b_g^2/n\right)}},$$

the following inequalities hold for C-SOBA (Alg. 1):

$$\begin{aligned}
\sum_{k=0}^K \mathcal{Y}^k \leq& \frac{8\mathcal{Y}^0}{\mu_g\beta} + \frac{16(1+\omega_\ell)K\beta\sigma^2}{\mu_g n} + \frac{24L_{y^\star}^2(1+\omega_u)K\alpha^2\sigma_1^2}{\mu_g n\beta} \\
&+ \left(\frac{24L_{y^\star}^2\alpha^2(1+\omega_u/n)}{\mu_g\beta} + \frac{16L_{y^\star}^2\alpha^2}{\mu_g^2\beta^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right]
\end{aligned}$$

$$+\frac{48L_{y^\star}^2\omega_u K\alpha^2 b_f^2}{\mu_g n\beta}+\left(\frac{32\omega_\ell K\beta}{n\mu_g}+\frac{48L_{y^\star}^2\rho^2\omega_u K\alpha^2}{n\mu_g\beta}\right)b_g^2, \tag{38}$$

$$\begin{aligned}\sum_{k=0}^{K}\mathcal{Z}^k \leq&\frac{4\mathcal{Z}^0}{\mu_g\gamma}+\frac{136L_1^2\mathcal{Y}^0}{\mu_g^3\beta}+\frac{272(1+\omega_\ell)L_1^2K\beta\sigma^2}{\mu_g^3 n}+\\&+\left(\frac{408L_1^2L_{y^\star}^2(1+\omega_u)\alpha^2}{\mu_g^3 n\beta}+\frac{8(1+\omega_\ell)\gamma}{\mu_g n}+\frac{8(L_{z^\star}^2+L_{zx}\rho)(1+\omega_u)\alpha^2}{\mu_g n\gamma}\right)\cdot K\sigma_1^2\\&+\left(\frac{408L_1^2L_{y^\star}^2(1+\omega_u/n)\alpha^2}{\mu_g^3\beta}+\frac{272L_1^2L_{y^\star}^2\alpha^2}{\mu_g^4\beta^2}+\frac{8(L_{z^\star}^2+L_{zx}\rho)(1+\omega_u/n)\alpha^2}{\mu_g\gamma}\right.\\&\left.+\frac{8L_{z^\star}^2\alpha^2}{\mu_g^2\gamma^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right]+\left(\frac{816L_1^2L_{y^\star}^2\omega_u K\alpha^2}{\mu_g^3 n\beta}+\frac{32\omega_\ell K\gamma}{n\mu_g}+\frac{16(L_{z^\star}^2+L_{zx}\rho)\omega_u K\alpha^2}{n\mu_g\gamma}\right)b_f^2\\&+\left(\frac{544L_1^2\omega_\ell K\beta}{n\mu_g^3}+\frac{816L_1^2L_{y^\star}^2\rho^2\omega_u K\alpha^2}{n\mu_g^3\beta}+\frac{16\rho^2\omega_\ell K\gamma}{n\mu_g}+\frac{8(L_{z^\star}^2+L_{zx}\rho)\rho^2\omega_u K\alpha^2}{n\mu_g\gamma}\right)b_g^2.\end{aligned} \tag{39}$$

Proof. We first separate $\mathcal{Y}^{k+1}$ into five parts:

$$\begin{aligned}\mathcal{Y}^{k+1}=&\mathbb{E}\left[\left\|y^{k+1}-y_\star^k\right\|_2^2\right]+\mathbb{E}\left[\left\|y_\star^{k+1}-y_\star^k\right\|_2^2\right]-2\mathbb{E}\left[\langle y^{k+1}-y_\star^k,y_\star^{k+1}-y_\star^k\rangle\right]\\=&\mathbb{E}\left[\left\|y^{k+1}-y_\star^k\right\|_2^2\right]+\mathbb{E}\left[\left\|y_\star^{k+1}-y_\star^k\right\|_2^2\right]-2\mathbb{E}\left[\langle y^k-y_\star^k,y_\star^{k+1}-y_\star^k\rangle\right]+2\beta\mathbb{E}\left[\left\langle\hat{D}_y^k,y_\star^{k+1}-y_\star^k\right\rangle\right]\\=&\mathbb{E}\left[\left\|y^{k+1}-y_\star^k\right\|_2^2\right]+\mathbb{E}\left[\left\|y_\star^{k+1}-y_\star^k\right\|_2^2\right]-2\mathbb{E}\left[\langle y^k-y_\star^k,\nabla y^\star(x^k)(x^{k+1}-x^k)\rangle\right]\\&-2\mathbb{E}\left[\langle y^k-y_\star^k,y_\star^{k+1}-y_\star^k-\nabla y^\star(x^k)(x^{k+1}-x^k)\rangle\right]+2\beta\mathbb{E}\left[\left\langle\hat{D}_y^k,y_\star^{k+1}-y_\star^k\right\rangle\right],\end{aligned} \tag{40}$$

where the existence of $\nabla y^\star(x^k)$ is guaranteed by Lemma B.7.

For the first part, by Assumptions 2.3 and 2.4, $\hat{D}_y^k$ is an unbiased estimator of $\nabla_y G^k$, thus

$$\begin{aligned}\mathbb{E}\left[\left\|y^{k+1}-y_\star^k\right\|_2^2\right]=&\mathbb{E}\left[\left\|y^k-\beta\hat{D}_y^k-y_\star^k\right\|_2^2\right]=\mathbb{E}\left[\left\|(y^k-y_\star^k-\beta\nabla_y G^k)-\beta(\hat{D}_y^k-\nabla_y G^k)\right\|_2^2\right]\\=&\mathbb{E}\left[\left\|y^k-y_\star^k-\beta\nabla_y G^k\right\|_2^2\right]+\mathbb{E}\left[\left\|\beta(\hat{D}_y^k-\nabla_y G^k)\right\|_2^2\right]\\\leq&(1-\beta\mu_g)^2\mathcal{Y}^k+\beta^2\mathbb{E}\left[\left\|\hat{D}_y^k-\nabla_y G^k\right\|_2^2\right]\\\leq&(1-\beta\mu_g)\mathcal{Y}^k+\beta^2\mathbb{E}\left[\left\|\hat{D}_y^k-\nabla_y G^k\right\|_2^2\right],\end{aligned} \tag{41}$$

where the first inequality uses Lemma B.1. For the second part, Lemma B.2 implies

$$\mathbb{E}\left[\left\|y_\star^{k+1}-y_\star^k\right\|_2^2\right]\leq L_{y^\star}^2\mathcal{X}_+^k=L_{y^\star}^2\alpha^2\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right]. \tag{42}$$

For the third part, we have

$$\begin{aligned}-2\mathbb{E}\left[\langle y^k-y_\star^k,\nabla y^\star(x^k)(x^{k+1}-x^k)\rangle\right]=&2\alpha\mathbb{E}\left[\langle y^k-y_\star^k,\nabla y^\star(x^k)\mathbb{E}_k(D_x^k)\rangle\right]\\\leq&\frac{\beta\mu_g}{2}\mathcal{Y}^k+\frac{2\alpha^2}{\beta\mu_g}L_{y^\star}^2\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right],\end{aligned} \tag{43}$$

where the equality uses the unbiasedness of $\hat{D}_x^k$ and the inequality uses Young's inequality and Lemma B.2. For the fourth part, we have

$$-2\mathbb{E}\left[\langle y^k-y_\star^k,y_\star^{k+1}-y_\star^k-\nabla y^\star(x^k)(x^{k+1}-x^k)\rangle\right]\leq L_{yx}\mathbb{E}\left[\|y^k-y_\star^k\|\|x^{k+1}-x^k\|^2\right]$$

$$
\begin{aligned}
\leq & \frac{L_{yx}^2}{4L_{y^\star}^2}\mathbb{E}\left[\mathbb{E}_k\left[\|y^k-y_\star^k\|^2\|x^{k+1}-x^k\|^2\right]\right]+L_{y^\star}^2\mathcal{X}_+^k \\
\leq & \frac{L_{yx}^2\alpha^2}{4L_{y^\star}^2}\mathbb{E}\left[\|y^k-y_\star^k\|^2\mathbb{E}_k\left[\left\|\hat{D}_x^k\right\|^2\right]\right]+L_{y^\star}^2\alpha^2\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right] \\
\leq & \frac{L_{yx}^2\left(B_x^2(1+\omega_u/n)+(1+\omega_u)\sigma_1^2/n+2\omega_u b_f^2/n+2\rho^2\omega_u b_g^2/n\right)\alpha^2}{4L_{y^\star}^2}\mathcal{Y}^k+L_{y^\star}^2\alpha^2\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right] \\
\leq & \frac{\beta\mu_g}{4}\mathcal{Y}^k+L_{y^\star}^2\alpha^2\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right],
\end{aligned}
\tag{44}
$$

where the first inequality uses Taylor's expansion, Cauchy's inequality and Lemma B.7, the second inequality uses Young's inequality, the third inequality uses the definition of x^{k+1}, the fourth inequality is due to

$$
\begin{aligned}
\mathbb{E}_k\left[\left\|\hat{D}_x^k\right\|^2\right] = & \left\|\mathbb{E}_k(D_x^k)\right\|^2+\mathbb{E}_k\left[\left\|D_x^k-\mathbb{E}_k(D_x^k)\right\|^2\right]+\mathbb{E}_k\left[\mathbb{E}\left[\left\|\hat{D}_x^k-D_x^k\right\|^2\middle|\mathcal{F}_x^k\right]\right] \\
\leq & B_x^2+\frac{\sigma_1^2}{n}+\frac{\omega_u}{n^2}\sum_{i=1}^n\mathbb{E}_k\left[\left\|D_{x,i}^k\right\|^2\right] \\
\leq & B_x^2+\frac{\sigma_1^2}{n}+\frac{\omega_u}{n}\left(B_x^2+\sigma_1^2+2b_f^2+2\rho^2b_g^2\right),
\end{aligned}
$$

and the last inequality is due to $\alpha\leq\sqrt{\frac{\mu_g L_{y^\star}^2\beta}{L_{yx}^2\left(B_x^2(1+\omega_u/n)+(1+\omega_u)\sigma_1^2/n+2\omega_u b_f^2/n+2\rho^2\omega_u b_g^2/n\right)}}$. For the fifth part, by Young's inequality and Lemma B.2 we have

$$
2\beta\mathbb{E}\left[\left\langle\hat{D}_y^k,y_\star^{k+1}-y_\star^k\right\rangle\right]\leq\beta^2\mathbb{E}\left[\left\|\hat{D}_y^k\right\|_2^2\right]+L_{y^\star}^2\alpha^2\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right].
\tag{45}
$$

Summing (40)(41)(42)(43)(44)(45) from $k=0$ to $K-1$ and applying Lemma B.6 we obtain

$$
\begin{aligned}
& \sum_{k=0}^K\mathcal{Y}^k \\
\leq & \frac{4\mathcal{Y}^0}{\beta\mu_g}+\frac{4(1+\omega_\ell)K\beta\sigma^2}{n\mu_g}+\frac{12L_{y^\star}^2\alpha^2}{\beta\mu_g}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right]+\frac{8L_{y^\star}^2\alpha^2}{\beta^2\mu_g^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right] \\
& +\frac{8\omega_\ell L_g^2\beta}{n\mu_g}\sum_{k=0}^K\mathcal{Y}^k+\frac{4\beta}{\mu_g}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\hat{D}_y^k\right\|_2^2\right]+\frac{8\omega_\ell K\beta b_g^2}{n\mu_g} \\
\leq & \frac{4\mathcal{Y}^0}{\beta\mu_g}+\frac{8(1+\omega_\ell)K\beta\sigma^2}{n\mu_g}+\frac{12L_{y^\star}^2(1+\omega_u)K\alpha^2\sigma_1^2}{\mu_g n\beta}+\left(\frac{12L_{y^\star}^2\alpha^2(1+\omega_u/n)}{\beta\mu_g}+\frac{8L_{y^\star}^2\alpha^2}{\beta^2\mu_g^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right] \\
& +\frac{4(1+4\omega_\ell/n)L_g^2\beta}{\mu_g}\sum_{k=0}^K\mathcal{Y}^k+\frac{24L_{y^\star}^2\omega_u K\alpha^2b_f^2}{\mu_g n\beta}+\left(\frac{16\omega_\ell K\beta}{n\mu_g}+\frac{24L_{y^\star}^2\rho^2\omega_u K\alpha^2}{n\mu_g\beta}\right)b_g
\end{aligned}
\tag{46}
$$

where the second inequality uses Lemma B.8. Using $\beta\leq\frac{\mu_g}{8(1+4\omega_\ell/n)L_g^2}$, (46) implies (38). Similarly, we separate $\mathcal{Z}^{k+1}$ into five parts:

$$
\begin{aligned}
\mathcal{Z}^{k+1}\leq & \mathbb{E}\left[\left\|\tilde{z}^{k+1}-z_\star^{k+1}\right\|_2^2\right] \\
= & \mathbb{E}\left[\left\|\tilde{z}^{k+1}-z_\star^k\right\|_2^2\right]+\mathbb{E}\left[\left\|z_\star^{k+1}-z_\star^k\right\|_2^2\right]-2\mathbb{E}\left[\left\langle z^k-z_\star^k,\nabla z^\star(x^k)(x^{k+1}-x^k)\right\rangle\right] \\
& -2\mathbb{E}\left[\left\langle z^k-z_\star^k,z_\star^{k+1}-z_\star^k-\nabla z^\star(x^k)(x^{k+1}-x^k)\right\rangle\right]+2\gamma\mathbb{E}\left[\left\langle\hat{D}_z^k,z_\star^{k+1}-z_\star^k\right\rangle\right],
\end{aligned}
\tag{47}
$$

where the inequality is due to Lemma B.2 and $\rho \geq C_f/\mu_g$. For the first part, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{z}^{k+1}-z_\star^k\right\|_2^2\right] =&\mathbb{E}\left[\left\|z^k-\gamma\mathbb{E}_k(D_z^k)-z_\star^k\right\|_2^2\right]+\gamma^2\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k)-\hat{D}_z^k\right\|_2^2\right]\\ =&\mathbb{E}\left[\left\|(z^k-z_\star^k)-\gamma\nabla_{yy}^2G^k(z^k-z_\star^k)-\gamma(\nabla_{yy}^2G^kz_\star^k+\nabla_yF^k)\right\|_2^2\right]+\gamma^2\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k)-\hat{D}_z^k\right\|_2^2\right]\\ \leq&(1+\gamma\mu_g)(1-\gamma\mu_g)^2\mathcal{Z}^k+\left(1+\frac{1}{\gamma\mu_g}\right)\gamma^2\cdot 2L_1^2\mathcal{Y}^k+\gamma^2\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k)-\hat{D}_z^k\right\|_2^2\right],\end{aligned}\tag{48}$$

where the inequality uses Young's inequality, $\|I-\gamma\nabla_{yy}^2G^k\|_2\leq 1-\gamma\mu_g$ and

$$\mathbb{E}\left[\left\|(\nabla_{yy}^2G^k-\nabla_{yy}^2G_\star^k)z_\star^k+(\nabla_yF^k-\nabla_yF_\star^k)\right\|_2^2\right]\leq 2\left(L_f^2+L_{g_{yy}}^2\rho^2\right)\mathcal{Y}^k.$$

For the second part, Lemma B.2 implies

$$\mathbb{E}\left[\left\|z_\star^{k+1}-z_\star^k\right\|_2^2\right]\leq L_{z^\star}^2\mathcal{X}_+^k=L_{z^\star}^2\alpha^2\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right].\tag{49}$$

For the third part, applying Young's inequality along with Lemma B.2 gives

$$-2\mathbb{E}\left[\left\langle z^k-z_\star^k,\nabla z^\star(x^k)(x^{k+1}-x^k)\right\rangle\right]\leq\frac{\gamma\mu_g}{2}\mathcal{Z}^k+\frac{2\alpha^2}{\gamma\mu_g}L_{z^\star}^2\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right].\tag{50}$$

For the fourth part, we have

$$-2\mathbb{E}\left[\left\langle z^k-z_\star^k,z_\star^{k+1}-z_\star^k-\nabla z^\star(x^k)(x^{k+1}-x^k)\right\rangle\right]\leq L_{zx}\mathbb{E}\left[\|z^k-z_\star^k\|\|x^{k+1}-x^k\|^2\right]\leq 2L_{zx}\rho\alpha^2\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right],\tag{51}$$

where the first inequality uses Taylor's expansion, Cauchy-Schwarz inequality and Lemma B.7, and the second inequality uses Lemma B.2 and $\rho\geq C_f/\mu_g$. For the fifth part, by Young's inequality and Lemma B.2 we have

$$2\gamma\mathbb{E}\left[\left\langle\hat{D}_z^k,z_\star^{k+1}-z_\star^k\right\rangle\right]\leq\gamma^2\mathbb{E}\left[\left\|\hat{D}_z^k\right\|_2^2\right]+L_{z^\star}^2\alpha^2\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right].\tag{52}$$

Summing (47)(48)(49)(50)(51)(52) from $K=0$ to $K-1$ and applying Lemma B.6 we achieve

$$\begin{aligned}\sum_{k=0}^K\mathcal{Z}^k\leq&\frac{2\mathcal{Z}^0}{\mu_g\gamma}+\frac{2(1+\omega_\ell)K\gamma\sigma_1^2}{\mu_gn}+\left(\frac{4L_{z^\star}^2\alpha^2}{\mu_g\gamma}+\frac{4L_{zx}\rho\alpha^2}{\mu_g\gamma}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right]+\frac{4L_{z^\star}^2\alpha^2}{\mu_g^2\gamma^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right]\\ &+\frac{2\gamma}{\mu_g}\mathbb{E}\left[\left\|\hat{D}_z^k\right\|_2^2\right]+\left(\frac{8L_1^2}{\mu_g^2}+\frac{12\omega_\ell L_1^2\gamma}{\mu_gn}\right)\sum_{k=0}^K\mathcal{Y}^k+\frac{12\omega_\ell L_g^2\gamma}{\mu_gn}\sum_{k=0}^K\mathcal{Z}^k+\frac{8\omega_\ell K\gamma b_f^2}{n\mu_g}+\frac{8C_f^2\omega_\ell K\gamma b_g^2}{n\mu_g^3}\\ \leq&\frac{2\mathcal{Z}^0}{\mu_g\gamma}+\left(\frac{4(1+\omega_\ell)K\gamma\sigma_1^2}{\mu_gn}+\frac{4(L_{z^\star}^2+L_{zx}\rho)(1+\omega_u)K\alpha^2\sigma_1^2}{\mu_gn\gamma}\right)\\ &+\left(\frac{4L_{z^\star}^2\alpha^2}{\mu_g^2\gamma^2}+\frac{4(L_{z^\star}^2+L_{zx}\rho)(1+\omega_u/n)\alpha^2}{\mu_g\gamma}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right]\\ &+\left(\frac{8L_1^2}{\mu_g^2}+\frac{6(1+4\omega_\ell/n)L_1^2\gamma}{\mu_g}\right)\sum_{k=0}^K\mathcal{Y}^k+\frac{6(1+4\omega_\ell/n)L_g^2\gamma}{\mu_g}\sum_{k=0}^K\mathcal{Z}^k\\ &+\left(\frac{16\omega_\ell K\gamma}{n\mu_g}+\frac{8(L_{z^\star}^2+L_{zx}\rho)\omega_uK\alpha^2}{n\mu_g\gamma}\right)b_f^2+\left(\frac{16\rho^2\omega_\ell K\gamma}{n\mu_g}+\frac{8(L_{z^\star}^2+L_{zx}\rho)\rho^2\omega_uK\alpha^2}{n\mu_g\gamma}\right)b_g^2,\end{aligned}\tag{53}$$

where the second inequality uses Lemma B.8 and $\rho\geq C_f/\mu_g$. Using $\gamma\leq\frac{\mu_g}{12(1+4\omega_\ell/n)L_g^2}$, we obtain

$$\sum_{k=0}^K\mathcal{Z}^k\leq\frac{4\mathcal{Z}^0}{\mu_g\gamma}+\left(\frac{8(1+\omega_\ell)K\gamma\sigma_1^2}{\mu_gn}+\frac{8(L_{z^\star}^2+L_{zx}\rho)(1+\omega_u)K\alpha^2\sigma_1^2}{\mu_gn\gamma}\right)$$

$$+\left(\frac{8L_{z^\star}^2\alpha^2}{\mu_g^2\gamma^2}+\frac{8(L_{z^\star}^2+L_{zx}\rho)(1+\omega_u/n)\alpha^2}{\mu_g\gamma}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right]$$
$$+\frac{17L_1^2}{\mu_g^2}\sum_{k=0}^{K}\mathcal{Y}^k+\left(\frac{32\omega_\ell K\gamma}{n\mu_g}+\frac{16(L_{z^\star}^2+L_{zx}C_f/\mu_g)\omega_u K\alpha^2}{n\mu_g\gamma}\right)b_f^2$$
$$+\left(\frac{32\rho^2\omega_\ell K\gamma}{n\mu_g}+\frac{16(L_{z^\star}^2+L_{zx}\rho)\rho^2\omega_u K\alpha^2}{n\mu_g\gamma}\right)b_g^2. \tag{54}$$

Applying (38) to (54) achieves (39). □

Now we are ready to prove 3.4. We first restate the theorem in a more detailed way.

Theorem B.10 (Convergence of C-SOBA). *Under the conditions that Assumptions 2.1, 2.2, 2.3, 2.4, 3.1, 3.2, 3.3 hold and* $\beta<\min\left\{\frac{2}{\mu_g+L_g},\frac{\mu_g n}{8(1+4\omega_\ell/n)L_g^2}\right\}$, $\gamma\le\min\left\{\frac{1}{L_g},\frac{\mu_g}{12(1+4\omega_\ell/n)L_g^2}\right\}$, $\rho\ge C_f/\mu_g$,

$$\alpha\le\min\left\{\frac{1}{5L_{\nabla\Phi}(1+\omega_u/n)},\sqrt{\frac{\mu_g\beta}{360L_{y^\star}^2(L_2^2+17\kappa_g^2L_1^2)(1+\omega_u/n)}},\sqrt{\frac{\mu_g\gamma}{120L_g^2(L_{z^\star}^2+L_{zx}\rho)(1+\omega_u/n)}},\right.$$
$$\left.\sqrt{\frac{\mu_g^2\beta^2}{240L_{y^\star}^2(L_2^2+17\kappa_g^2L_1^2)}},\sqrt{\frac{\mu_g^2\gamma^2}{120L_g^2L_{z^\star}^2}},\sqrt{\frac{\mu_gL_{y^\star}^2\beta}{L_{yx}^2\left(B_x^2(1+\omega_u/n)+(1+\omega_u)\sigma_1^2/n+2\omega_u b_f^2/n+2\rho^2\omega_u b_g^2/n\right)}}\right\}, \tag{55}$$

C-SOBA (Alg. 1) converges as

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right]$$
$$\le\frac{2\Delta_\Phi^0}{K\alpha}+\frac{24(L_2^2+17\kappa_g^2L_1^2)\Delta_y^0}{\mu_gK\beta}+\frac{12L_g^2\Delta_z^0}{\mu_gK\gamma}+\frac{48(L_2^2+17\kappa_g^2L_1^2)(1+\omega_\ell)\beta\sigma^2}{\mu_gn}+\left(\frac{L_{\nabla\Phi}(1+\omega_u)\alpha}{n}+\frac{24L_g^2(1+\omega_\ell)\gamma}{\mu_gn}\right.$$
$$\left.+\frac{72L_{y^\star}^2(L_2^2+17\kappa_g^2L_1^2)(1+\omega_u)\alpha^2}{\mu_gn\beta}+\frac{24L_g^2(L_{z^\star}^2+L_{zx}\rho)(1+\omega_u)\alpha^2}{\mu_gn\gamma}\right)\sigma_1^2$$
$$+\left(\frac{2L_{\nabla\Phi}\omega_u\alpha}{n}+\frac{96L_g^2\omega_\ell\gamma}{\mu_gn}+\frac{144L_{y^\star}^2(L_2^2+17\kappa_g^2L_1^2)\omega_u\alpha^2}{\mu_gn\beta}+\frac{48L_g^2(L_{z^\star}^2+L_{zx}\rho)\omega_u\alpha^2}{n\mu_g\gamma}\right)b_f^2$$
$$+\left(\frac{2L_{\nabla\Phi}\rho^2\omega_u\alpha}{n}+\frac{96(L_2^2+17\kappa_g^2L_1^2)\omega_\ell\beta}{\mu_gn}+\frac{48L_g^2\rho^2\omega_\ell\gamma}{\mu_gn}+\frac{144L_{y^\star}^2\rho^2(L_2^2+16\kappa_g^2L_1^2)\omega_u\alpha^2}{\mu_gn\beta}\right.$$
$$\left.+\frac{24L_g^2\rho^2(L_{z^\star}^2+L_{zx}\rho)\omega_u\alpha^2}{\mu_gn\gamma}\right)b_g^2. \tag{56}$$

If we further choose parameters as

$$\beta=\left(\frac{\mu_g+L_g}{2}+\frac{8(1+4\omega_\ell/n)L_g^2}{\mu_g}+\sqrt{\frac{2K\left((1+\omega_\ell)\sigma^2+2\omega_\ell b_g^2\right)}{n\Delta_y^0}}\right)^{-1},$$
$$\gamma=\left(L_g+\frac{12(1+4\omega_\ell/n)L_g^2}{\mu_g}+\sqrt{\frac{2K\left((1+\omega_\ell)\sigma_1^2+4\omega_\ell b_f^2+2(C_f^2/\mu_g^2)\omega_\ell b_g^2\right)}{n\Delta_z^0}}\right)^{-1},$$
$$\rho=\frac{C_f}{\mu_g},$$

$$\alpha = \Bigg(5L_{\nabla\Phi}(1+\omega_u/n) + \sqrt{\frac{360L_{y^\star}^2(L_2^2+17\kappa_g^2L_1^2)(1+\omega_u/n)}{\mu_g\beta}} + \sqrt{\frac{120L_g^2(L_{z^\star}^2+L_{zx}C_f/\mu_g)(1+\omega_u/n)}{\mu_g\gamma}}$$
$$+ \sqrt{\frac{240L_{y^\star}^2(L_2^2+17\kappa_g^2L_1^2)}{\mu_g^2\beta^2}} + \sqrt{\frac{120L_g^2L_{z^\star}^2}{\mu_g^2\gamma^2}}$$
$$+ \sqrt{\frac{L_{yx}^2\left(B_x^2(1+\omega_u/n)+(1+\omega_u)\sigma_1^2/n+2\omega_u b_f^2/n+2C_f^2\omega_u b_g^2/(\mu_g^2 n)\right)}{\mu_g L_{y^\star}^2\beta}}\Bigg)^{-1},$$

C-SOBA (Alg. 1) converges as order

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right] = \mathcal{O}\Bigg(\frac{\sqrt{(1+\omega_\ell+\omega_u)\Delta}\sigma + \sqrt{(\omega_\ell+\omega_u)\Delta}(b_f+b_g)}{\sqrt{nK}}$$
$$+ \frac{\Delta^{3/4}(\sqrt[4]{1+\omega_\ell}\sqrt{\sigma}+\sqrt[4]{\omega_\ell}\sqrt{b_g})\left(\sqrt{1+\omega_u}\sigma+\sqrt{\omega_u}(b_f+b_g)+\sqrt{n+\omega_u}B_x\right)}{(nK)^{3/4}}$$
$$+ \frac{\sqrt{(1+\omega_u)(1+\omega_\ell/n)}\Delta\sigma + \sqrt{\omega_u(1+\omega_\ell/n)}(b_f+b_g)}{\sqrt{n}K}$$
$$+ \frac{\sqrt{(1+\omega_\ell/n)(1+\omega_u/n)}\Delta B_x}{K} + \frac{(1+\omega_\ell/n+\omega_u/n)\Delta}{K}\Bigg),$$

where $\Delta_\Phi^0 \triangleq \Phi(x^0)$, $\Delta_y^0 \triangleq \|y^0 - y_\star^0\|_2^2$, $\Delta_z^0 \triangleq \|z^0 - z_\star^0\|_2^2$, *and* $\Delta \triangleq \max\{\Delta_\Phi^0, \Delta_x^0, \Delta_y^0, \Delta_z^0\}$.

Proof. By $L_{\nabla\Phi}$-smoothness of Φ (Lemma B.2), we have

$$\begin{aligned}
&\mathbb{E}\left[\Phi(x^{k+1})\right]\\
\leq&\mathbb{E}\left[\Phi(x^k)\right] + \mathbb{E}\left[\langle\nabla\Phi(x^k), x^{k+1}-x^k\rangle\right] + \frac{L_{\nabla\Phi}}{2}\mathcal{X}_+^k\\
=&\mathbb{E}\left[\Phi(x^k)\right] - \alpha\mathbb{E}\left[\mathbb{E}_k\left[\left\langle\nabla\Phi(x^k), \hat{D}_x^k\right\rangle\right]\right] + \frac{L_{\nabla\Phi}\alpha^2}{2}\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right]\\
=&\mathbb{E}\left[\Phi(x^k)\right] - \frac{\alpha}{2}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right] - \frac{\alpha}{2}\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right] + \frac{\alpha}{2}\mathbb{E}\left[\left\|\nabla\Phi(x^k)-\mathbb{E}_k(D_x^k)\right\|_2^2\right] + \frac{L_{\nabla\Phi}\alpha^2}{2}\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right].
\end{aligned} \tag{57}$$

Summing (57) from $k=0$ to $K-1$, we obtain

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right] \leq \frac{2\Phi(x^0)}{\alpha} - \sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right] + \sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)-\mathbb{E}_k(D_x^k)\right\|_2^2\right] + L_{\nabla\Phi}\alpha\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\hat{D}_x^k\right\|_2^2\right]. \tag{58}$$

Applying Lemma B.4, B.8, B.9 to (58), we obtain

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right]\\
\leq&\frac{2\Delta_\Phi^0}{K\alpha} + \frac{24(L_2^2+17\kappa_g^2L_1^2)\Delta_y^0}{\mu_g K\beta} + \frac{12L_g^2\Delta_z^0}{\mu_g K\gamma} + \frac{48(L_2^2+17\kappa_g^2L_1^2)(1+\omega_\ell)\beta\sigma^2}{\mu_g n} + \Bigg(\frac{L_{\nabla\Phi}(1+\omega_u)\alpha}{n} + \frac{24L_g^2(1+\omega_\ell)\gamma}{\mu_g n}\\
&+ \frac{72L_{y^\star}^2(L_2^2+17\kappa_g^2L_1^2)(1+\omega_u)\alpha^2}{\mu_g n\beta} + \frac{24L_g^2(L_{z^\star}^2+L_{zx}\rho)(1+\omega_u)\alpha^2}{\mu_g n\gamma}\Bigg)\sigma_1^2 - \frac{C_1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)\right\|_2^2\right]
\end{aligned}$$

$$+\left(\frac{2L_{\nabla\Phi}\omega_u\alpha}{n}+\frac{96L_g^2\omega_\ell\gamma}{\mu_g n}+\frac{144L_{y^\star}^2(L_2^2+17\kappa_g^2L_1^2)\omega_u\alpha^2}{\mu_g n\beta}+\frac{48L_g^2(L_{z^\star}^2+L_{zx}\rho)\omega_u\alpha^2}{n\mu_g\gamma}\right)b_f^2$$
$$+\left(\frac{2L_{\nabla\Phi}\rho^2\omega_u\alpha}{n}+\frac{96(L_2^2+17\kappa_g^2L_1^2)\omega_\ell\beta}{\mu_g n}+\frac{48L_g^2\rho^2\omega_\ell\gamma}{\mu_g n}+\frac{144L_{y^\star}^2\rho^2(L_2^2+16\kappa_g^2L_1^2)\omega_u\alpha^2}{\mu_g n\beta}\right.$$
$$\left.+\frac{24L_g^2\rho^2(L_{z^\star}^2+L_{zx}\rho)\omega_u\alpha^2}{\mu_g n\gamma}\right)b_g^2, \tag{59}$$

where

$$C_1\triangleq 1-\left(L_{\nabla\Phi}(1+\omega_u/n)\alpha+\frac{72L_{y^\star}^2(L_2^2+17\kappa_g^2L_1^2)(1+\omega_u/n)\alpha^2}{\mu_g\beta}+\frac{24L_g^2(L_{z^\star}^2+L_{zx}\rho)(1+\omega_u/n)\alpha^2}{\mu_g\gamma}\right.$$
$$\left.+\frac{48L_{y^\star}^2(L_2^2+17\kappa_g^2L_1^2)\alpha^2}{\mu_g^2\beta^2}+\frac{24L_g^2L_{z^\star}^2\alpha^2}{\mu_g^2\gamma^2}\right). \tag{60}$$

Note that (55) implies $C_1\geq 0$, (56) is a direct result of (60). □

B.2. Proof of Theorem 4.2

Before proving Theorem 4.2, we need a few additional lemmas.

Lemma B.11 (Lower Level Convergence). *Under Assumptions 2.1, 2.2, 2.3, 2.4, 4.1, when* $\beta<\min\left\{\frac{2}{\mu_g+L_g},\frac{n\mu_g}{8\omega_\ell L_g^2}\right\}$, $\gamma\leq\min\left\{\frac{1}{L_g},\frac{n\mu_g}{36\omega_\ell L_g^2}\right\}$, $\rho\geq C_f/\mu_g$, *the following inequalities hold for CM-SOBA (Alg. 1):*

$$\sum_{k=0}^K\mathcal{Y}^k\leq\frac{2\mathcal{Y}^0}{\beta\mu_g}+\frac{4\beta K(1+\omega_\ell)\sigma^2}{n\mu_g}+\frac{8\beta K\omega_\ell b_g^2}{n\mu_g}+\frac{4L_{y^\star}^2}{\beta^2\mu_g^2}\sum_{k=0}^{K-1}\mathcal{X}_+^k, \tag{61}$$
$$\sum_{k=0}^K\mathcal{Z}^k\leq\frac{4\mathcal{Z}^0}{\gamma\mu_g}+\frac{50L_1^2\mathcal{Y}^0}{\beta\mu_g^3}+\frac{100\beta K(1+\omega_\ell)L_1^2\sigma^2}{n\mu_g^3}+\frac{6K(1+\omega_\ell)\gamma\sigma_1^2}{\mu_g n}$$
$$+\left(\frac{100L_{y^\star}^2L_1^2}{\beta^2\mu_g^4}+\frac{12L_{z^\star}^2}{\gamma^2\mu_g^2}\right)\sum_{k=0}^{K-1}\mathcal{X}_+^k+\frac{(200\beta L_1^2+24\gamma\rho^2\mu_g^2)K\omega_\ell b_g^2}{n\mu_g^3}+\frac{24K\omega_\ell\gamma b_f^2}{\mu_g n}. \tag{62}$$

Proof. By Assumptions 2.3 and 2.4, $\hat{D}_y^k$ is an unbiased estimator of $\nabla_y G^k$, thus

$$\begin{aligned}\mathbb{E}\left[\left\|y^{k+1}-y_\star^k\right\|_2^2\right]=&\mathbb{E}\left[\left\|y^k-\beta\hat{D}_y^k-y_\star^k\right\|_2^2\right]=\mathbb{E}\left[\left\|(y^k-y_\star^k-\beta\nabla_y G^k)-\beta(\hat{D}_y^k-\nabla_y G^k)\right\|_2^2\right]\\=&\mathbb{E}\left[\left\|y^k-y_\star^k-\beta\nabla_y G^k\right\|_2^2\right]+\mathbb{E}\left[\left\|\beta(\hat{D}_y^k-\nabla_y G^k)\right\|_2^2\right]\\\leq&(1-\beta\mu_g)^2\mathcal{Y}^k+\beta^2\mathbb{E}\left[\left\|\hat{D}_y^k-\nabla_y G^k\right\|_2^2\right],\end{aligned} \tag{63}$$

where the inequality uses Lemma B.1. Consequently, we have

$$\begin{aligned}\mathcal{Y}^{k+1}\leq&(1+\beta\mu_g)\mathbb{E}\left[\left\|y^{k+1}-y_\star^k\right\|_2^2\right]+\left(1+\frac{1}{\beta\mu_g}\right)\mathbb{E}\left[\left\|y_\star^{k+1}-y_\star^k\right\|_2^2\right]\\\leq&(1-\beta\mu_g)\mathcal{Y}^k+2\beta^2\mathbb{E}\left[\left\|\hat{D}_y^k-\nabla_y G^k\right\|_2^2\right]+\frac{2L_{y^\star}^2}{\beta\mu_g}\mathcal{X}_+^k,\end{aligned} \tag{64}$$

where the first inequality uses Young's inequality, and the second inequality uses (63), $\beta<1/\mu_g$ and Lemma B.2. Summing (64) from $k=0$ to $K-1$ and applying Lemma B.6, we obtain

$$\sum_{k=0}^K\mathcal{Y}^k\leq\frac{\mathcal{Y}^0}{\beta\mu_g}+\frac{2\beta}{\mu_g}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\hat{D}_y^k-\nabla_y G^k\right\|_2^2\right]+\frac{2L_{y^\star}^2}{\beta^2\mu_g^2}\sum_{k=0}^{K-1}\mathcal{X}_+^k$$

$$\leq \frac{\mathcal{Y}^0}{\beta\mu_g} + \frac{2\beta K(1+\omega_\ell)\sigma^2}{n\mu_g} + \frac{4\beta K\omega_\ell b_g^2}{n\mu_g} + \frac{2L_{y^\star}^2}{\beta^2\mu_g^2}\sum_{k=0}^{K-1}\mathcal{X}_+^k + \frac{4\omega_\ell L_g^2\beta}{n\mu_g}\sum_{k=0}^{K}\mathcal{Y}^k. \tag{65}$$

By $\beta \leq \frac{n\mu_g}{8\omega_\ell L_g^2}$, (65) implies (61). Similarly,

$$\begin{aligned}
\mathbb{E}\left[\left\|z^{k+1}-z_\star^k\right\|_2^2\right] \leq& \mathbb{E}\left[\left\|\tilde{z}^{k+1}-z_\star^k\right\|_2^2\right] = \mathbb{E}\left[\left\|z^k-\gamma\mathbb{E}_k(D_z^k)-z_\star^k\right\|_2^2\right] + \gamma^2\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k)-\hat{D}_z^k\right\|_2^2\right] \\
=& \mathbb{E}\left[\left\|(z^k-z_\star^k)-\gamma\nabla_{yy}^2G^k(z^k-z_\star^k)-\gamma(\nabla_{yy}^2G^kz_\star^k+\nabla_yF^k)\right\|_2^2\right] + \gamma^2\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k)-\hat{D}_z^k\right\|_2^2\right] \\
\leq& (1+\gamma\mu_g)(1-\gamma\mu_g)^2\mathcal{Z}^k + \left(1+\frac{1}{\gamma\mu_g}\right)\gamma^2\cdot 2L_1^2\mathcal{Y}^k + \gamma^2\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k)-\hat{D}_z^k\right\|_2^2\right],
\end{aligned} \tag{66}$$

where the first inequality uses $\rho \geq C_f/\mu_g$ and Lemma B.2, the second inequality uses Young's inequality, $\|I-\gamma\nabla_{yy}^2G^k\|_2 \leq 1-\gamma\mu_g$ and

$$\mathbb{E}\left[\left\|(\nabla_{yy}^2G^k-\nabla_{yy}^2G_\star^k)z_\star^k+(\nabla_yF^k-\nabla_yF_\star^k)\right\|_2^2\right] \leq 2\left(L_f^2+L_{g_{yy}}^2\frac{C_f^2}{\mu_g^2}\right)\mathcal{Y}^k.$$

Consequently,

$$\begin{aligned}
\mathcal{Z}^{k+1} \leq& \left(1+\frac{\gamma\mu_g}{2}\right)\mathbb{E}\left[\left\|z^{k+1}-z_\star^k\right\|_2^2\right] + \left(1+\frac{2}{\gamma\mu_g}\right)\mathbb{E}\left[\left\|z_\star^{k+1}-z_\star^k\right\|_2^2\right] \\
\leq& \left(1-\frac{\gamma\mu_g}{2}\right)\mathcal{Z}^k + \frac{6L_1^2\gamma}{\mu_g}\mathcal{Y}^k + \frac{3\gamma^2}{2}\mathbb{E}\left[\left\|\hat{D}_z^k-\mathbb{E}_k(D_z^k)\right\|_2^2\right] + \frac{3L_{z^\star}^2}{\gamma\mu_g}\mathcal{X}_+^k,
\end{aligned} \tag{67}$$

where the first inequality uses Young's inequality and the second inequality uses (66), $\gamma \leq 1/\mu_g$ and Lemma B.2. Summing (67) from $K=0$ to $K-1$, we achieve

$$\begin{aligned}
\sum_{k=0}^{K}\mathcal{Z}^k \leq& \frac{2\mathcal{Z}^0}{\gamma\mu_g} + \frac{12L_1^2}{\mu_g^2}\sum_{k=0}^{K}\mathcal{Y}^k + \frac{3\gamma}{\mu_g}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\hat{D}_z^k-\mathbb{E}_k(D_z^k)\right\|_2^2\right] + \frac{6L_{z^\star}^2}{\gamma^2\mu_g^2}\sum_{k=0}^{K-1}\mathcal{X}_+^k \\
\leq& \frac{2\mathcal{Z}^0}{\gamma\mu_g} + \frac{3K(1+\omega_\ell)\gamma\sigma_1^2}{\mu_g n} + \frac{6L_{z^\star}^2}{\gamma^2\mu_g^2}\sum_{k=0}^{K-1}\mathcal{X}_+^k + \left(\frac{12L_1^2}{\mu_g^2}+\frac{18\omega_\ell L_1^2\gamma}{n\mu_g}\right)\sum_{k=0}^{K}\mathcal{Y}^k + \frac{18\omega_\ell L_g^2\gamma}{n\mu_g}\sum_{k=0}^{K}\mathcal{Z}^k + \frac{12K\omega_\ell\gamma b_f^2}{\mu_g n} \\
&+ \frac{12\rho^2K\omega_\ell\gamma b_g^2}{n\mu_g} \\
\leq& \frac{2\mathcal{Z}^0}{\gamma\mu_g} + \frac{3K(1+\omega_\ell)\gamma\sigma_1^2}{\mu_g n} + \frac{6L_{z^\star}^2}{\gamma^2\mu_g^2}\sum_{k=0}^{K-1}\mathcal{X}_+^k + \frac{25L_1^2}{2\mu_g^2}\sum_{k=0}^{K}\mathcal{Y}^k + \frac{1}{2}\sum_{k=0}^{K}\mathcal{Z}^k + \frac{12K\omega_\ell\gamma b_f^2}{\mu_g n} + \frac{12\rho^2K\omega_\ell\gamma b_g^2}{n\mu_g},
\end{aligned} \tag{68}$$

where the second inequality uses Lemma B.6 and $\rho \geq C_f/\mu_g$, the third inequality uses $\gamma \leq \frac{n\mu_g}{36\omega_\ell L_g^2}$. (68) further implies

$$\sum_{k=0}^{K}\mathcal{Z}^k \leq \frac{4\mathcal{Z}^0}{\gamma\mu_g} + \frac{6K(1+\omega_\ell)\gamma\sigma_1^2}{\mu_g n} + \frac{12L_{z^\star}^2}{\gamma^2\mu_g^2}\sum_{k=0}^{K-1}\mathcal{X}_+^k + \frac{25L_1^2}{\mu_g^2}\sum_{k=0}^{K}\mathcal{Y}^k + \frac{24K\omega_\ell\gamma b_f^2}{\mu_g n} + \frac{24\rho^2K\omega_\ell\gamma b_g^2}{n\mu_g}. \tag{69}$$

Applying (61) to (69), we achieve (62). □

Lemma B.12 (Global Vanilla Compression Error in Upper Level)**.** *Under Assumptions 2.1, 2.2, 2.3, 2.4, 4.1, the following inequality holds for CM-SOBA (Alg. 1):*

$$\begin{aligned}
\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\hat{D}_x^k-\mathbb{E}_k(D_x^k)\right\|_2^2\right] \leq& \frac{(1+\omega_u)K\sigma_1^2}{n} + \frac{6\omega_u}{n}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right] + \frac{6\omega_uL_2^2}{n}\sum_{k=0}^{K}\mathcal{Y}^k + \frac{6\omega_uL_g^2}{n}\sum_{k=0}^{K}\mathcal{Z}^k \\
&+ \frac{6\omega_uKb_f^2}{n} + \frac{6\rho^2\omega_uKb_g^2}{n}.
\end{aligned} \tag{70}$$

Proof. By Assumptions 2.3 and 2.4, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{D}_x^k-\mathbb{E}_k(D_x^k)\right\|_2^2\right]=&\mathbb{E}\left[\left\|(\hat{D}_x^k-D_x^k)+(D_x^k-\mathbb{E}_k(D_x^k))\right\|_2^2\right]=\mathbb{E}\left[\left\|\hat{D}_x^k-D_x^k\right\|_2^2\right]+\mathbb{E}\left[\left\|D_x^k-\mathbb{E}_k(D_x^k)\right\|_2^2\right]\\ =&\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left[\left\|\mathcal{C}_i^u(D_{x,i}^k)-D_{x,i}^k\right\|_2^2\right]+\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left[\left\|D_{x,i}^k-\mathbb{E}_k(D_{x,i}^k)\right\|_2^2\right]\\ \le&\frac{\omega_u}{n^2}\sum_{i=1}^n\mathbb{E}\left[\left\|D_{x,i}^k\right\|_2^2\right]+\frac{\sigma_1^2}{n},\end{aligned}\tag{71}$$

where the inequality uses Assumption 2.4 and Lemma B.3. We next bound the second moment of $D_{x,i}^k$.

$$\begin{aligned}\mathbb{E}\left[\left\|D_{x,i}^k\right\|_2^2\right]=&\mathbb{E}\left[\left\|(D_{x,i}^k-\mathbb{E}_k(D_{x,i}^k))+\mathbb{E}_k(D_{x,i}^k)\right\|_2^2\right]\\ =&\mathbb{E}\left[\left\|D_{x,i}^k-\mathbb{E}_k(D_{x,i}^k)\right\|_2^2\right]+\mathbb{E}\left[\left\|\mathbb{E}_k(D_{x,i}^k)\right\|_2^2\right]\\ \le&\sigma_1^2+6\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right]+6\mathbb{E}\left[\left\|\nabla_xF_{\star,i}^k-\nabla_xF_\star^k\right\|_2^2\right]+6\mathbb{E}\left[\left\|(\nabla_{xy}^2G_{\star,i}^k-\nabla_{xy}^2G_\star^k)z^k\right\|_2^2\right]\\ &+6\mathbb{E}\left[\left\|(\nabla_{xy}^2G_i^k-\nabla_{xy}^2G_{\star,i}^k)z^k\right\|_2^2\right]+6\mathbb{E}\left[\left\|\nabla_{xy}^2G_\star^k(z^k-z_\star^k)\right\|_2^2\right]+6\mathbb{E}\left[\left\|\nabla_xF_i^k-\nabla_xF_{\star,i}^k\right\|_2^2\right]\\ \le&\sigma_1^2+6\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right]+6\mathbb{E}\left[\left\|\nabla_xF_{\star,i}^k-\nabla_xF_\star^k\right\|_2^2\right]+6\rho^2\mathbb{E}\left[\left\|\nabla_{xy}^2G_{\star,i}^k-\nabla_{xy}^2G_\star^k\right\|_2^2\right]+6L_2^2\mathcal{Y}^k\\ &+6L_g^2\mathcal{Z}^k,\end{aligned}\tag{72}$$

where the first inequality uses Lemma B.3 and Cauchy-Schwarz inequality, and the second inequality uses Assumption 2.1 and Lemma B.2. Combining (71) (72) and applying Assumption 4.1, we obtain

$$\mathbb{E}\left[\left\|\hat{D}_x^k-\mathbb{E}_k(D_x^k)\right\|_2^2\right]\le\frac{(1+\omega_u)\sigma_1^2}{n}+\frac{6\omega_u}{n}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right]+\frac{6\omega_uL_2^2}{n}\mathcal{Y}^k+\frac{6\omega_uL_g^2}{n}\mathcal{Z}^k+\frac{6\omega_ub_f^2}{n}+\frac{6\rho^2\omega_ub_g^2}{n}.\tag{73}$$

Summing (73) from $k=0$ to $K-1$ achieves (70). □

Lemma B.13 (Momentum-Gradient Bias)**.** *Under Assumptions 2.1, 2.2, 2.3, 2.4, 4.1, assuming $\rho\ge C_f/\mu_g$, the following inequality holds for CM-SOBA (Alg. 1):*

$$\begin{aligned}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|h_x^k-\nabla\Phi(x^k)\right\|_2^2\right]\le&\frac{\|h_x^0-\nabla\Phi(x^0)\|^2}{\theta}+\frac{(1+\omega_u)K\theta\sigma_1^2}{n}+\frac{6\omega_u\theta}{n}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right]\\ &+6L_2^2\left(1+\frac{\omega_u\theta}{n}\right)\sum_{k=0}^K\mathcal{Y}^k+6L_g^2\left(1+\frac{\omega_u\theta}{n}\right)\sum_{k=0}^K\mathcal{Z}^k+\frac{2L_{\nabla\Phi}^2}{\theta^2}\sum_{k=0}^{K-1}\mathcal{X}_+^k\\ &+\frac{6\omega_uK\theta b_f^2}{n}+\frac{6\rho^2\omega_uK\theta b_g^2}{n}.\end{aligned}\tag{74}$$

Further assuming $\beta<\min\left\{\frac{2}{\mu_g+L_g},\frac{n\mu_g}{8\omega_\ell L_g^2}\right\}$, $\gamma\le\min\left\{\frac{1}{L_g},\frac{n\mu_g}{36\omega_\ell L_g^2}\right\}$, $\theta\le\min\left\{1,\frac{n}{12\omega_u}\right\}$ and applying Lemma B.11, we have

$$\begin{aligned}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|h_x^k-\nabla\Phi(x^k)\right\|_2^2\right]\le&\frac{\|h_x^0-\nabla\Phi(x^0)\|^2}{\theta}+\frac{13(L_2^2+25\kappa_g^2L_1^2)\mathcal{Y}^0}{\beta\mu_g}+\frac{26L_g^2\mathcal{Z}^0}{\gamma\mu_g}\\ &+\frac{26(L_2^2+25\kappa_g^2L_1^2)(1+\omega_\ell)K\beta\sigma^2}{n\mu_g}+\left(\frac{(1+\omega_u)\theta}{n}+\frac{39L_g^2(1+\omega_\ell)\gamma}{\mu_gn}\right)\cdot K\sigma_1^2\\ &+\left(\frac{2L_{\nabla\Phi}^2}{\theta^2}+\frac{26L_{y^\star}^2(L_2^2+25\kappa_g^2L_1^2)}{\beta^2\mu_g^2}+\frac{78L_{z^\star}^2L_g^2}{\gamma^2\mu_g^2}\right)\sum_{k=0}^{K-1}\mathcal{X}_+^k\end{aligned}$$

$$+\left(\frac{6\rho^2\omega_u\theta}{n}+\frac{52(L_2^2+25\kappa_g^2L_1^2)\omega_\ell\beta}{n\mu_g}+\frac{156L_g^2\rho^2\omega_\ell\gamma}{n\mu_g}\right)\cdot Kb_g^2$$
$$+\left(\frac{6\omega_u\theta}{n}+\frac{156L_g^2\omega_\ell\gamma}{n\mu_g}\right)\cdot Kb_f^2+\frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right].$$

Proof. Note that $\hat{D}_x^k$ and x^{k+1} are mutually independent conditioned on $\mathcal{F}^k$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|h_x^{k+1}-\nabla\Phi(x^{k+1})\right\|_2^2\right]\\
=&\mathbb{E}\left[\left\|(1-\theta)(h_x^k-\nabla\Phi(x^k))+\theta(\hat{D}_x^k-\mathbb{E}_k(D_x^k))+\theta(\mathbb{E}_k(D_x^k)-\nabla\Phi(x^k))+(\nabla\Phi(x^k)-\nabla\Phi(x^{k+1})\right\|_2^2\right]\\
=&\mathbb{E}\left[\left\|(1-\theta)(h_x^k-\nabla\Phi(x^k))+\theta(\mathbb{E}_k(D_x^k)-\nabla\Phi(x^k))+(\nabla\Phi(x^k)-\nabla\Phi(x^{k+1})\right\|_2^2\right]+\theta^2\mathbb{E}\left[\left\|\hat{D}_x^k-\mathbb{E}_k(D_x^k)\right\|_2^2\right].
\end{aligned}\tag{75}$$

By Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left\|(1-\theta)(h_x^k-\nabla\Phi(x^k))+\theta(\mathbb{E}_k(D_x^k)-\nabla\Phi(x^k))+(\nabla\Phi(x^k)-\nabla\Phi(x^{k+1})\right\|_2^2\right]\\
\le&(1-\theta)\mathbb{E}\left[\left\|h_x^k-\nabla\Phi(x^k)\right\|_2^2\right]+\theta\mathbb{E}\left[\left\|(\mathbb{E}_k(D_x^k)-\nabla\Phi(x^k))+\frac{1}{\theta}\cdot(\nabla\Phi(x^k)-\nabla\Phi(x^{k+1}))\right\|_2^2\right]\\
\le&(1-\theta)\mathbb{E}\left[\left\|h_x^k-\nabla\Phi(x^k)\right\|_2^2\right]+2\theta\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)-\nabla\Phi(x^k)\right\|_2^2\right]+\frac{2}{\theta}\mathbb{E}\left[\left\|\nabla\Phi(x^k)-\nabla\Phi(x^{k+1})\right\|_2^2\right]\\
\le&(1-\theta)\mathbb{E}\left[\left\|h_x^k-\nabla\Phi(x^k)\right\|_2^2\right]+2\theta\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)-\nabla\Phi(x^k)\right\|_2^2\right]+\frac{2L_{\nabla\Phi}^2}{\theta}\mathcal{X}_+^k,
\end{aligned}\tag{76}$$

where the second inequality uses Cauchy-Schwarz inequality, and the third inequality uses Lemma B.2. Summing (75)(76) from $k=0$ to $K-1$ and applying Lemma B.4 and B.12, we obtain (74). □

Now we are ready to prove Theorem 4.2. We first restate the theorem in a more detailed way.

Theorem B.14 (Convergence of CM-SOBA)**.** *Under Assumptions 2.1, 2.2, 2.3, 2.4, 4.1 and assuming* $\beta<\min\left\{\frac{2}{\mu_g+L_g},\frac{\mu_g n}{8\omega_\ell L_g^2}\right\}$, $\gamma\le\min\left\{\frac{1}{L_g},\frac{\mu_g n}{36\omega_\ell L_g^2}\right\}$, $\theta\le\min\left\{1,\frac{n}{12\omega_u}\right\}$, $\rho\ge C_f\mu_g$, $\alpha\le\min\{\frac{1}{2L_{\nabla\Phi}},C_2\}$ *with*

$$C_2^{-2}\triangleq 2\cdot\left(\frac{2L_{\nabla\Phi}^2}{\theta^2}+\frac{26(L_2^2+25\kappa_g^2L_1^2)L_{y^\star}^2}{\beta^2\mu_g^2}+\frac{78L_g^2L_{z^\star}^2}{\gamma^2\mu_g^2}\right),$$

CM-SOBA (Alg. 1) converges as

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right]\\
\le&\frac{4\Delta_\Phi^0}{K\alpha}+\frac{2\Delta_x^0}{K\theta}+\frac{26(L_2^2+25\kappa_g^2L_1^2)\Delta_y^0}{\mu_gK\beta}+\frac{52L_g^2\Delta_z^0}{\mu_gK\gamma}+\frac{52(1+\omega_\ell)(L_2^2+25\kappa_g^2L_1^2)\beta}{\mu_g n}\cdot\sigma^2\\
&+\left(\frac{2(1+\omega_u)\theta}{n}+\frac{78L_g^2(1+\omega_\ell)\gamma}{\mu_g n}\right)\cdot\sigma_1^2+\left(\frac{12\rho^2\omega_u\theta}{n}+\frac{104(L_2^2+25\kappa_g^2L_1^2)\omega_\ell\beta}{n\mu_g}+\frac{312L_g^2\rho^2\omega_\ell\gamma}{n\mu_g}\right)\cdot b_g^2\\
&+\left(\frac{12\omega_u\theta}{n}+\frac{312L_g^2\omega_\ell\gamma}{n\mu_g}\right)\cdot b_f^2.
\end{aligned}\tag{77}$$

If we further choose parameters as

$$\alpha=\frac{1}{2L_{\nabla\Phi}+C_2^{-1}},$$

$$\beta = \left(\frac{\mu_g + L_g}{2} + \frac{8\omega_\ell L_g^2}{\mu_g n} + \sqrt{\frac{2K\left((1+\omega_\ell)\sigma^2 + 2\omega_\ell b_g^2\right)}{n\Delta_y^0}} \right)^{-1},$$

$$\gamma = \left(L_g + \frac{36\omega_\ell L_g^2}{\mu_g n} + \sqrt{\frac{3K\left((1+\omega_\ell)\sigma_1^2 + 4\omega_\ell b_f^2 + 4\omega_\ell (C_f^2/\mu_g^2) b_g^2\right)}{2n\Delta_z^0}} \right)^{-1},$$

$$\theta = \left(1 + \frac{12\omega_u}{n} + \sqrt{\frac{K\left((1+\omega_u)\sigma_1^2 + 6\omega_u b_f^2 + 6\omega_u (C_f^2/\mu_g^2) b_g^2\right)}{n\Delta_x^0}} \right)^{-1},$$

$$\rho = \frac{C_f}{\mu_g},$$

CM-SOBA (Alg. 1) converges as order

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla\Phi(x^K)\right\|_2^2\right] = \mathcal{O}\left(\frac{\sqrt{(1+\omega_u+\omega_\ell)\Delta}\sigma + \sqrt{(\omega_u+\omega_\ell)\Delta}(b_f+b_g)}{\sqrt{nK}} + \frac{(1+\omega_u/n+\omega_\ell/n)\Delta}{K} \right)$$

where we define $\Delta_\Phi^0 \triangleq \Phi(x^0)$, $\Delta_x^0 \triangleq \|h_x^0 - \nabla\Phi(x^0)\|_2^2$, $\Delta_y^0 \triangleq \|y^0 - y_\star^0\|_2^2$, $\Delta_z^0 \triangleq \|z^0 - z_\star^0\|_2^2$, *and* $\Delta \triangleq \max\{\Delta_\Phi^0, \Delta_x^0, \Delta_y^0, \Delta_z^0\}$.

Proof. By $L_{\nabla\Phi}$-smoothness of Φ (Lemma B.2), we have

$$\begin{aligned}
&\mathbb{E}\left[\Phi(x^{k+1})\right] \\
\leq&\mathbb{E}\left[\Phi(x^k)\right] + \mathbb{E}\left[\langle\nabla\Phi(x^k), x^{k+1}-x^k\rangle\right] + \frac{L_{\nabla\Phi}}{2}\mathbb{E}\left[\left\|x^{k+1}-x^k\right\|_2^2\right] \\
=&\mathbb{E}\left[\Phi(x^k)\right] + \mathbb{E}\left[\left\langle \frac{h_x^k}{2}, x^{k+1}-x^k\right\rangle\right] + \mathbb{E}\left[\left\langle \nabla\Phi(x^k) - \frac{h_x^k}{2}, x^{k+1}-x^k\right\rangle\right] + \frac{L_{\nabla\Phi}}{2}\mathbb{E}\left[\left\|x^{k+1}-x^k\right\|_2^2\right] \\
=&\mathbb{E}\left[\Phi(x^k)\right] - \left(\frac{1}{2\alpha} - \frac{L_{\nabla\Phi}}{2}\right)\mathcal{X}_+^k + \frac{\alpha}{2}\mathbb{E}\left[\left\|\nabla\Phi(x^k) - h_x^k\right\|_2^2\right] - \frac{\alpha}{2}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right].
\end{aligned} \tag{78}$$

Summing (78) from $k=0$ to $K-1$, we have

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right] \leq \frac{2\Phi(x^0)}{\alpha} - \left(\frac{1}{\alpha^2} - \frac{L_{\nabla\Phi}}{\alpha}\right)\sum_{k=0}^{K-1}\mathcal{X}_+^k + \sum_{k=0}^{K-1}\mathbb{E}\left[\left\|h_x^k - \nabla\Phi(x^k)\right\|_2^2\right]. \tag{79}$$

By the choice of α, we have

$$\frac{1}{\alpha^2} - \frac{L_{\nabla\Phi}}{\alpha} \geq \frac{1}{2\alpha^2} \geq \frac{1}{2}C_2^{-2},$$

thus by applying Lemma B.13 to (79) we obtain (77). □

B.3. Proof of Theorem 5.1

Lemma B.15 (Bounded Lower Level Updates)**.** *Under Assumptions 2.1, 2.2, the following inequalities hold for Alg. 2:*

$$\sum_{k=0}^{K-1}\mathcal{Y}_+^k \leq 6\sum_{k=0}^{K}\mathcal{Y}^k + 3L_{y^\star}^2\sum_{k=0}^{K-1}\mathcal{X}_+^k, \tag{80}$$

$$\sum_{k=0}^{K-1}\mathcal{Z}_+^k \leq 6\sum_{k=0}^{K}\mathcal{Z}^k + 3L_{z^\star}^2\sum_{k=0}^{K-1}\mathcal{X}_+^k. \tag{81}$$

Proof. By Cauchy-Schwarz inequality and $L_{y^\star}$-Lipschitz continuity of $y^\star(x)$ (Lemma B.2), we have

$$\begin{aligned}\mathcal{Y}_+^k =&\mathbb{E}\left[\left\|(y^{k+1}-y_\star^{k+1})+(y_\star^{k+1}-y_\star^k)+(y_\star^k-y^k)\right\|_2^2\right]\\ \le&3\mathcal{Y}^{k+1}+3\mathbb{E}\left[\left\|y_\star^{k+1}-y_\star^k\right\|_2^2\right]+3\mathcal{Y}^k\\ \le&3\mathcal{Y}^{k+1}+3\mathcal{Y}^k+3L_{y_\star}^2\mathcal{X}_+^k. \end{aligned}\tag{82}$$

Summing (82) from $k=0$ to $K-1$ obtains (80). Similary, (81) is achieved by applying Cauchy-Schwarz inequality and $L_{z^\star}$-Lipschitz continuity of $z^\star(x)$. □

The following Lemma describes the contractive property when multiplying $(1+\omega)^{-1}$ to an ω-unbiased compressor.

Lemma B.16 ((He et al., 2023a), Lemma 1)**.** *Assume $\mathcal{C}:\mathbb{R}^{d_C}\to\mathbb{R}^{d_C}$ is an ω-unbiased compressor, then for any $x\in\mathbb{R}^{d_C}$, it holds that*

$$\mathbb{E}\left[\left\|\frac{1}{1+\omega}\mathcal{C}(x)-x\right\|_2^2\right]\le\left(1-\frac{1}{1+\omega}\right)\|x\|_2^2.$$

Lemma B.17 (Bounded Difference of Local Update Directions in Lower Level)**.** *Under Assumptions 2.1 and 2.3, the following inequalities hold for EF-SOBA (Alg. 2):*

$$\mathbb{E}\left[\left\|D_{y,i}^{k+1}-D_{y,i}^k\right\|_2^2\right]\le 2L_g^2\mathcal{X}_+^k+2L_g^2\mathcal{Y}_+^k+3\sigma^2,\tag{83}$$

$$\mathbb{E}\left[\left\|D_{z,i}^{k+1}-D_{z,i}^k\right\|_2^2\right]\le 6L_1^2\mathcal{X}_+^k+6L_1^2\mathcal{Y}_+^k+6L_g^2\mathcal{Z}_+^k+3\sigma_1^2.\tag{84}$$

Proof. For the first inequality, we have

$$\begin{aligned}\mathbb{E}\left[\left\|D_{y,i}^k-D_{y,i}^{k-1}\right\|_2^2\right]=&\mathbb{E}\left[\mathbb{E}_k\left[\|D_{y,i}^k-D_{y,i}^{k-1}\|^2\right]\right]\le\mathbb{E}\left[\left\|\mathbb{E}_k(D_{y,i}^k)-D_{y,i}^{k-1}\right\|_2^2\right]+\sigma^2\\ \le&2\mathbb{E}\left[\left\|\mathbb{E}_k(D_{y,i}^k)-\mathbb{E}_{k-1}(D_{y,i}^{k-1})\right\|_2^2\right]+2\mathbb{E}\left[\left\|D_{y,i}^{k-1}-\mathbb{E}_{k-1}(D_{y,i}^{k-1})\right\|_2^2\right]+\sigma^2\\ \le&2L_g^2\mathcal{X}_+^{k-1}+2L_g^2\mathcal{Y}_+^{k-1}+3\sigma^2,\end{aligned}\tag{85}$$

which is exactly (83), where the first inequality uses Lemma B.3, the second inequality uses Cauchy-Schwarz inequality, the third inequality uses Assumption 2.1 and Lemma B.3. Similarly, we have

$$\mathbb{E}\left[\left\|D_{z,i}^k-D_{z,i}^{k-1}\right\|_2^2\right]\le 2\mathbb{E}\left[\left\|\mathbb{E}_k(D_{z,i}^k)-\mathbb{E}_{k-1}(D_{z,i}^{k-1})\right\|_2^2\right]+3\sigma_1^2.\tag{86}$$

Since

$$\mathbb{E}_k(D_{z,i}^k)-\mathbb{E}_{k-1}(D_{z,i}^{k-1})=\nabla_{yy}^2G_i^k(z^k-z^{k-1})+(\nabla_{yy}^2G_i^k-\nabla_{yy}^2G_i^{k-1})z^{k-1}+(\nabla_yF_i^k-\nabla_yF_i^{k-1}),$$

by Cauchy-Schwarz inequality and Assumption 2.1 we have

$$\mathbb{E}\left[\left\|\mathbb{E}_k(D_{z,i}^k)-\mathbb{E}_{k-1}(D_{z,i}^{k-1})\right\|_2^2\right]\le 3L_g^2\mathcal{Z}_+^{k-1}+\left(3L_{g_{yy}}^2\rho^2+3L_f^2\right)\left(\mathcal{X}_+^{k-1}+\mathcal{Y}_+^{k-1}\right),$$

which together with (86) leads to (84). □

Lemma B.18 (Memory Bias in Lower Level)**.** *Under Assumptions 2.1, 2.2, 2.3, 2.4, and assuming $\delta_\ell=(1+\omega_\ell)^{-1}$, the following inequalities hold for EF-SOBA (Alg. 2):*

$$\begin{aligned}\sum_{k=0}^{K-2}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|m_{y,i}^{k+1}-D_{y,i}^k\right\|_2^2\right]\le&\frac{2\omega_\ell}{n}\sum_{i=1}^n\left\|\mathbb{E}(D_{y,i}^0)-m_{y,i}^0\right\|_2^2+12\omega_\ell(1+\omega_\ell)L_3^2\sum_{k=0}^{K-2}\mathcal{X}_+^k+72\omega_\ell(1+\omega_\ell)L_g^2\sum_{k=0}^{K-1}\mathcal{Y}^k\\ &+18\omega_\ell(1+\omega_\ell)(K-1)\sigma^2,\end{aligned}\tag{87}$$

$$\begin{aligned}\sum_{k=0}^{K-2}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|m_{z,i}^{k+1}-D_{z,i}^k\right\|_2^2\right]\le&\frac{2\omega_\ell}{n}\sum_{i=1}^n\left\|\mathbb{E}(D_{z,i}^0)-m_{z,i}^0\right\|_2^2+36\omega_\ell(1+\omega_\ell)L_4^2\sum_{k=0}^{K-2}\mathcal{X}_+^k+216\omega_\ell(1+\omega_\ell)L_1^2\sum_{k=0}^{K-1}\mathcal{Y}^k\\ &+216\omega_\ell(1+\omega_\ell)L_g^2\sum_{k=0}^{K-1}\mathcal{Z}^k+18\omega_\ell(1+\omega_\ell)(K-1)\sigma_1^2.\end{aligned}\tag{88}$$

Proof. By the choice of δ_ℓ, we have

$$\begin{aligned}\mathbb{E}\left[\left\|m_{y,i}^{k+1}-D_{y,i}^{k}\right\|_2^2\right] =&\mathbb{E}\left[\left\|m_{y,i}^{k}+\frac{1}{1+\omega_\ell}\mathcal{C}_i^\ell(D_{y,i}^{k}-m_{y,i}^{k})-D_{y,i}^{k}\right\|_2^2\right]\\ \le&\left(1-\frac{1}{1+\omega_\ell}\right)\mathbb{E}\left[\left\|D_{y,i}^{k}-m_{y,i}^{k}\right\|_2^2\right]\\ \le&\left(1-\frac{1}{2(1+\omega_\ell)}\right)\mathbb{E}\left[\left\|m_{y,i}^{k}-D_{y,i}^{k-1}\right\|_2^2\right]+3\omega_\ell\mathbb{E}\left[\left\|D_{y,i}^{k}-D_{y,i}^{k-1}\right\|_2^2\right],\end{aligned}\tag{89}$$

where the first inequality uses Lemma B.16, and the second inequality uses Young's inequality. Applying Lemma B.17 to (89), we obtain

$$\mathbb{E}\left[\left\|m_{y,i}^{k+1}-D_{y,i}^{k}\right\|_2^2\right]\le\left(1-\frac{1}{2(1+\omega_\ell)}\right)\mathbb{E}\left[\left\|m_{y,i}^{k}-D_{y,i}^{k-1}\right\|_2^2\right]+6\omega_\ell L_g^2\mathcal{X}_+^{k-1}+6\omega_\ell L_g^2\mathcal{Y}_+^{k-1}+9\omega_\ell\sigma^2.\tag{90}$$

For the initial term, we have

$$\mathbb{E}\left[\left\|m_{y,i}^{1}-D_{y,i}^{0}\right\|_2^2\right]\le\left(1-\frac{1}{1+\omega_\ell}\right)\mathbb{E}\left[\left\|D_{y,i}^{0}-m_{y,i}^{0}\right\|_2^2\right]\le\frac{\omega_\ell}{1+\omega_\ell}\mathbb{E}\left[\left\|\mathbb{E}(D_{y,i}^{0})-m_{y,i}^{0}\right\|_2^2\right]+\frac{\omega_\ell\sigma^2}{1+\omega_\ell},\tag{91}$$

where the first inequality uses Lemma B.16, and the second inequality uses Lemma B.3. Averaging (90) from $i=1$ to n and summing from $k=1$ to $K-2$ and applying (91), we reach

$$\begin{aligned}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|m_{y,i}^{k+1}-D_{y,i}^{k}\right\|_2^2\right]\le&\frac{2\omega_\ell}{n}\sum_{i=1}^{n}\left\|\mathbb{E}(D_{y,i}^{0})-m_{y,i}^{0}\right\|_2^2+12\omega_\ell(1+\omega_\ell)L_g^2\sum_{k=0}^{K-1}\mathcal{X}_+^k\\&+12\omega_\ell(1+\omega_\ell)L_g^2\sum_{k=0}^{K-1}\mathcal{Y}_+^k+18\omega_\ell(1+\omega_\ell)K\sigma^2.\end{aligned}\tag{92}$$

Applying Lemma B.15 to (92), we reach (87). Similarly, we have

$$\mathbb{E}\left[\left\|m_{z,i}^{k+1}-D_{z,i}^{k}\right\|_2^2\right]\le\left(1-\frac{1}{2(1+\omega_\ell)}\right)\mathbb{E}\left[\left\|m_{z,i}^{k}-D_{z,i}^{k-1}\right\|_2^2\right]+3\omega_\ell\mathbb{E}\left[\left\|D_{z,i}^{k}-D_{z,i}^{k-1}\right\|_2^2\right].\tag{93}$$

Applying Lemma B.17 to (93) leads to

$$\mathbb{E}\left[\left\|m_{z,i}^{k+1}-D_{z,i}^{k}\right\|_2^2\right]\le\left(1-\frac{1}{2(1+\omega_\ell)}\right)\mathbb{E}\left[\left\|m_{z,i}^{k}-D_{z,i}^{k-1}\right\|_2^2\right]+18\omega_\ell L_g^2\mathcal{Z}_+^{k-1}+18\omega_\ell L_1^2\left(\mathcal{X}_+^{k-1}+\mathcal{Y}_+^{k-1}\right)+9\omega_\ell\sigma_1^2.\tag{94}$$

For the initial term, we have

$$\mathbb{E}\left[\left\|m_{z,i}^{1}-D_{y,i}^{0}\right\|_2^2\right]\le\left(1-\frac{1}{1+\omega_\ell}\right)\mathbb{E}\left[\left\|D_{z,i}^{0}-m_{z,i}^{0}\right\|_2^2\right]\le\frac{\omega_\ell}{1+\omega_\ell}\mathbb{E}\left[\left\|\mathbb{E}(D_{z,i}^{0})-m_{z,i}^{0}\right\|_2^2\right]+\frac{\omega_\ell\sigma_1^2}{1+\omega_\ell},\tag{95}$$

where the first inequality uses Lemma B.16, and the second inequality uses Lemma B.3. Averaging from $i=1$ to n, summing from $k=1$ to $K-2$, applying (95) and Lemma B.15, we reach (88). □

Lemma B.19 (Local EF21 Compression Error in Lower Level)**.** *Under Assumptions 2.1, 2.2, 2.3, 2.4, assuming* $\delta_\ell=(1+\omega_\ell)^{-1}$*, the following inequalities hold for EF-SOBA (Alg. 2):*

$$\begin{aligned}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\hat{D}_{y,i}^{k}-D_{y,i}^{k}\right\|_2^2\right]\le&\frac{\omega_\ell(1+4\omega_\ell)}{n}\sum_{i=1}^{n}\left\|\mathbb{E}(D_{y,i}^{0})-m_{y,i}^{0}\right\|_2^2+4\omega_\ell\omega_1L_3^2\sum_{k=0}^{K-1}\mathcal{X}_+^k+24\omega_\ell\omega_1L_g^2\sum_{k=0}^{K}\mathcal{Y}^k\\&+6\omega_\ell\omega_1K\sigma^2.\end{aligned}\tag{96}$$

$$\begin{aligned}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\hat{D}_{z,i}^{k}-D_{z,i}^{k}\right\|_2^2\right]\le&\frac{\omega_\ell(1+4\omega_\ell)}{n}\sum_{i=1}^{n}\left\|\mathbb{E}(D_{z,i}^{0})-m_{z,i}^{0}\right\|_2^2+12\omega_\ell\omega_1L_4^2\sum_{k=0}^{K-1}\mathcal{X}_+^k+72\omega_\ell\omega_1L_1^2\sum_{k=0}^{K}\mathcal{Y}^k\\&+72\omega_\ell\omega_1L_g^2\sum_{k=0}^{K}\mathcal{Z}^k+6\omega_\ell\omega_1K\sigma_1^2.\end{aligned}\tag{97}$$

Proof. By the definition of $\hat{D}_{y,i}^k$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{D}_{y,i}^k - D_{y,i}^k\right\|_2^2\right] =& \mathbb{E}\left[\left\|m_{y,i}^k + \mathcal{C}(D_{y,i}^k - m_{y,i}^k) - D_{y,i}^k\right\|_2^2\right] \leq \omega_\ell \mathbb{E}\left[\left\|D_{y,i}^k - m_{y,i}^k\right\|_2^2\right] \\ \leq& 2\omega_\ell \mathbb{E}\left[\left\|m_{y,i}^k - D_{y,i}^{k-1}\right\|_2^2\right] + 2\omega_\ell \mathbb{E}\left[\left\|D_{y,i}^k - D_{y,i}^{k-1}\right\|_2^2\right] \\ \leq& 2\omega_\ell \mathbb{E}\left[\left\|m_{y,i}^k - D_{y,i}^{k-1}\right\|_2^2\right] + 4\omega_\ell L_g^2 \mathcal{X}_+^{k-1} + 4\omega_\ell L_g^2 \mathcal{Y}_+^{k-1} + 6\omega_\ell \sigma^2, \end{aligned} \tag{98}$$

where the first inequality uses Assumption 2.4, the second inequality uses Cauchy-Schwarz inequality, the third inequality uses Cauchy-Schwarz inequality and Lemma B.17. For the initial term, Assumption 2.3 implies

$$\mathbb{E}\left[\left\|\hat{D}_{y,i}^0 - D_{y,i}^0\right\|_2^2\right] \leq (1+\omega_\ell)\mathbb{E}\left[\left\|m_{y,i}^1 - D_{y,i}^0\right\|_2^2\right] \leq \omega_\ell \left\|\mathbb{E}(D_{y,i}^0) - m_{y,i}^0\right\|_2^2 + \omega_\ell \sigma^2. \tag{99}$$

Averaging (98) from $i = 1$ to n, summing from $k = 1$ to $K-1$, applying (99), Lemma B.15 and B.18, we obtain (96). Similarly, we have

$$\mathbb{E}\left[\left\|\hat{D}_{z,i}^k - D_{z,i}^k\right\|_2^2\right] \leq 2\omega_\ell \mathbb{E}\left[\left\|m_{z,i}^k - D_{z,i}^{k-1}\right\|_2^2\right] + 12\omega_\ell L_1^2 \mathcal{X}_+^{k-1} + 12\omega_\ell L_1^2 \mathcal{Y}_+^{k-1} + 12\omega_\ell L_g^2 \mathcal{Z}_+^{k-1} + 6\omega_\ell \sigma_1^2, \tag{100}$$

and

$$\mathbb{E}\left[\left\|\hat{D}_{z,i}^0 - D_{z,i}^0\right\|_2^2\right] \leq \omega_\ell \left\|\mathbb{E}(D_{z,i}^0) - m_{z,i}^0\right\|_2^2 + \omega_\ell \sigma_1^2. \tag{101}$$

Averaging (100) from $i = 1$ to n, summing from $k = 1$ to $K-1$, applying (101), Lemma B.15 and B.18, we obtain (97). □

Lemma B.20 (Global EF21 Compression Error in Lower Level)**.** *Under Assumptions 2.1, 2.2, 2.3, 2.4, assuming $\delta_\ell = (1+\omega_\ell)^{-1}$, the following inequalities hold for EF-SOBA (Alg. 2):*

$$\begin{aligned}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla_y G^k - \hat{D}_y^k\right\|_2^2\right] \leq& \frac{\omega_\ell(1+4\omega_\ell)}{n^2} \sum_{i=1}^n \left\|\mathbb{E}(D_{y,i}^0) - m_{y,i}^0\right\|_2^2 + \frac{(1+6\omega_\ell\omega_1)K\sigma^2}{n} + \frac{4\omega_\ell\omega_1 L_3^2}{n} \sum_{k=0}^{K-1} \mathcal{X}_+^k \\ &+ \frac{24\omega_\ell\omega_1 L_g^2}{n} \sum_{k=0}^{K} \mathcal{Y}^k, \end{aligned} \tag{102}$$

$$\begin{aligned}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k) - \hat{D}_z^k\right\|_2^2\right] \leq& \frac{\omega_\ell(1+4\omega_\ell)}{n^2} \sum_{i=1}^n \left\|\mathbb{E}(D_{z,i}^0) - m_{z,i}^0\right\|_2^2 + \frac{(1+6\omega_\ell\omega_1)K\sigma_1^2}{n} + \frac{12\omega_\ell\omega_1 L_4^2}{n} \sum_{k=0}^{K-1} \mathcal{X}_+^k \\ &+ \frac{72\omega_\ell\omega_1 L_1^2}{n} \sum_{k=0}^{K} \mathcal{Y}^k + \frac{72\omega_\ell\omega_1 L_g^2}{n} \sum_{k=0}^{K} \mathcal{Z}^k. \end{aligned} \tag{103}$$

Proof. By Assumption 2.3, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla_y G^k - \hat{D}_y^k\right\|_2^2\right] =& \mathbb{E}\left[\left\|(\hat{D}_y^k - D_y^k) + (D_y^k - \nabla_y G^k)\right\|_2^2\right] = \mathbb{E}\left[\left\|\hat{D}_y^k - D_y^k\right\|_2^2\right] + \mathbb{E}\left[\left\|D_y^k - \nabla_y G^k\right\|_2^2\right] \\ \leq& \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}\left[\left\|\hat{D}_{y,i}^k - D_{y,i}^k\right\|_2^2\right] + \frac{1}{n}\sigma^2. \end{aligned} \tag{104}$$

Summing (104) from $k = 0$ to $K-1$ and applying Lemma B.19, we obtain (102). Similarly, we have

$$\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k) - \hat{D}_z^k\right\|_2^2\right] \leq \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}\left[\left\|\hat{D}_{z,i}^k - D_{z,i}^k\right\|_2^2\right] + \frac{1}{n}\sigma_1^2. \tag{105}$$

Summing (105) from $k = 0$ to $K-1$ and applying Lemma B.19, we obtain (103). □

Lemma B.21 (Lower Level Convergence). *Under Assumptions 2.1, 2.2, 2.3, 2.4, assuming* $\beta < \min\left\{\frac{2}{\mu_g+L_g}, \frac{\mu_g n}{96\omega_\ell\omega_1 L_g^2}\right\}$, $\gamma \le \min\left\{\frac{1}{L_g}, \frac{\mu_g n}{432\omega_\ell\omega_1 L_g^2}\right\}$, $\rho \ge C_f/\mu_g$, $\delta_\ell = (1+\omega_\ell)^{-1}$, *the following inequalities hold for EF-SOBA (Alg. 2):*

$$\begin{aligned}\sum_{k=0}^{K}\mathcal{Y}^k \le& \frac{2\mathcal{Y}^0}{\beta\mu_g} + \frac{4\beta\omega_\ell(1+4\omega_\ell)}{\mu_g n^2}\sum_{i=1}^{n}\left\|\mathbb{E}(D_{y,i}^0) - m_{y,i}^0\right\|_2^2 + \left(\frac{16\omega_\ell L_3^2\omega_1\beta}{\mu_g n} + \frac{4L_{y_\star^2}}{\beta^2\mu_g^2}\right)\sum_{k=0}^{K-1}\mathcal{X}_+^k \\ &+ \frac{4\beta K(1+6\omega_\ell\omega_1)\sigma^2}{\mu_g n}, \end{aligned} \tag{106}$$

$$\begin{aligned}\sum_{k=0}^{K}\mathcal{Z}^k \le& \frac{4\mathcal{Z}^0}{\gamma\mu_g} + \frac{50L_1^2\mathcal{Y}^0}{\beta\mu_g^3} + \frac{6\gamma\omega_\ell(1+4\omega_\ell)}{\mu_g n^2}\sum_{i=1}^{n}\left\|\mathbb{E}(D_{z,i}^0) - m_{z,i}^0\right\|_2^2 + \frac{100L_1^2\beta\omega_\ell(1+4\omega_\ell)}{\mu_g^2 n^2}\sum_{i=1}^{n}\left\|\mathbb{E}(D_{y,i}^0) - m_{y,i}^0\right\|_2^2 \\ &+ \frac{100K(1+6\omega_\ell\omega_1)L_1^2\beta\sigma^2}{\mu_g^3 n} + \frac{6K(1+6\omega_\ell\omega_1)\gamma\sigma_1^2}{\mu_g n} \\ &+ \left(\frac{400\omega_\ell\omega_1 L_1^2L_3^2\beta}{\mu_g^3 n} + \frac{100L_1^2L_{y^\star}^2}{\beta^2\mu_g^4} + \frac{72\omega_\ell\omega_1 L_4^2\gamma}{\mu_g n} + \frac{12L_{z^\star}^2}{\gamma^2\mu_g^2}\right)\sum_{k=0}^{K-1}\mathcal{X}_+^k. \end{aligned} \tag{107}$$

Proof. By Assumptions 2.3 and 2.4, $\hat{D}_y^k$ is an unbiased estimator of $\nabla_y G^k$, thus we have

$$\begin{aligned}\mathbb{E}\left[\left\|y^{k+1} - y_\star^k\right\|_2^2\right] =& \mathbb{E}\left[\left\|y^k - \beta\hat{D}_y^k - y_\star^k\right\|_2^2\right] = \mathbb{E}\left[\left\|(y^k - y_\star^k - \beta\nabla_y G^k) - \beta(\hat{D}_y^k - \nabla_y G^k)\right\|_2^2\right] \\ =& \mathbb{E}\left[\left\|y^k - y_\star^k - \beta\nabla_y G^k\right\|_2^2\right] + \mathbb{E}\left[\left\|\beta(\hat{D}_y^k - \nabla_y G^k)\right\|_2^2\right] \\ \le& (1-\beta\mu_g)^2\mathcal{Y}^k + \beta^2\mathbb{E}\left[\left\|\hat{D}_y^k - \nabla_y G^k\right\|_2^2\right], \end{aligned} \tag{108}$$

where the inequality uses Lemma B.1. Consequently,

$$\begin{aligned}\mathcal{Y}^{k+1} \le& (1+\beta\mu_g)\mathbb{E}\left[\left\|y^{k+1} - y_\star^k\right\|_2^2\right] + \left(1 + \frac{1}{\beta\mu_g}\right)\mathbb{E}\left[\left\|y_\star^{k+1} - y_\star^k\right\|_2^2\right] \\ \le& (1-\beta\mu_g)\mathcal{Y}^k + 2\beta^2\mathbb{E}\left[\left\|\hat{D}_y^k - \nabla_y G^k\right\|_2^2\right] + \frac{2L_{y^\star}^2}{\beta\mu_g}\mathcal{X}_+^k, \end{aligned} \tag{109}$$

where the first inequality uses Young's inequality, the second inequality uses (109), Lemma B.2 and $\beta < 1/\mu_g$. Summing (109) from $k=0$ to $K-1$ and applying Lemma B.20, we obtain

$$\begin{aligned}\sum_{k=0}^{K}\mathcal{Y}^k \le& \frac{\mathcal{Y}^0}{\beta\mu_g} + \frac{2\beta}{\mu_g}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\hat{D}_y^k - \nabla_y G^k\right\|_2^2\right] + \frac{2L_{y^\star}^2}{\beta^2\mu_g^2}\sum_{k=0}^{K-1}\mathcal{X}_+^k \\ \le& \frac{\mathcal{Y}^0}{\beta\mu_g} + \frac{2\beta\omega_\ell(1+4\omega_\ell)}{n^2\mu_g}\sum_{i=1}^{n}\left\|\mathbb{E}(D_{y,i}^0) - m_{y,i}^0\right\|_2^2 + \frac{2\beta K(1+6\omega_\ell\omega_1)\sigma^2}{n\mu_g} + \left(\frac{8\omega_\ell\omega_1 L_3^2\beta}{n\mu_g} + \frac{2L_{y^\star}^2}{\beta^2\mu_g^2}\right)\sum_{k=0}^{K-1}\mathcal{X}_+^k \\ &+ \frac{48\omega_\ell\omega_1 L_g^2\beta}{n\mu_g}\sum_{k=0}^{K}\mathcal{Y}^k. \end{aligned} \tag{110}$$

By $\beta \le \frac{n\mu_g}{96\omega_\ell\omega_1 L_g^2}$, (110) implies (106). Similarly, by $\rho \ge C_f/\mu_g$ and Lemma B.2, we have

$$\begin{aligned}\mathbb{E}\left[\left\|z^{k+1} - z_\star^k\right\|_2^2\right] \le& \mathbb{E}\left[\left\|\tilde{z}^{k+1} - z_\star^k\right\|_2^2\right] = \mathbb{E}\left[\left\|z^k - \gamma\mathbb{E}_k(D_z^k) - z_\star^k\right\|_2^2\right] + \gamma^2\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k) - \hat{D}_z^k\right\|_2^2\right] \\ =& \mathbb{E}\left[\left\|(z^k - z_\star^k) - \gamma\nabla_{yy}^2 G^k(z^k - z_\star^k) - \gamma(\nabla_{yy}^2 G^k z_\star^k + \nabla_y F^k)\right\|_2^2\right] + \gamma^2\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k) - \hat{D}_z^k\right\|_2^2\right] \\ \le& (1+\gamma\mu_g)(1-\gamma\mu_g)^2\mathcal{Z}^k + \left(1 + \frac{1}{\gamma\mu_g}\right)\gamma^2\cdot 2L_1^2\mathcal{Y}^k + \gamma^2\mathbb{E}\left[\left\|\mathbb{E}_k(D_z^k) - \hat{D}_z^k\right\|_2^2\right], \end{aligned} \tag{111}$$

where the last inequality uses Young's inequality, $\|I-\gamma\nabla^2_{yy}G^k\|_2\le 1-\gamma\mu_g$ and

$$\mathbb{E}\left[\left\|(\nabla^2_{yy}G^k-\nabla^2_{yy}G^k_\star)z^k_\star+(\nabla_y F^k-\nabla_y F^k_\star)\right\|_2^2\right]\le 2\left(L_f^2+L_{g_{yy}}^2\frac{C_f^2}{\mu_g^2}\right)\mathcal{Y}^k.$$

Consequently,

$$\begin{aligned}\mathcal{Z}^{k+1}\le&\left(1+\frac{\gamma\mu_g}{2}\right)\mathbb{E}\left[\left\|z^{k+1}-z^k_\star\right\|_2^2\right]+\left(1+\frac{2}{\gamma\mu_g}\right)\mathbb{E}\left[\left\|z^{k+1}_\star-z^k_\star\right\|_2^2\right]\\ \le&\left(1-\frac{\gamma\mu_g}{2}\right)\mathcal{Z}^k+\frac{6L_1^2\gamma}{\mu_g}\mathcal{Y}^k+\frac{3\gamma^2}{2}\mathbb{E}\left[\left\|\hat{D}^k_z-\mathbb{E}_k(D^k_z)\right\|_2^2\right]+\frac{3L_{z^\star}^2}{\gamma\mu_g}\mathcal{X}^k_+,\end{aligned}\tag{112}$$

where the first inequality uses Young's inequality, the second inequality uses (111), Lemma B.2 and $\gamma\le 1/\mu_g$. Summing (112) from $K=0$ to $K-1$ and applying Lemma B.20 we achieve

$$\begin{aligned}&\sum_{k=0}^K\mathcal{Z}^k\\ \le&\frac{2\mathcal{Z}^0}{\gamma\mu_g}+\frac{12L_1^2}{\mu_g^2}\sum_{k=0}^K\mathcal{Y}^k+\frac{3\gamma}{\mu_g}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\hat{D}^k_z-\mathbb{E}_k(D^k_z)\right\|_2^2\right]+\frac{6L_{z^\star}^2}{\gamma^2\mu_g^2}\sum_{k=0}^{K-1}\mathcal{X}^k_+\\ \le&\frac{2\mathcal{Z}^0}{\gamma\mu_g}+\frac{3\gamma\omega_\ell(1+4\omega_\ell)}{\mu_g n^2}\sum_{i=1}^n\left\|\mathbb{E}(D^0_{z,i})-m^0_{z,i}\right\|_2^2+\frac{3K(1+6\omega_\ell\omega_1)\gamma\sigma_1^2}{\mu_g n}+\left(\frac{36\omega_\ell\omega_1L_4^2\gamma}{\mu_g n}+\frac{6L_{z^\star}^2}{\gamma^2\mu_g^2}\right)\sum_{k=0}^{K-1}\mathcal{X}^k_+\\ &+\left(\frac{12L_1^2}{\mu_g^2}+\frac{216\omega_\ell\omega_1L_1^2\gamma}{n\mu_g}\right)\sum_{k=0}^K\mathcal{Y}^k+\frac{216\omega_\ell\omega_1L_g^2\gamma}{n\mu_g}\sum_{k=0}^K\mathcal{Z}^k\\ \le&\frac{2\mathcal{Z}^0}{\gamma\mu_g}+\frac{3\gamma\omega_\ell(1+4\omega_\ell)}{\mu_g n^2}\sum_{i=1}^n\left\|\mathbb{E}(D^0_{z,i})-m^0_{z,i}\right\|_2^2+\frac{3K(1+6\omega_\ell\omega_1)\gamma\sigma_1^2}{\mu_g n}+\left(\frac{36\omega_\ell\omega_1L_4^2\gamma}{\mu_g n}+\frac{6L_{z^\star}^2}{\gamma^2\mu_g^2}\right)\sum_{k=0}^{K-1}\mathcal{X}^k_+\\ &+\frac{25L_1^2}{2\mu_g^2}\sum_{k=0}^K\mathcal{Y}^k+\frac{1}{2}\sum_{k=0}^K\mathcal{Z}^k,\end{aligned}\tag{113}$$

where the last inequality uses $\gamma\le\frac{\mu_g n}{432\omega_\ell\omega_1L_g^2}$. This further implies

$$\begin{aligned}\sum_{k=0}^K\mathcal{Z}^k\le&\frac{4\mathcal{Z}^0}{\gamma\mu_g}+\frac{6\gamma\omega_\ell(1+4\omega_\ell)}{\mu_g n^2}\sum_{i=1}^n\left\|\mathbb{E}(D^0_{z,i})-m^0_{z,i}\right\|_2^2+\frac{6K(1+6\omega_\ell\omega_1)\gamma\sigma_1^2}{\mu_g n}+\left(\frac{72\omega_\ell\omega_1L_4^2\gamma}{\mu_g n}+\frac{12L_{z^\star}^2}{\gamma^2\mu_g^2}\right)\sum_{k=0}^{K-1}\mathcal{X}^k_+\\ &+\frac{25L_1^2}{\mu_g^2}\sum_{k=0}^K\mathcal{Y}^k.\end{aligned}\tag{114}$$

Applying (106) to (114), we achieve (107). □

Lemma B.22 (Momentum-Gradient Bias)**.** *Under Assumptions 2.1, 2.2, 2.3, assuming $\rho\ge C_f/\mu_g$, the following inequality holds for EF-SOBA (Alg. 2):*

$$\sum_{k=0}^K\mathbb{E}\left[\left\|h^k_x-\nabla\Phi(x^k)\right\|_2^2\right]\le\frac{\|h^0_x-\nabla\Phi(x^0)\|^2}{\theta}+\frac{K\theta\sigma_1^2}{n}+6L_2^2\sum_{k=0}^K\mathcal{Y}^k+6L_g^2\sum_{k=0}^K\mathcal{Z}^k+\frac{2L_{\nabla\Phi}^2}{\theta^2}\sum_{k=0}^{K-1}\mathcal{X}^k_+.\tag{115}$$

Proof. Note that x^{k+1} is conditionally independent of D^k_x given filtration $\mathcal{F}^k$, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|h^{k+1}_x-\nabla\Phi(x^{k+1})\right\|_2^2\right]\\ =&\mathbb{E}\left[\left\|(1-\theta)h^k_x+\theta D^k_x-\nabla\Phi(x^{k+1})\right\|_2^2\right]\end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}\left[\mathbb{E}_k\left[(1-\theta)(h_x^k-\nabla\Phi(x^k))+\theta(D_x^k-\mathbb{E}_k(D_x^k))+\theta(\mathbb{E}_k(D_x^k)-\nabla\Phi(x^k))+(\nabla\Phi(x^k)-\nabla\Phi(x^{k+1}))\right]\right]\\
=&\mathbb{E}\left[\left\|(1-\theta)(h_x^k-\nabla\Phi(x^k))+\theta(\mathbb{E}_k(D_x^k)-\nabla\Phi(x^k))+(\nabla\Phi(x^k)-\nabla\Phi(x^{k+1}))\right\|_2^2\right]+\theta^2\mathbb{E}\left[\left\|D_x^k-\mathbb{E}_k(D_x^k)\right\|_2^2\right]\\
\le&(1-\theta)\mathbb{E}\left[\left\|h_x^k-\nabla\Phi(x^k)\right\|_2^2\right]+\theta\mathbb{E}\left[\left\|(\mathbb{E}_k(D_x^k)-\nabla\Phi(x^k))+\frac{1}{\theta}(\nabla\Phi(x^k)-\nabla\Phi(x^{k+1}))\right\|_2^2\right]+\frac{\theta^2\sigma_1^2}{n}\\
\le&(1-\theta)\mathbb{E}\left[\left\|h_x^k-\nabla\Phi(x^k)\right\|_2^2\right]+2\theta\mathbb{E}\left[\left\|\mathbb{E}_k(D_x^k)-\nabla\Phi(x^k)\right\|_2^2\right]+\frac{2L_{\nabla\Phi}^2}{\theta}\mathcal{X}_+^k+\frac{\theta^2\sigma_1^2}{n},
\end{aligned}
\tag{116}
$$

where the first inequality uses Jensen's inequality and Lemma B.3, the second inequality uses Cauchy-Schwarz inequality and Lemma B.2. Summing (116) from $k=0$ to $K-1$ and applying Lemma B.4, we achieve (115). □

Lemma B.23 (Local Momentum Bias)**.** *Under Assumptions 2.1, 2.2, 2.3, the following inequality holds for EF-SOBA (Alg. 2):*

$$
\begin{aligned}
\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|h_{x,i}^k-\mathbb{E}_k(D_{x,i}^k)\right\|_2^2\right]\le&\frac{1}{\theta n}\sum_{i=1}^{n}\|h_{x,i}^0-\mathbb{E}(D_{x,i}^0)\|^2+\frac{6L_5^2}{\theta^2}\sum_{k=0}^{K-1}\mathcal{X}_+^k+\frac{36L_2^2}{\theta^2}\sum_{k=0}^{K}\mathcal{Y}^k+\frac{36L_g^2}{\theta^2}\sum_{k=0}^{K}\mathcal{Z}^k\\
&+2K\sigma_1^2.
\end{aligned}
\tag{117}
$$

Proof. By the update rule of $h_{x,i}^k$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|h_{x,i}^k-\mathbb{E}_k(D_{x,i}^k)\right\|_2^2\right]\\
=&\mathbb{E}\left[\left\|(1-\theta)(h_{x,i}^{k-1}-\mathbb{E}_{k-1}(D_{x,i}^{k-1}))+\theta(D_{x,i}^{k-1}-\mathbb{E}_{k-1}(D_{x,i}^{k-1}))+(\mathbb{E}_{k-1}(D_{x,i}^{k-1})-\mathbb{E}_k(D_{x,i}^k))\right\|_2^2\right]\\
\le&(1-\theta)\mathbb{E}\left[\left\|h_{x,i}^{k-1}-\mathbb{E}(D_{x,i}^{k-1})\right\|_2^2\right]+2\theta\sigma_1^2+\frac{2}{\theta}\mathbb{E}\left[\left\|\mathbb{E}_k(D_{x,i}^k)-\mathbb{E}_{k-1}(D_{x,i}^{k-1})\right\|_2^2\right],
\end{aligned}
\tag{118}
$$

where the inequality uses Jensen's inequality, Cauchy-Schwarz inequality and Lemma B.3. For the last term, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbb{E}_k(D_{x,i}^k)-\mathbb{E}_{k-1}(D_{x,i}^{k-1})\right\|_2^2\right]\\
=&\mathbb{E}\left[\left\|(\nabla_{xy}^2G_i^k-\nabla_{xy}^2G_i^{k-1})z^k+\nabla_{xy}^2G_i^{k-1}(z^k-z^{k-1})+(\nabla F_i^k-\nabla F_i^{k-1})\right\|_2^2\right]\\
\le&\left(3L_{g_{xy}}^2\rho^2+3L_f^2\right)\left(\mathcal{X}_+^{k-1}+\mathcal{Y}_+^{k-1}\right)+3L_g^2\mathcal{Z}_+^{k-1},
\end{aligned}
\tag{119}
$$

where the inequality uses Cauchy-Schwarz inequality and Assumption 2.1. Averaging (118) from $i=1$ to n, summing from $k=1$ to $K-1$ and applying (119), we obtain

$$
\begin{aligned}
\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|h_{x,i}^k-\mathbb{E}_k(D_{x,i}^k)\right\|_2^2\right]\le&\frac{1}{\theta n}\sum_{i=1}^{n}\|h_{x,i}^0-\mathbb{E}(D_{x,i}^0)\|^2+2K\sigma_1^2+\frac{6L_2^2}{\theta^2}\sum_{k=0}^{K-1}\mathcal{X}_+^k+\frac{6L_2^2}{\theta^2}\sum_{k=0}^{K-1}\mathcal{Y}_+^k\\
&+\frac{6L_g^2}{\theta^2}\sum_{k=0}^{K-1}\mathcal{Z}_+^k.
\end{aligned}
\tag{120}
$$

By applying Lemma B.15 to (120), we obtain (117). □

Lemma B.24 (Global EF21 Compression Error in Upper Level)**.** *Under Assumptions 2.1, 2.2, 2.3, 2.4, assuming* $\delta_u=(1+\omega_u)^{-1}$*, and* $m_{x,i}^0=h_{x,i}^0$ *for* $i=1,\cdots,n$*, the following inequality holds for EF-SOBA (Alg. 2):*

$$
\begin{aligned}
\sum_{k=0}^{K}\mathbb{E}\left[\left\|\hat{h}_x^k-h_x^k\right\|_2^2\right]\le&\frac{6\omega_u(1+\omega_u)\theta}{n}\sum_{i=1}^{n}\|h_{x,i}^0-\mathbb{E}(D_{x,i}^0)\|^2+36\omega_u(1+\omega_u)L_5^2\sum_{k=0}^{K-1}\mathcal{X}_+^k+216\omega_u(1+\omega_u)L_2^2\sum_{k=0}^{K}\mathcal{Y}^k\\
&+216\omega_u(1+\omega_u)L_g^2\sum_{k=0}^{K}\mathcal{Z}^k+18\omega_u(1+\omega_u)K\theta^2\sigma_1^2,
\end{aligned}
\tag{121}
$$

where we define $h_x^k\triangleq\frac{1}{n}\sum_{i=1}^{n}h_{x,i}^k$.

Proof. By Lemma B.16, we have

$$\mathbb{E}\left[\left\|m_{x,i}^{k+1}-h_{x,i}^{k+1}\right\|_2^2\right]=\mathbb{E}\left[\left\|m_{x,i}^{k}+\frac{\mathcal{C}_i^u(h_{x,i}^{k+1}-m_{x,i}^k)}{1+\omega_u}-h_{x,i}^{k+1}\right\|_2^2\right]\leq\left(1-\frac{1}{1+\omega_u}\right)\mathbb{E}\left[\left\|h_{x,i}^{k+1}-m_{x,i}^k\right\|_2^2\right]. \quad (122)$$

For $\mathbb{E}\left[\left\|h_{x,i}^{k+1}-m_{x,i}^k\right\|_2^2\right]$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|h_{x,i}^{k+1}-m_{x,i}^k\right\|_2^2\right]\\
=&\mathbb{E}\left[\left\|(h_{x,i}^k-m_{x,i}^k)+\theta(D_{x,i}^k-\mathbb{E}(D_{x,i}^k))+\theta(\mathbb{E}(D_{x,i}^k)-h_{x,i}^k)\right\|_2^2\right]\\
\leq&\left(1+\frac{1}{2(1+\omega_u)}\right)\mathbb{E}\left[\left\|m_{x,i}^k-h_{x,i}^k\right\|_2^2\right]+(1+2(1+\omega_u))\theta^2\sigma_1^2+(1+2(1+\omega_u))\theta^2\mathbb{E}\left[\left\|h_{x,i}^k-\mathbb{E}(D_{x,i}^k)\right\|_2^2\right], \quad (123)
\end{aligned}$$

where the inequality uses Young's inequality and Assumption 2.3. Combining (122)(123) achieves

$$\mathbb{E}\left[\left\|m_{x,i}^{k+1}-h_{x,i}^{k+1}\right\|_2^2\right]\leq\left(1-\frac{1}{2(1+\omega_u)}\right)\mathbb{E}\left[\left\|m_{x,i}^k-h_{x,i}^k\right\|_2^2\right]+3\omega_u\theta^2\sigma_1^2+3\omega_u\theta^2\mathbb{E}\left[\left\|h_{x,i}^k-\mathbb{E}(D_{x,i}^k)\right\|_2^2\right]. \quad (124)$$

Averaging (124) from $i=1$ to n and summing from $k=0$ to $K-1$ gives

$$\begin{aligned}
\sum_{k=0}^{K}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|m_{x,i}^k-h_{x,i}^k\right\|_2^2\right]\leq&\frac{2(1+\omega_u)}{n}\sum_{i=1}^{n}\|m_{x,i}^0-h_{x,i}^0\|^2+6\omega_u(1+\omega_u)K\theta^2\sigma_1^2\\
&+6\omega_u(1+\omega_u)\theta^2\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|h_{x,i}^k-\mathbb{E}(D_{x,i}^k)\right\|_2^2\right]. \quad (125)
\end{aligned}$$

Applying Lemma B.23 to (125) and noting that $m_{x,i}^0=h_{x,i}^0$, we obtain

$$\begin{aligned}
&\sum_{k=0}^{K}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|m_{x,i}^k-h_{x,i}^k\right\|_2^2\right]\\
\leq&18\omega_u(1+\omega_u)K\theta^2\sigma_1^2+\frac{6\omega_u(1+\omega_u)\theta}{n}\sum_{i=1}^{n}\|h_{x,i}^0-\mathbb{E}(D_{x,i}^0)\|^2+36\omega_u(1+\omega_u)L_5^2\sum_{k=0}^{K-1}\mathcal{X}_+^k\\
&+216\omega_u(1+\omega_u)L_2^2\sum_{k=0}^{K}\mathcal{Y}^k+216\omega_u(1+\omega_u)L_g^2\sum_{k=0}^{K}\mathcal{Z}^k. \quad (126)
\end{aligned}$$

By the update rules we have $\hat{h}_x^k=\frac{1}{n}\sum_{i=1}^n m_{x,i}^k$, thus (121) is a direct result of (126) by applying the following inequality:

$$\sum_{k=0}^{K}\mathbb{E}\left[\left\|\hat{h}_x^k-h_x^k\right\|_2^2\right]\leq\sum_{k=0}^{K}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|m_{x,i}^k-h_{x,i}^k\right\|_2^2\right].$$

□

Lemma B.25 (Update Direction Bias in Upper Level)**.** *Under Assumptions 2.1, 2.2, 2.3, 2.4, assuming $\rho\geq C_f/\mu_g$, $\delta_u=(1+\omega_u)^{-1}$, and $m_{x,i}^0=h_{x,i}^0$ for $i=1,\cdots,n$, the following inequality holds for EF-SOBA (Alg. 2):*

$$\begin{aligned}
\sum_{k=0}^{K}\mathbb{E}\left[\left\|\hat{h}_x^k-\nabla\Phi(x^k)\right\|_2^2\right]\leq&\frac{2\|h_x^0-\nabla\Phi(x^0)\|^2}{\theta}+\frac{12\omega_u(1+\omega_u)\theta}{n}\sum_{i=1}^{n}\|h_{x,i}^0-\mathbb{E}(D_{x,i}^0)\|^2\\
&+\left(\frac{4L_{\nabla\Phi}^2}{\theta^2}+72\omega_u(1+\omega_u)L_5^2\right)\sum_{k=0}^{K-1}\mathcal{X}_+^k+12\omega_2L_2^2\sum_{k=0}^{K}\mathcal{Y}^k
\end{aligned}$$

$$+ 12\omega_2 L_g^2 \sum_{k=0}^{K} \mathcal{Z}^k + \left(\frac{2K\theta}{n} + 36\omega_u(1+\omega_u)K\theta^2\right)\sigma_1^2. \tag{127}$$

Further assume $\beta < \min\left\{\frac{2}{\mu_g + L_g}, \frac{\mu_g n}{96\omega_\ell\omega_1 L_g^2}\right\}$, $\gamma \le \min\left\{\frac{1}{L_g}, \frac{\mu_g n}{432\omega_\ell\omega_1 L_g^2}\right\}$ *and applying Lemma B.21, we have*

$$\begin{aligned}
&\sum_{k=0}^{K}\mathbb{E}\left[\left\|\hat{h}_x^k - \nabla\Phi(x^k)\right\|_2^2\right] \\
\le&\frac{2}{\theta}\cdot\|h_x^0 - \nabla\Phi(x^0)\|^2 + \frac{12\omega_u(1+\omega_u)\theta}{n}\sum_{i=1}^{n}\|h_{x,i}^0 - \mathbb{E}(D_{x,i}^0)\|^2 + \frac{24\omega_2(L_2^2+25\kappa_g^2L_1^2)}{\mu_g\beta}\cdot\mathcal{Y}^0 + \frac{48\omega_2 L_g^2}{\mu_g\gamma}\cdot\mathcal{Z}^0 \\
&+ \frac{48\omega_2(1+6\omega_\ell\omega_1)(L_2^2+25\kappa_g^2L_1^2)K\beta}{\mu_g n}\cdot\sigma^2 + \left(\frac{2\theta}{n} + 36\omega_u(1+\omega_u)\theta^2 + \frac{72\omega_2(1+6\omega_\ell\omega_1)L_g^2\gamma}{\mu_g n}\right)K\cdot\sigma_1^2 \\
&+ \frac{48\omega_2\omega_\ell(1+4\omega_\ell)(L_2^2+25\kappa_g^2L_1^2)\beta}{\mu_g n^2}\sum_{i=1}^{n}\left\|\mathbb{E}(D_{y,i}^0) - m_{y,i}^0\right\|_2^2 + \frac{72\omega_2\omega_\ell(1+4\omega_\ell)L_g^2\gamma}{\mu_g n^2}\sum_{i=1}^{n}\left\|\mathbb{E}(D_{z,i}^0) - m_{z,i}^0\right\|_2^2 \\
&+ \left[\frac{4L_{\nabla\Phi}^2}{\theta^2} + 72\omega_u(1+\omega_u)L_5^2 + 12\omega_2(L_2^2+25\kappa_g^2L_1^2)\left(\frac{4L_{y^\star}^2}{\beta^2\mu_g^2} + \frac{16\omega_\ell\omega_1 L_3^2\beta}{\mu_g n}\right)\right. \\
&\left.+ 12\omega_2 L_g^2\left(\frac{12L_{z^\star}^2}{\gamma^2\mu_g^2} + \frac{72\omega_\ell\omega_1 L_4^2\gamma}{\mu_g n}\right)\right]\cdot\sum_{k=0}^{K-1}\mathcal{X}_+^k.
\end{aligned} \tag{128}$$

Proof. By Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\left\|\hat{h}_x^k - \nabla\Phi(x^k)\right\|_2^2\right] \le 2\mathbb{E}\left[\left\|\hat{h}_x^k - h_x^k\right\|_2^2\right] + 2\mathbb{E}\left[\left\|h_x^k - \nabla\Phi(x^k)\right\|_2^2\right]. \tag{129}$$

Summing (129) from $k=0$ to K and applying Lemma B.22 and B.24, we obtain (127). □

Now we are ready to prove Theorem 5.1. We first restate the theorem in a more detailed way.

Theorem B.26 (Convergence of EF-SOBA). *Under Assumptions 2.1, 2.2, 2.3, 2.4, assuming* $\beta < \min\left\{\frac{2}{\mu_g + L_g}, \frac{\mu_g n}{96\omega_\ell\omega_1 L_g^2}\right\}$, $\gamma \le \min\left\{\frac{1}{L_g}, \frac{\mu_g n}{432\omega_\ell\omega_1 L_g^2}\right\}$, $\rho \ge C_f/\mu_g$, $\delta_\ell = (1+\omega_\ell)^{-1}$, $\delta_u = (1+\omega_u)^{-1}$, $m_{x,i}^0 = h_{x,i}^0$ *for* $i = 1, \cdots, n$, $\alpha \le \min\{\frac{1}{2L_{\nabla\Phi}}, C_3\}$ *with*

$$\begin{aligned}
C_3^{-2} \triangleq& 2\cdot\left[\frac{4L_{\nabla\Phi}^2}{\theta^2} + 72\omega_u(1+\omega_u)L_5^2 + 12\omega_2(L_2^2+25\kappa_g^2L_1^2)\left(\frac{4L_{y^\star}^2}{\beta^2\mu_g^2} + \frac{16\omega_\ell\omega_1 L_3^2\beta}{\mu_g n}\right)\right. \\
&\left.+ 12\omega_2 L_g^2\left(\frac{12L_{z^\star}^2}{\gamma^2\mu_g^2} + \frac{72\omega_\ell\omega_1 L_4^2\gamma}{\mu_g n}\right)\right],
\end{aligned}$$

EF-SOBA (Alg. 2) converges as

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right] \le& \frac{2\Delta_\Phi^0}{K\alpha} + \frac{2\Delta_x^0}{K\theta} + \frac{12\omega_u(1+\omega_u)\theta\Delta_h^0}{K} + \frac{24\omega_2(L_2^2+25\kappa_g^2L_1^2)\Delta_y^0}{\mu_g K\beta} \\
&+ \frac{48\omega_2 L_g^2\Delta_z^0}{\mu_g K\gamma} + \frac{48\omega_2(1+6\omega_\ell\omega_1)(L_2^2+25\kappa_g^2L_1^2)\beta}{\mu_g n}\cdot\sigma^2 \\
&+ \left(\frac{2\theta}{n} + 36\omega_u(1+\omega_u)\theta^2 + \frac{72\omega_2(1+6\omega_\ell\omega_1)L_g^2\gamma}{\mu_g n}\right)\cdot\sigma_1^2 \\
&+ \frac{48\omega_2\omega_\ell(1+4\omega_\ell)(L_2^2+25\kappa_g^2L_1^2)\beta}{\mu_g K n^2}\sum_{i=1}^{n}\left\|\mathbb{E}(D_{y,i}^0) - m_{y,i}^0\right\|_2^2
\end{aligned}$$

$$+\frac{72\omega_2\omega_\ell(1+4\omega_\ell)L_g^2\gamma}{\mu_g K n^2}\sum_{i=1}^n\left\|\mathbb{E}(D_{z,i}^0)-m_{z,i}^0\right\|_2^2. \tag{130}$$

If we further choose parameters as

$$\begin{aligned}
\alpha =&\frac{1}{2L_{\nabla\Phi}+C_3^{-1}},\quad \rho=C_f/\mu_g,\quad \delta_\ell=\frac{1}{1+\omega_\ell},\quad \delta_u=\frac{1}{1+\omega_u},\\
\beta =&\left(\frac{\mu_g+L_g}{2}+\frac{96\omega_\ell\omega_1 L_g^2}{\mu_g n}+\sqrt{\frac{2\omega_\ell(1+4\omega_\ell)\Delta_{m_y}^0}{n\Delta_y^0}}+\sqrt{\frac{2K\sigma^2(1+6\omega_\ell\omega_1)}{n\Delta_y^0}}\right)^{-1},\\
\gamma =&\left(L_g+\frac{432\omega_\ell\omega_1 L_g^2}{\mu_g n}+\sqrt{\frac{3\omega_\ell(1+4\omega_\ell)\Delta_{m_z}^0}{2n\Delta_z^0}}+\sqrt{\frac{K\sigma_1^2(1+6\omega_\ell\omega_1)}{n\Delta_z^0}}\right)^{-1},\\
\theta =&\left(1+\sqrt{\frac{6\omega_u(1+\omega_u)\Delta_h^0}{\Delta_x^0}}+\sqrt{\frac{K\sigma_1^2}{n\Delta_x^0}}+\sqrt[3]{\frac{18\omega_u(1+\omega_u)K\sigma_1^2}{\Delta_x^0}}\right)^{-1},
\end{aligned}$$

EF-SOBA (Alg. 2) converges as order

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right]=&\mathcal{O}\left(\frac{(1+\omega_u)^2(1+\omega_\ell)^{3/2}\sqrt{\Delta}\sigma}{\sqrt{nK}}+\frac{\omega_u^{1/3}(1+\omega_u)^{1/3}\Delta^{2/3}\sigma^{2/3}}{K^{2/3}}+\frac{(1+\omega_u)\Delta}{K}\right.\\
&\left.+\frac{(1+\omega_u)^2\sqrt{\omega_\ell(1+\omega_\ell)}\Delta}{\sqrt{n}K}+\frac{(1+\omega_u)^2\omega_\ell^3\Delta}{nK}\right),
\end{aligned}$$

where we define $\Delta_\Phi^0\triangleq\Phi(x^0)$, $\Delta_x^0\triangleq\|h_x^0-\nabla\Phi(x^0)\|_2^2$, $\Delta_h^0\triangleq\frac{1}{n}\sum_{i=1}^n\|h_{x,i}^0-\mathbb{E}(D_{x,i}^0)\|_2^2$, $\Delta_y^0\triangleq\|y^0-y_\star^0\|_2^2$, $\Delta_z^0\triangleq\|z^0-z_\star^0\|_2^2$, $\Delta_{m_y}^0\triangleq\frac{1}{n}\sum_{i=1}^n\left\|\mathbb{E}(D_{y,i}^0)-m_{y,i}^0\right\|_2^2$, $\Delta_{m_z}^0\triangleq\frac{1}{n}\sum_{i=1}^n\left\|\mathbb{E}(D_{z,i}^0)-m_{z,i}^0\right\|_2^2$, *and* $\Delta\triangleq\max\{\Delta_\Phi^0,\Delta_x^0,\Delta_h^0,\Delta_y^0,\Delta_z^0,\Delta_{m_y}^0,\Delta_{m_z}^0\}$.

Proof. By $L_{\nabla\Phi}$-smoothness of Φ (Lemma B.2), we have

$$\begin{aligned}
&\mathbb{E}\left[\Phi(x^{k+1})\right]\\
\le&\mathbb{E}\left[\Phi(x^k)\right]+\mathbb{E}\left[\langle\nabla\Phi(x^k),x^{k+1}-x^k\rangle\right]+\frac{L_{\nabla\Phi}}{2}\mathbb{E}\left[\left\|x^{k+1}-x^k\right\|_2^2\right]\\
=&\mathbb{E}\left[\Phi(x^k)\right]+\mathbb{E}\left[\left\langle\frac{1}{2}\hat{h}_x^k,x^{k+1}-x^k\right\rangle\right]+\mathbb{E}\left[\left\langle\nabla\Phi(x^k)-\frac{1}{2}\hat{h}_x^k,x^{k+1}-x^k\right\rangle\right]+\frac{L_{\nabla\Phi}}{2}\mathbb{E}\left[\left\|x^{k+1}-x^k\right\|_2^2\right]\\
=&\mathbb{E}\left[\Phi(x^k)\right]-\left(\frac{1}{2\alpha}-\frac{L_{\nabla\Phi}}{2}\right)\mathcal{X}_+^k+\frac{\alpha}{2}\mathbb{E}\left[\left\|\nabla\Phi(x^k)-\hat{h}_x^k\right\|_2^2\right]-\frac{\alpha}{2}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right].
\end{aligned}\tag{131}$$

Summing (131) from $k=0$ to $K-1$, we have

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right]\le\frac{2\Phi(x^0)}{\alpha}-\left(\frac{1}{\alpha^2}-\frac{L_{\nabla\Phi}}{\alpha}\right)\sum_{k=0}^{K-1}\mathcal{X}_+^k+\sum_{k=0}^{K}\mathbb{E}\left[\left\|\hat{h}_x^k-\nabla\Phi(x^k)\right\|_2^2\right]. \tag{132}$$

By the choice of α, we have

$$\frac{1}{\alpha^2}-\frac{L_{\nabla\Phi}}{\alpha}\ge\frac{1}{2\alpha^2}\ge\frac{1}{2}C_3^{-2},$$

thus by applying Lemma B.25 to (132) we obtain (130). □

C. Convergence Acceleration

In this section, we present the algorithmic design as well as convergence results of the two variants mentioned in Sec. 6, namely CM-SOBA-MSC and EF-SOBA-MSC.

C.1. Algorithm Design

In this subsection, we propose two variants of the proposed algorithms to help further improve the convergence rate. These variants are both based on the multi-step compression (MSC) technique (He et al., 2023a) as shown in Alg. 3.

Algorithm 3 MSC Module

Input: vector x, ω-unbiased compressor $\mathcal{C}$, communication rounds R;
Initialize $v^0 = 0$;
for $r = 1, \cdots, R$ **do**
  Send compressed vector $\mathcal{C}(x - v^{r-1})$ to the receiver;
  Update $v^r = v^{r-1} + (1+\omega)^{-1}\mathcal{C}(x - v^{r-1})$;
end for
Output: $\mathrm{MSC}(x; \mathcal{C}, R) \triangleq v^R / \left(1 - (\omega/(1+\omega))^R\right)$.

Note that MSC contains a loop of length R, we may also introduce an R-time sampling step to balance the computation and communication complexities. Specifically, we use the following steps instead of (6):

$$
\begin{aligned}
D_{x,i}^k &\triangleq \frac{1}{R}\sum_{r=1}^{R} \nabla_{xy}^2 G(x^k, y^k; \xi_i^{k,r}) z^k + \nabla_x F(x^k, y^k; \phi_i^{k,r}), \\
D_{y,i}^k &\triangleq \frac{1}{R}\sum_{r=1}^{R} \nabla_y G(x^k, y^k; \xi_i^{k,r}), \\
D_{z,i}^k &\triangleq \frac{1}{R}\sum_{r=1}^{R} \nabla_{yy}^2 G(x^k, y^k; \xi_i^{k,r}) z^k + \nabla_y F(x^k, y^k; \phi_i^{k,r}).
\end{aligned} \tag{133}
$$

Intuitively, by replacing $\mathcal{C}_i^\ell(\cdot), \mathcal{C}_i^u(\cdot)$ with $\mathrm{MSC}\left(\cdot; \mathcal{C}_i^\ell, R\right)$, $\mathrm{MSC}\left(\cdot; \mathcal{C}_i^u, R\right)$, (6) with (133), the outer loop of the double-loop variants are equivalent to the original algorithm except for reducing the gradient/Jacobian sampling variance σ^2 by R times, as well as reducing the compression variance ω_u, ω_ℓ to $\omega_u\left(\omega_u/(1+\omega_u)\right)^R$ and $\omega_\ell\left(\omega_\ell/(1+\omega_\ell)\right)^R$, see Lemma C.2. For convenience, we name the so-generated variants of CM-SOBA and EF-SOBA as CM-SOBA-MSC and EF-SOBA-MSC, respectively.

A detailed description of CM-SOBA-MSC and EF-SOBA-MSC is in Algorithms 4 and 5, respectively.

Algorithm 4 CM-SOBA-MSC Algorithm

Input: α, β, γ, θ, ρ, R, x^0, y^0, $z^0(\|z^0\|_2 \le \rho)$, h_x^0;
for $k = 0, 1, \cdots, K-1$ **do**
  on worker:
  Compute $D_{x,i}^k$, $D_{y,i}^k$, $D_{z,i}^k$ as in (133);
  Send $\mathrm{MSC}\left(D_{x,i}^k; \mathcal{C}_i^u, R\right)$, $\mathrm{MSC}\left(D_{y,i}^k; \mathcal{C}_i^\ell, R\right)$, and $\mathrm{MSC}\left(D_{z,i}^k; \mathcal{C}_i^\ell, R\right)$ to server;
  on server:
    $\hat{D}_x^k = \frac{1}{n}\sum_{i=1}^n \mathrm{MSC}\left(D_{x,i}^k; \mathcal{C}_i^u, R\right)$,
    $\hat{D}_y^k = \frac{1}{n}\sum_{i=1}^n \mathrm{MSC}\left(D_{y,i}^k; \mathcal{C}_i^\ell, R\right)$,
    $\hat{D}_z^k = \frac{1}{n}\sum_{i=1}^n \mathrm{MSC}\left(D_{z,i}^k; \mathcal{C}_i^\ell, R\right)$;
    $x^{k+1} = x^k - \alpha \cdot h_x^k$;
    $y^{k+1} = y^k - \beta \cdot \hat{D}_y^k$;
    $\tilde{z}^{k+1} = z^k - \gamma \cdot \hat{D}_z^k$;
    $z^{k+1} = \mathrm{Clip}(\tilde{z}^{k+1}, \rho)$;
    $h_x^{k+1} = (1-\theta)h_x^k + \theta \cdot \hat{D}_x^k$;
  Broadcast $x^{k+1}, y^{k+1}, z^{k+1}$ to all workers;
end for

Algorithm 5 EF-SOBA-MSC Algorithm

Input: α, β, γ, θ, ρ, δ_u, δ_ℓ, x^0, y^0, $z^0(\|z^0\|_2 \leq \rho)$, $\{m_{x,i}^0\}$, $\{m_{y,i}^0\}$, $\{m_{z,i}^0\}$, $\{h_{x,i}^0\}$, $\hat{h}_x^0 = \frac{1}{n}\sum_{i=1}^n m_{x,i}^0$, $m_y^0 = \frac{1}{n}\sum_{i=1}^n m_{y,i}^0$, $m_z^0 = \frac{1}{n}\sum_{i=1}^n m_{z,i}^0$;
for $k = 0, 1, \cdots, K-1$ **do**
  on worker:
  Compute $D_{x,i}^k$, $D_{y,i}^k$, $D_{z,i}^k$ as in (133);
  $h_{x,i}^{k+1} = (1-\theta)h_{x,i}^k + \theta \cdot D_x^k$;
  $m_{x,i}^{k+1} = m_{x,i}^k + \delta_u \cdot \mathrm{MSC}\left(h_{x,i}^{k+1} - m_{x,i}^k; \mathcal{C}_i^u, R\right)$,
  $m_{y,i}^{k+1} = m_{y,i}^k + \delta_\ell \cdot \mathrm{MSC}\left(D_{y,i}^k - m_{y,i}^k; \mathcal{C}_i^\ell, R\right)$,
  $m_{z,i}^{k+1} = m_{z,i}^k + \delta_\ell \cdot \mathrm{MSC}\left(D_{z,i}^k - m_{z,i}^k; \mathcal{C}_i^\ell, R\right)$;
  Send $\mathrm{MSC}\left(h_{x,i}^{k+1} - m_{x,i}^k; \mathcal{C}_i^u, R\right), \mathrm{MSC}\left(D_{y,i}^k - m_{y,i}^k; \mathcal{C}_i^\ell, R\right), \mathrm{MSC}\left(D_{z,i}^k - m_{z,i}^k; \mathcal{C}_i^\ell, R\right)$ to server;
  on server:
  $\hat{D}_y^k = m_y^k + \frac{1}{n}\sum_{i=1}^n \mathrm{MSC}\left(D_{y,i}^k - m_{y,i}^k; \mathcal{C}_i^\ell, R\right)$,
  $\hat{D}_z^k = m_z^k + \frac{1}{n}\sum_{i=1}^n \mathrm{MSC}\left(D_{z,i}^k - m_{z,i}^k; \mathcal{C}_i^\ell, R\right)$;
  $x^{k+1} = x^k - \alpha \cdot \hat{h}_x^k$;
  $y^{k+1} = y^k - \beta \cdot \hat{D}_y^k$;
  $\tilde{z}^{k+1} = z^k - \gamma \cdot \hat{D}_z^k$;
  $z^{k+1} = \mathrm{Clip}(\tilde{z}^{k+1}, \rho)$;
  $\hat{h}_x^{k+1} = \hat{h}_x^k + \frac{\delta_u}{n}\sum_{i=1}^n \mathrm{MSC}\left(h_{x,i}^{k+1} - m_{x,i}^k; \mathcal{C}_i^u, R\right)$;
  $m_y^{k+1} = m_y^k + \frac{\delta_\ell}{n}\sum_{i=1}^n \mathrm{MSC}\left(D_{y,i}^k - m_{y,i}^k; \mathcal{C}_i^\ell, R\right)$;
  $m_z^{k+1} = m_z^k + \frac{\delta_\ell}{n}\sum_{i=1}^n \mathrm{MSC}\left(D_{z,i}^k - m_{z,i}^k; \mathcal{C}_i^\ell, R\right)$;
  Broadcast $x^{k+1}, y^{k+1}, z^{k+1}$ to all workers;
end for

C.2. Convergence of CM-SOBA-MSC and EF-SOBA-MSC

Similar to Lemma B.3, we have the following lemma for the gradient accumulation mechanism.

Lemma C.1 (Reduced Variance)**.** *Under Assumption 2.3, we have the following variance bounds for CM-SOBA-MSC (Alg. 4) and EF-SOBA-MSC (Alg. 5):*

$$\begin{aligned}
&\mathrm{Var}\left[D_{y,i}^k \mid \mathcal{F}^k\right] \leq \tilde{\sigma}^2, \quad \mathrm{Var}\left[D_y^k \mid \mathcal{F}^k\right] \leq \tilde{\sigma}^2/n;\\
&\mathrm{Var}\left[D_{x,i}^k \mid \mathcal{F}^k\right] \leq \tilde{\sigma}_1^2, \quad \mathrm{Var}\left[D_x^k \mid \mathcal{F}^k\right] \leq \tilde{\sigma}_1^2/n;\\
&\mathrm{Var}\left[D_{z,i}^k \mid \mathcal{F}^k\right] \leq \tilde{\sigma}_1^2, \quad \mathrm{Var}\left[D_z^k \mid \mathcal{F}^k\right] \leq \tilde{\sigma}_1^2/n.
\end{aligned}$$

The following lemma describes the property of the MSC module (Alg. 3).

Lemma C.2 ((He et al., 2023a), Lemma 2)**.** *Assume $\mathcal{C}$ is an ω-unbiased compressor, and R is any positive integer. $\mathrm{MSC}(\cdot; C, R)$ is then an $\tilde{\omega}$-unbiased compressor with*

$$\tilde{\omega} = \omega\left(\frac{\omega}{1+\omega}\right)^R.$$

Consequently, $\mathrm{MSC}(\cdot; \mathcal{C}_i^u, R)$ is equivalent to an $\tilde{\omega}_u$-unbiased compressor. Similarly, $\mathrm{MSC}\left(\cdot; \mathcal{C}_i^\ell, R\right)$ is equivalent to an $\tilde{\omega}_\ell$-unbiased compressor. The following is a technical lemma.

Lemma C.3 ((He et al., 2023a), Lemma 11)**.** *For $R \geq 4(1+\omega)\ln(4(1+\omega))$, it holds that*

$$R\left(\frac{\omega}{1+\omega}\right)^R \leq \left(\frac{\omega}{1+\omega}\right)^{R/2}.$$

We have the following convergence result for CM-SOBA-MSC (Alg. 4).

Theorem C.4 (Convergence of CM-SOBA-MSC). *Under Assumptions 2.1, 2.2, 2.3, 2.4, 4.1 and assuming* $\beta < \min\left\{\frac{2}{\mu_g+L_g}, \frac{\mu_g n}{8\tilde{\omega}_\ell L_g^2}\right\}$, $\gamma \le \min\left\{\frac{1}{L_g}, \frac{\mu_g n}{36\tilde{\omega}_\ell L_g^2}\right\}$, $\theta \le \min\left\{1, \frac{n}{12\tilde{\omega}_u}\right\}$, $\rho \ge C_f/\mu_g$, $\alpha \le \min\{\frac{1}{2L_{\nabla\Phi}}, C_2\}$ *with*

$$C_2^{-2} \triangleq 2\cdot\left(\frac{2L_{\nabla\Phi}^2}{\theta^2} + \frac{26(L_2^2+25\kappa_g^2L_1^2)L_{y^\star}^2}{\beta^2\mu_g^2} + \frac{78L_g^2L_{z^\star}^2}{\gamma^2\mu_g^2}\right),$$

CM-SOBA-MSC (Alg. 4) converges as

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right]\\
\le&\frac{4\Delta_\Phi^0}{K\alpha} + \frac{2\Delta_x^0}{K\theta} + \frac{26(L_2^2+25\kappa_g^2L_1^2)\Delta_y^0}{\mu_g K\beta} + \frac{52L_g^2\Delta_z^0}{\mu_g K\gamma} + \frac{52(1+\tilde{\omega}_\ell)(L_2^2+25\kappa_g^2L_1^2)\beta}{\mu_g n}\cdot\tilde{\sigma}^2\\
&+\left(\frac{2(1+\tilde{\omega}_u)\theta}{n} + \frac{78L_g^2(1+\tilde{\omega}_\ell)\gamma}{\mu_g n}\right)\cdot\tilde{\sigma}_1^2 + \left(\frac{12\rho^2\tilde{\omega}_u\theta}{n} + \frac{104(L_2^2+25\kappa_g^2L_1^2)\tilde{\omega}_\ell\beta}{n\mu_g} + \frac{312L_g^2\rho^2\tilde{\omega}_\ell\gamma}{n\mu_g}\right)\cdot b_g^2\\
&+\left(\frac{12\tilde{\omega}_u\theta}{n} + \frac{312L_g^2\tilde{\omega}_\ell\gamma}{n\mu_g}\right)\cdot b_f^2. \qquad (134)
\end{aligned}$$

If we further choose parameters as

$$\begin{aligned}
R =& \left\lceil 4(1+\omega_u+\omega_\ell)\ln\left(4(1+\omega_u+\omega_\ell) + \frac{(1+\omega_u+\omega_\ell)^2(b_f^2+b_g^2)^2}{\sigma^4}\right)\right\rceil,\\
\alpha =& \frac{1}{2L_{\nabla\Phi}+C_2^{-1}},\\
\beta =& \left(\frac{\mu_g+L_g}{2} + \frac{8\tilde{\omega}_\ell L_g^2}{\mu_g n} + \sqrt{\frac{2K\left((1+\tilde{\omega}_\ell)\tilde{\sigma}^2+2\tilde{\omega}_\ell b_g^2\right)}{n\Delta_y^0}}\right)^{-1},\\
\gamma =& \left(L_g + \frac{36\tilde{\omega}_\ell L_g^2}{\mu_g n} + \sqrt{\frac{3K\left((1+\tilde{\omega}_\ell)\tilde{\sigma}_1^2+4\tilde{\omega}_\ell b_f^2+4\tilde{\omega}_\ell(C_f^2/\mu_g^2)b_g^2\right)}{2n\Delta_z^0}}\right)^{-1},\\
\theta =& \left(1 + \frac{12\tilde{\omega}_u}{n} + \sqrt{\frac{K\left((1+\tilde{\omega}_u)\tilde{\sigma}_1^2+6\tilde{\omega}_u b_f^2+6\tilde{\omega}_u(C_f^2/\mu_g^2)b_g^2\right)}{n\Delta_x^0}}\right)^{-1},\\
\rho =& \frac{C_f}{\mu_g},
\end{aligned}$$

CM-SOBA-MSC (Alg. 4) converges as order

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right] = \mathcal{O}\left(\frac{\sqrt{\Delta}\sigma}{\sqrt{nT}} + \frac{(1+\omega_u+\omega_\ell)\Delta\tilde{\Theta}(1)}{T}\right),$$

where $T \triangleq KR$ *is the total number of iterations of CM-SOBA-MSC (Alg. 4),* $\tilde{\Theta}$ *hides logarithmic terms independent of* T*, and* Δ *is as defined in Theorem B.14.*

Proof. By Lemma C.1 and C.2, the outer loop of CM-SOBA-MSC (Alg. 4) is equivalent to CM-SOBA (Alg. 1), except for using gradient/Jacobian oracles and unbiased compressors with variance reduced by a factor of R. Thus, (134) is a direct corollary of Theorem B.14. By applying Lemma C.3, the choice of R implies

$$\tilde{\omega}_u \le 1, \quad \tilde{\omega}_\ell \le 1, \quad R\tilde{\omega}_u \le \frac{\sigma^2}{b_f^2+b_g^2}, \quad R\tilde{\omega}_\ell \le \frac{\sigma^2}{b_f^2+b_g^2}.$$

Consequently, by the choice of α, β, γ, θ, ρ and R, CM-SOBA-MSC (Alg. 4) converges as

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right] =&\mathcal{O}\left(\frac{\sqrt{(1+\tilde{\omega}_u+\tilde{\omega}_\ell)\Delta}\sigma+\sqrt{R(\tilde{\omega}_u+\tilde{\omega}_\ell)\Delta}(b_f+b_g)}{\sqrt{nT}}+\frac{(1+\tilde{\omega}_u/n+\tilde{\omega}_\ell/n)\Delta R}{T}\right)\\ =&\mathcal{O}\left(\frac{\sqrt{\Delta}\sigma}{\sqrt{nT}}+\frac{(1+\omega_u+\omega_\ell)\Delta\tilde{\Theta}(1)}{T}\right).\end{aligned}$$

□

For EF-SOBA-MSC (Alg. 5), we have the following notations for convenience:

$$\tilde{\omega}_1 \triangleq 1+6\tilde{\omega}_\ell(1+\tilde{\omega}_\ell),\quad \tilde{\omega}_2 \triangleq 1+36\tilde{\omega}_u(1+\tilde{\omega}_u).$$

Now we are ready to prove Theorem 6.1. The detailed result is as follows.

Theorem C.5 (Convergence of EF-SOBA-MSC)**.** *Under assumptions 2.1, 2.2, 2.3, 2.4 and assuming* $\rho \geq C_f/\mu_g$, $\beta < \min\left\{\frac{2}{\mu_g+L_g}, \frac{\mu_g n}{96\tilde{\omega}_\ell\tilde{\omega}_1}L_g^2\right\}$, $\gamma \leq \min\left\{\frac{1}{L_g}, \frac{\mu_g n}{432\tilde{\omega}_\ell\tilde{\omega}_1 L_g^2}\right\}$, $\delta_\ell = (1+\tilde{\omega}_\ell)^{-1}$, $\delta_u = (1+\tilde{\omega}_u)^{-1}$, $m_{x,i}^0 = h_{x,i}^0$ *for* $i=1,\cdots,n$, $\alpha \leq \min\left\{\frac{1}{2L_{\nabla\Phi}}, C_4\right\}$ *with*

$$\begin{aligned}C_4^{-2} \triangleq& 2\cdot\left[\frac{4L_{\nabla\Phi}^2}{\theta^2}+72\tilde{\omega}_u(1+\tilde{\omega}_u)L_5^2+12\tilde{\omega}_2(L_2^2+25\kappa_g^2L_1^2)\left(\frac{4L_{y^\star}^2}{\beta^2\mu_g^2}+\frac{16\tilde{\omega}_\ell\tilde{\omega}_1L_3^2\beta}{\mu_g n}\right)\right.\\ &\left.+12\tilde{\omega}_2L_g^2\left(\frac{12L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{72\tilde{\omega}_\ell\tilde{\omega}_1L_4^2\gamma}{\mu_g n}\right)\right],\end{aligned}$$

EF-SOBA-MSC (Alg. 5) converges as

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right] \leq&\frac{2\Delta_\Phi^0}{K\alpha}+\frac{2\Delta_x^0}{K\theta}+\frac{12\tilde{\omega}_u(1+\tilde{\omega}_u)\theta\Delta_h^0}{K}+\frac{24\tilde{\omega}_2(L_2^2+25\kappa_g^2L_1^2)\Delta_y^0}{\mu_gK\beta}\\ &+\frac{48\tilde{\omega}_2L_g^2\Delta_z^0}{\mu_gK\gamma}+\frac{48\tilde{\omega}_2(1+6\tilde{\omega}_\ell\tilde{\omega}_1)(L_2^2+25\kappa_g^2L_1^2)\beta}{\mu_g n}\cdot\tilde{\sigma}^2\\ &+\left(\frac{2\theta}{n}+36\tilde{\omega}_u(1+\tilde{\omega}_u)\theta^2+\frac{72\tilde{\omega}_2(1+6\tilde{\omega}_\ell\tilde{\omega}_1)L_g^2\gamma}{\mu_g n}\right)\cdot\tilde{\sigma}_1^2\\ &+\frac{48\tilde{\omega}_2\tilde{\omega}_\ell(1+4\tilde{\omega}_\ell)(L_2^2+25\kappa_g^2L_1^2)\beta}{\mu_gKn^2}\sum_{i=1}^n\left\|\mathbb{E}(D_{y,i}^0)-m_{y,i}^0\right\|_2^2\\ &+\frac{72\tilde{\omega}_2\tilde{\omega}_\ell(1+4\tilde{\omega}_\ell)L_g^2\gamma}{\mu_gKn^2}\sum_{i=1}^n\left\|\mathbb{E}(D_{z,i}^0)-m_{z,i}^0\right\|_2^2.\end{aligned}\tag{135}$$

If we further choose parameters as

$$\begin{aligned}&R=\left\lceil 4(1+\omega_u+\omega_\ell)\ln\left(4(1+\omega_u+\omega_\ell)+\frac{(1+\omega_u)^2\Delta n^3}{\sigma^2}\right)\right\rceil,\\ &\alpha=\frac{1}{2L_{\nabla\Phi}+C_4^{-1}},\quad \rho=C_f/\mu_g,\quad \delta_\ell=\frac{1}{1+\tilde{\omega}_\ell},\quad \delta_u=\frac{1}{1+\tilde{\omega}_u},\\ &\beta=\left(\frac{\mu_g+L_g}{2}+\frac{96\tilde{\omega}_\ell\tilde{\omega}_1L_g^2}{\mu_g n}+\sqrt{\frac{2\tilde{\omega}_\ell(1+4\tilde{\omega}_\ell)\Delta_{m_y}^0}{n\Delta_y^0}}+\sqrt{\frac{2K\tilde{\sigma}^2(1+6\tilde{\omega}_\ell\tilde{\omega}_1)}{n\Delta_y^0}}\right)^{-1},\\ &\gamma=\left(L_g+\frac{432\tilde{\omega}_\ell\tilde{\omega}_1L_g^2}{\mu_g n}+\sqrt{\frac{3\tilde{\omega}_\ell(1+4\tilde{\omega}_\ell)\Delta_{m_z}^0}{2n\Delta_z^0}}+\sqrt{\frac{K\tilde{\sigma}_1^2(1+6\tilde{\omega}_\ell\tilde{\omega}_1)}{n\Delta_y^0}}\right)^{-1},\end{aligned}$$

$$\theta = \left(1 + \sqrt{\frac{6\tilde{\omega}_u(1+\tilde{\omega}_u)\Delta_h^0}{\Delta_x^0}} + \sqrt{\frac{K\tilde{\sigma}_1^2}{n\Delta_x^0}} + \sqrt[3]{\frac{18\tilde{\omega}_u(1+\tilde{\omega}_u)K\tilde{\sigma}_1^2}{\Delta_x^0}}\right)^{-1},$$

EF-SOBA-MSC (Alg. 5) converges as order

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right] = \mathcal{O}\left(\frac{\sqrt{\Delta}\sigma}{\sqrt{nT}} + \frac{(1+\omega_u+\omega_\ell)\Delta\tilde{\Theta}(1)}{T}\right),$$

where $T : KR$ is the total number of iterations of EF-SOBA-MSC (Alg. 5), $\tilde{\Theta}$ hides logarithmic terms independent of T, and Δ is as defined in Theorem B.26.

Proof. By Lemma C.1 and C.2, the outer loop of EF-SOBA-MSC (Alg. 5) is equivalent to EF-SOBA (Alg. 2), except for using gradient/Jacobian oracles and unbiased compressors with variance reduced by a factor of R. Thus, (135) is a direct corollary of Theorem B.26. By applying Lemma C.3, the choice of R implies

$$\tilde{\omega}_u \le 1, \quad \tilde{\omega}_\ell \le 1, \quad R\tilde{\omega}_u \le \frac{\sigma}{\sqrt{\Delta}n^{3/2}}.$$

Consequently, by the choice of δ_u, δ_ℓ, α, β, γ, θ, ρ and R, EF-SOBA-MSC (Alg. 5) converges as

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(x^k)\right\|_2^2\right] =& \mathcal{O}\left(\frac{(1+\tilde{\omega}_u)^2(1+\tilde{\omega}_\ell)^{3/2}\sqrt{\Delta}\sigma}{\sqrt{nT}} + \frac{(R\tilde{\omega}_u)^{1/3}(1+\tilde{\omega}_u)^{1/3}\Delta^{2/3}\sigma^{2/3}}{T^{2/3}} + \frac{(1+\tilde{\omega}_u)\Delta R}{T}\right.\\ &\left. + \frac{(1+\tilde{\omega}_u)^2\sqrt{\tilde{\omega}_\ell(1+\tilde{\omega}_\ell)}\Delta R}{\sqrt{n}T} + \frac{(1+\tilde{\omega}_u)^2\tilde{\omega}_\ell^3\Delta R}{nT}\right).\\ =& \mathcal{O}\left(\frac{\sqrt{\Delta}\sigma}{\sqrt{nT}} + \frac{(1+\omega_u+\omega_\ell)\Delta\tilde{\Theta}(1)}{T}\right).\end{aligned}$$

□

D. Experimental Specifications

D.1. Hyper-Representation

Problem formulation. Following (Franceschi et al., 2018), the hyper-representation problem can be formulated as:

$$\begin{aligned}&\min_{\lambda}\ L(\lambda) = \frac{1}{|\mathcal{D}_v|}\sum_{\xi\in\mathcal{D}_v} L(\omega^*(\lambda), \lambda; \xi)\\ &s.t.\ \omega^*(\lambda) = \arg\min_{\omega}\frac{1}{|\mathcal{D}_\tau|}\sum_{\eta\in\mathcal{D}_\tau} L(\omega, \lambda; \eta)\end{aligned}$$

where L stands for the cross entropy loss here, $\mathcal{D}_v$ and $\mathcal{D}_\tau$ denote the validation set and training set, respectively. Hyper-representation consists of two nested problems, where the upper-level optimizes the intermediate representation parameter λ to obtain better feature representation on validation data, and the lower-level optimizes the weights ω of the downstream tasks on training data.

Datasets and model architecture. For MNIST, we use a 2-layer multilayer perceptron (MLP) with 200 hidden units. Therefore, the upper problem optimizes the hidden layer with 157,000 parameters, and the lower problem optimizes the output layer with 2,010 parameters. For CIFAR-10, we train the 7-layer LeNet (Lecun et al., 1998), where we treat the last fully connected layer's parameters as lower-level variables and the rest layers' parameters as upper-level variables.

Hyperparameter settings. According to the optimal relation shown in (10), we set the compression parameter $K = 200$ for lower-level and $K = 2000$ for upper-level. The dataset is partitioned to 10 workers both under homogeneous and heterogeneous distributions. The batch size of workers' stochastic oracle is 512 for MNIST and 1000 for CIFAR-10. The moving average parameter θ of CM-SOBA and EF-SOBA is 0.1. We optimize the stepsizes for all compared algorithms via grid search, each ranging from $[0.001, 0.05, \ldots, 0.5]$, which is summarized in Table 2.

Table 2. Stepsize selection for experiments of hyper-representation

Algorithm	Dataset	Stepsize $[\alpha, \beta, \gamma]$
NC-SOBA	MNIST	[0.5, 0.1, 0.01]
C-SOBA	MNIST	[0.1, 0.1, 0.01]
CM-SOBA	MNIST	[0.1, 0.1, 0.01]
EF-SOBA	MNIST	[0.5, 0.1, 0.01]
NC-SOBA	CIFAR-10	[0.1, 0.001, 0.001]
C-SOBA	CIFAR-10	[0.05, 0.001, 0.001]

D.2. Hyperparameter Optimization

Problem formulation. Hyperparameter optimization can be formulated as:

$$\min_{\lambda} \ L(\lambda) = \frac{1}{|\mathcal{D}_v|} \sum_{\xi \in \mathcal{D}_v} L(\omega^*(\lambda); \xi)$$
$$s.t. \ \omega^*(\lambda) = \arg\min_{\omega} \frac{1}{|\mathcal{D}_\tau|} \sum_{\eta \in \mathcal{D}_\tau} (L(\omega; \eta) + R(\omega, \lambda))$$

where L is the loss function, $R(\omega, \lambda)$ is a regularizer, $\mathcal{D}_v$ and $\mathcal{D}_\tau$ denote the validation set and training set. To perform logistic regression with regularization, following (Pedregosa, 2016; Grazzi et al., 2020; Chen et al., 2022), we define $L = \log(1 + e^{-yx^T\omega})$ and $R = \frac{1}{2}\sum_{i=1}^{p} e^{\lambda_i}\omega_i^2$ on synthetic dataset. For MNIST, we have the model parameter $\omega \in \mathbb{R}^{p\times c}$ with $p = 784$ and $c = 10$. Following (Grazzi et al., 2020), we set L as the cross entropy loss and $R = \frac{1}{cp}\sum_{j=1}^{c}\sum_{i=1}^{p} e^{\lambda_i}\omega_{ij}^2$.

Datasets. We construct synthetic heterogeneous data by a linear model $y = \text{sign}(x^T\omega + \epsilon \cdot z)$, where $\epsilon = 0.1$ is the noise rate and $z \in \mathbb{R}$ is the noise vector sampled from standard normal distribution. The distribution of $x \in \mathbb{R}^{100}$ on worker i is $N(0, i^2)$ if $i\%2 = 0$ otherwise $\chi^2(i)$. Additionally, we assume there are 5 workers with 500 training data and 500 validation data respectively. For MNIST, We partition it to 10 workers under both homogeneous and heterogeneous data distributions.

Hyperparameter settings. For the experiments in this study, where the upper and lower levels share the same compressed dimension, we use a uniform compression parameter for both levels: $K = 10$ for the MNIST dataset and $K = 20$ for the synthetic dataset. The batch size for the synthetic dataset is 50 and for MNIST is 512. We optimize the stepsizes for all compared algorithms via grid search, each ranging from $[0.001, 0.05, \ldots, 0.5]$, which is summarized in Table 3.

Table 3. Stepsize selection for experiments of hyperparameter optimization

Algorithm	Dataset	Stepsize $[\alpha, \beta, \gamma]$
NC-SOBA	Synthetic	[0.1, 0.01, 0.001]
C-SOBA	Synthetic	[0.05, 0.01, 0.001]
CM-SOBA	Synthetic	[0.05, 0.01, 0.001]
EF-SOBA	Synthetic	[0.1, 0.01, 0.001]
NC-SOBA	MNIST	[0.1, 0.1, 0.1]
C-SOBA	MNIST	[0.1, 0.1, 0.1]
CM-SOBA	MNIST	[0.1, 0.1, 0.1]
EF-SOBA	MNIST	[0.1, 0.1, 0.1]

Additional results on synthetic data. It can be seen from Figure 5 that under the heterogeneous data distribution, our proposed EF-SOBA outperforms with a similar convergence rate and much fewer communication bits. These results are also consistent with those on MNIST in Figure 3, which implies the broad application of our algorithms on various datasets and problem setups.

D.3. Additional Results

Results on CIFAR-10. Under the experimental setup in Appendix D.1, we evaluate the performance of compressed algorithms on CIFAR-10 under homogeneous data distributions. We only draw the result of C-SOBA here to compare with NC-SOBA because from results on MNIST we can see other algorithms perform worse than C-SOBA under the homogeneous data distributions. From Figure 6, it can be seen that with nearly $10\times$ communication bits savings, our compressed algorithm converges to the same test accuracy as non-compressed algorithm. It validates the effectiveness of our

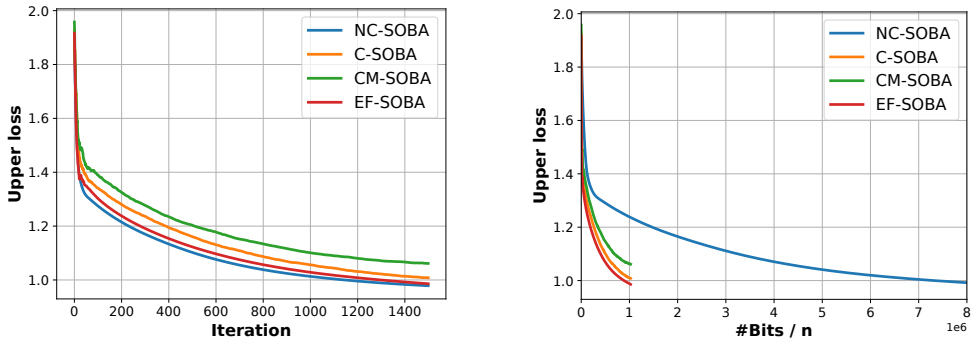

Figure 5. Hyperparameter optimization on synthetic heterogeneous data.

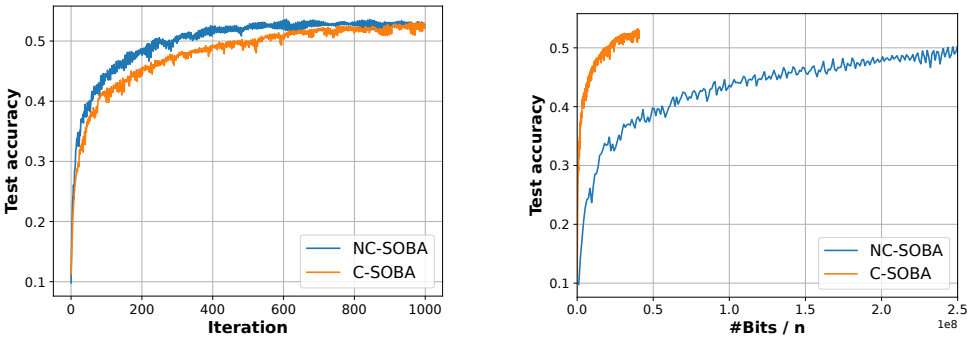

Figure 6. Hyper-representation on CIFAR-10 under homogeneous data distributions.

proposed algorithms even under complicated model architecture and large dataset. Notice that the backbone test accuracy is not satisfactory here, we suspect that it's because the bilevel structure of the hyper-representation problem brings challenges to the training.

More comparison baselines. In our study, we evaluate our compression algorithms for distributed stochastic bilevel optimization against the SOBA algorithm, which serves as our non-compression baseline (referred to as NC-SOBA). Furthermore, we include FedNest(Tarzanagh et al., 2022) as another non-compression baseline for comparison with SOBA. We conduct hyper-representation experiments on the MNIST dataset, employing an MLP backbone and utilizing homogeneous data distributions. We implement FedNest based on its publicly available source code. As illustrated in Fig. 7, it is evident that NC-SOBA achieves faster convergence in terms of communication bits compared to FedNest.

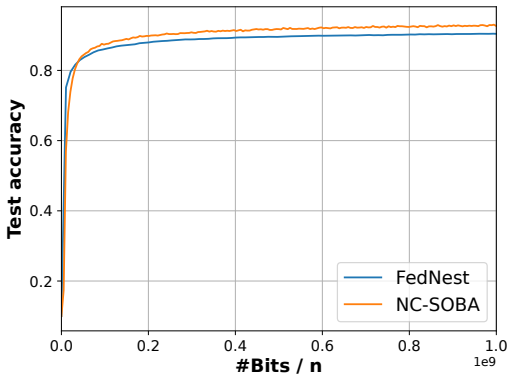

Figure 7. Hyper-representation on MNIST under homogeneous data distributions.

Tuning the momentum parameter in CM-SOBA. In Fig. 2, CM-SOBA performs inferiorly to C-SOBA due to the momentum parameter θ being set to a fixed value of 0.1, without further optimization, which may lead to sub-optimal results. To mitigate this limitation, we conducted additional experiments employing a refined approach for selecting the momentum parameter. The results are depicted in Fig. 8, illustrating that with an appropriately tuned momentum parameter, CM-SOBA indeed outperforms C-SOBA. This underscores the significance of momentum parameter optimization for the effective implementation of CM-SOBA.

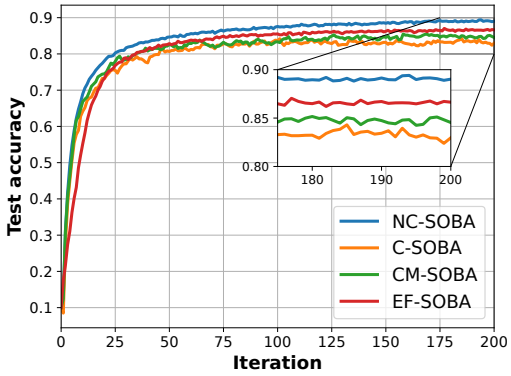

Figure 8. Hyper-representation on MNIST under heterogeneous data distributions.

D.4. Ablation Studies

Ablation on MSC rounds. We propose algorithm variants in Sec. 6 to enhance theoretical convergence rate, by utilizing the multi-step compression and gradient accumulation mechanism. It's worth noting that when CM-SOBA (Alg. 1) is a special case of CM-SOBA-MSC (Alg. 4) with $R = 1$. Same thing happens to EF-SOBA (Alg. 2) and EF-SOBA-MSC (Alg. 5). Thus a natural question is, how should we select R in practice, and whether $R > 1$ can be more effective than $R = 1$? To address this issue, we conduct ablation experiments on the hyperparameter optimization task on MNIST dataset. The problem formulation, data and stepsizes are set consistent with Appendix D.2.

Figure 9 displays the loss curve of CM-SOBA-MSC (left) and EF-SOBA-MSC (right) with different R's in the hyperparameter optimization task on MNIST under heterogeneous data distributions. With $R = 2$, both algorithms perform better than those with $R = 1, 4, 5$. We demonstrate that when R is too small, the gradient bias induced by compression error and sampling randomness slows down the convergence, while a much larger R trades communication/computation savings to update directions with little improvement, making it less effective. Generally speaking, there is a trade-off in the selection of R, and we recommend choosing suitable R's by cross validation.

One can also observe from Figure 9 that EF-SOBA-MSC with $R = 1$ (which is exactly EF-SOBA) has a worse performance than CM-SOBA-MSC with $R = 1$ (which is exactly CM-SOBA), even if the data is constructed heterogeneously. This

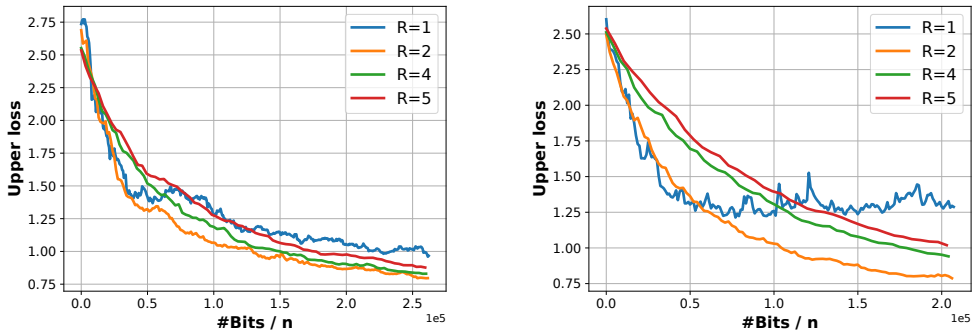

Figure 9. Ablation on MSC rounds R for CM-SOBA-MSC (left) and EF-SOBA-MSC (right), conducted on hyperparameter optimization task on MNIST heterogeneous data.

phenomenon is consistent with our convergence results, that EF-SOBA is more susceptible to large ω's than CM-SOBA. Consequently, we recommend using EF-SOBA-MSC with $R > 1$ when severely aggressive compressors are applied.

Table 4. Compressor choices under different strategies.

Strategy	K for lower-level rand-K	K for upper-level rand-K	Communicated entries per iter
$\sqrt{d_x} : \sqrt{d_y}$	6	68	80
$1:1$	1	78	80

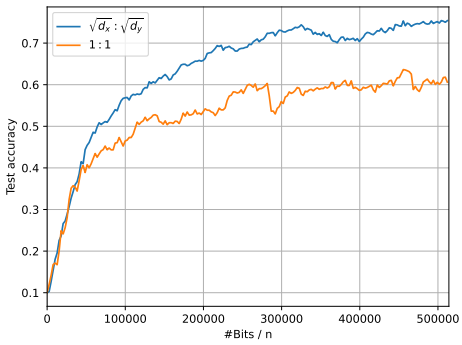

Figure 10. Ablation on compressor choices conducted on hyper-presentation optimization task on MNIST heterogeneous data.

Ablation on compressor choices. We evaluate various compressor choices while maintaining consistent per-round communication cost constraints. In Fig. 10, we present the performance comparison of C-SOBA on the hyper-presentation problem using the MNIST dataset, where $d_x = 157000$ and $d_y = 2010$. All experiments employ identical learning rates: $\alpha =$ 2e-2, $\beta =$ 8e-3, and $\gamma =$ 8e-4. The Rand-K compressors, outlined in Table 4, are selected based on two strategies:

- Strategy 1: $\frac{\omega_u}{\omega_l} \approx \sqrt{\frac{d_x}{2d_y}} = \Theta\left(\sqrt{\frac{d_x}{d_y}}\right)$;
- Strategy 2: $\frac{\omega_u}{\omega_l} \approx \frac{1}{2} = \Theta\left(\frac{1}{1}\right)$.

It is evident that Strategy 1 (as recommended in (10)) outperforms Strategy 2.