Decentralized Stochastic Bilevel Optimization with Improved Per-Iteration Complexity

Xuxing Chen* Minhui Huang† Shiqian Ma‡ Krishnakumar Balasubramanian§

October 23, 2022

Abstract

Bilevel optimization recently has received tremendous attention due to its great success in solving important machine learning problems like meta learning, reinforcement learning, and hyperparameter optimization. Extending single-agent training on bilevel problems to the decentralized setting is a natural generalization, and there has been a flurry of work studying decentralized bilevel optimization algorithms. However, it remains unknown how to design the distributed algorithm with sample complexity and convergence rate comparable to SGD for stochastic optimization, and at the same time without directly computing the exact Hessian or Jacobian matrices. In this paper we propose such an algorithm. More specifically, we propose a novel decentralized stochastic bilevel optimization (DSBO) algorithm that only requires first order stochastic oracle, Hessian-vector product and Jacobian-vector product oracle. The sample complexity of our algorithm matches the currently best known results for DSBO, and the advantage of our algorithm is that it does not require estimating the full Hessian and Jacobian matrices, thereby having improved per-iteration complexity.

1 Introduction

Many machines learning problems can be formulated as a bilevel optimization problem of the form,

$$\begin{aligned} \min_{x\in\mathbb{R}^p} \quad & \Phi(x) = f(x, y^*(x)) \\ \text{s.t.} \quad & y^*(x) = \arg\min_{y\in\mathbb{R}^q} g(x, y), \end{aligned} \tag{1}$$

where we minimize the upper level function f with respect to x subject to the constraint that $y^*(x)$ is the minimizer of the lower level function. Its applications can range from classical optimization problems like compositional optimization (Chen et al., 2021) to modern machine learning problems such as reinforcement learning (Hong et al., 2020), meta learning (Ji et al., 2020; Snell et al., 2017; Bertinetto et al., 2018; Rajeswaran et al., 2019), hyperparameter optimization (Pedregosa, 2016; Franceschi et al., 2018), etc. State-of-the-art bilevel optimization algorithms with non-asymptotic analyses include BSA (Ghadimi and Wang, 2018), TTSA (Hong et al., 2020), StocBiO (Ji et al., 2020), ALSET (Chen et al., 2021), to name a few.

*Department of Mathematics, University of California, Davis. xuxchen@ucdavis.edu

†Department of Electrical and Computer Engineering, University of California, Davis. mhhuang@ucdavis.edu

‡Department of Computational Applied Mathematics and Operations Research, Rice University, and Department of Mathematics, University of California, Davis (on leave). sqma@rice.edu

§Department of Statistics, University of California, Davis. kbala@ucdavis.edu.

Decentralized bilevel optimization aims at solving bilevel problems in a decentralized setting, which provides additional benefits such as faster convergence, data privacy preservation and robustness to low network bandwidth compared to the centralized setting and the single-agent training (Lian et al., 2017). For example, decentralized meta learning, which is a special case of decentralized bilevel optimization, arise naturally in the context of medical data analysis in the context of protecting patient privacy; see, for example, Zhang et al. (2019); Altae-Tran et al. (2017); Kayaalp et al. (2022). Motivated by such applications, the works of Lu et al. (2022); Chen et al. (2022b); Yang et al. (2022); Gao et al. (2022) proposed and analyzed various decentralized stochastic bilevel optimization (DSBO) algorithms.

From a mathematical perspective, DSBO aims at solving the following problem in a distributed setting:

$$\begin{aligned} \min_{x\in\mathbb{R}^p} \quad & \Phi(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x, y^*(x)) \\ \text{s.t.} \quad & y^*(x) = \arg\min_{y\in\mathbb{R}^q} g(x,y) := \frac{1}{n}\sum_{i=1}^{n} g_i(x,y), \end{aligned} \tag{2}$$

where $x \in \mathbb{R}^p, y \in \mathbb{R}^q$. f_i is possibly nonconvex and g_i is strongly convex in y. Here n denotes the number of agents, and agent i only has access to stochastic oracles of f_i, g_i. The local objectives f_i and g_i are defined as:

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

$\mathcal{D}_{f_i}$ and $\mathcal{D}_{g_i}$ represent the data distributions used to generate the objectives for agent i, and each agent only has access to f_i and g_i. In practice we can replace the expectation by empirical loss, and then use samples to approximate the gradients in the updates. Existing works on DSBO require computing the full Hessian (or Jacobian) matrices in the hypergradient estimation, whose per-iteration complexity is $\mathcal{O}(q^2)$ (or $\mathcal{O}(pq)$). In problems like hyperparameter estimation, the lower level corresponds to learning the parameters of a model. When considering modern overparametrized models, the order of q is hence extremely large. Hence, to reduce the per-iteration complexity, it is of great interest to have each iteration based only on Hessian-vector (or Jacobian-vector) products, whose complexity is $\mathcal{O}(q)$ (or $\mathcal{O}(p)$); see, for example, Pearlmutter (1994).

1.1 Our contributions

Our contributions in this work are as follows.

- In Section 3.2, we propose a novel method to estimate the global hypergradient. Our method estimates the product of the inverse of the Hessian and vectors directly, without computing the full Hessian or Jacobian matrices, and thus improves the previous overall complexity on hypergradient estimation from $\mathcal{O}(Nq^2)$ to $\mathcal{O}(Nq)$, where N is the total steps of the hypergradient estimation subroutine.
- We design a DSBO algorithm (see Algorithm 3), and in Theorem 3.3 and Corollary 3.1 we show the sample complexity is of order $\mathcal{O}(\epsilon^{-2}\log\frac{1}{\epsilon})$, which matches the currently best known results of the single-agent bilevel optimization (Chen et al., 2021). Our proof relies on weaker assumptions comparing to Yang et al. (2022), and is based on carefully combining moving average stochastic gradient estimation analyses with the decentralized bilevel algorithm analyses.
- We conduct experiments on several machine learning problems. Our numerical results show the efficiency of our algorithm in both the synthetic and the real-world problems. Moreover,

since our algorithm does not store the full Hessian or Jacobian matrices, both the space complexity and the communication complexity are improved comparing to Chen et al. (2022b); Yang et al. (2022).

1.2 Related work

Bilevel optimization. Different from classical constrained optimization, bilevel optimization restricts certain variables to be the minimizer of the lower level function, which is more applicable in modern machine learning problems like meta learning (Snell et al., 2017; Bertinetto et al., 2018; Rajeswaran et al., 2019) and hyperparameter optimization (Pedregosa, 2016; Franceschi et al., 2018). In recent years, Ghadimi and Wang (2018) gave the first non-asymptotic analysis of the bilevel stochastic approximation methods, which attracted much attention to study more efficient bilevel optimization algorithms including AID-based (Domke, 2012; Pedregosa, 2016; Gould et al., 2016; Ghadimi and Wang, 2018; Grazzi et al., 2020; Ji et al., 2021), ITD-based (Domke, 2012; Maclaurin et al., 2015; Franceschi et al., 2018; Grazzi et al., 2020; Ji et al., 2021), and Neumann series-based (Chen et al., 2021; Hong et al., 2020; Ji et al., 2021) methods. These methods only require access to first order stochastic oracles and matrix-vector product (Hessian-vector and Jacobian-vector) oracles, which demonstrate great potential in solving bilevel optimization problems and achieve $\mathcal{O}(\epsilon^{-2}\log(\frac{1}{\epsilon}))$ sample complexity (Chen et al., 2021) that matches the result of SGD for single level stochastic optimization ignoring the log factors. Moreover, under stronger assumptions and variance reduction techniques, better complexity bounds are obtained (Guo et al., 2021; Khanduri et al., 2021; Yang et al., 2021; Chen et al., 2022a).

Decentralized optimization. Extending optimization algorithms from a single-agent setting to a multi-agent setting has been studied extensively in recent years thanks to the modern parallel computing. Decentralized optimization, which does not require a central node, serves as an important part of distributed optimization. Because of data heterogeneity and the absence of a central node, decentralized optimization is more challenging and each node communicates with neighbors to exchange information and solve a finite-sum optimization problem. Under certain scenarios, decentralized algorithms are more preferable comparing to centralized ones since the former preserve data privacy (Ram et al., 2009; Yan et al., 2012; Wu et al., 2017; Koloskova et al., 2020) and have been proved useful when the network bandwidth is low (Lian et al., 2017).

Decentralized stochastic bilevel optimization. To make bilevel optimization applicable in parallel computing, recent work started to focus on distributed bilevel optimization. FEDNEST (Tarzanagh et al., 2022) and FedBiO (Li et al., 2022) impose federated learning, which is essentially a centralized setting, on bilevel optimization. There are also some existing works (Lu et al., 2022; Chen et al., 2022b; Yang et al., 2022; Gao et al., 2022) on DSBO. They can be classified to two categories: global DSBO and personalized DSBO. The problem (2) that we consider in this paper is a global DSBO. The personalized DSBO (Lu et al., 2022; Gao et al., 2022) replaces $y^*(x)$ by the local one $y_i^*(x) = \arg\min_{y\in\mathbb{R}^q} g_i(x,y)$ in (2):

$$\begin{aligned} \min_{x\in\mathbb{R}^p} \quad & \Phi(x) = \frac{1}{n}\sum_{i=1}^n f_i(x, y_i^*(x)) \\ \text{s.t.} \quad & y_i^*(x) = \arg\min_{y\in\mathbb{R}^q} g_i(x,y). \end{aligned} \tag{3}$$

To solve global DSBO (2), Chen et al. (2022b) proposes a JHIP oracle to estimate the Jacobian-Hessian-inverse product while Yang et al. (2022) introduces a Hessian-inverse estimation subroutine based on Neumann series approach which can be dated back to Ghadimi and Wang (2018).

Algorithm	Problem Type	Network	Samples	Full Jacobian	Mini-batch
FEDNEST	FBO	Centralized	$\tilde{\mathcal{O}}(\epsilon^{-2})$	No	No
SPDB	P-DSBO	Decentralized	$\tilde{\mathcal{O}}(\epsilon^{-2})$	No	Yes
VRBO	P-DSBO	Decentralized	$\tilde{\mathcal{O}}(\epsilon^{-2})$	No	No
DSBO-JHIP	G-DSBO	Decentralized	$\mathcal{O}(\epsilon^{-3})$	Yes	No
DSBO	G-DSBO	Decentralized	$\tilde{\mathcal{O}}(\epsilon^{-2})$	Yes	No
MA-DSBO	G-DSBO	Decentralized	$\tilde{\mathcal{O}}(\epsilon^{-2})$	No	No

Table 1: We compare our Algorithm 3 (MA-DSBO) with existing distributed bilevel optimization algorithms: FEDNEST (Tarzanagh et al., 2022), SPDB (Lu et al., 2022), VRBO (Gao et al., 2022), DSBO-JHIP (Chen et al., 2022b), and DSBO (Yang et al., 2022). The problem types include Federated Bilevel Optimization (FBO), Personalized-Decentralized Bilevel Optimization (P-DSBO), and Global-Decentralized Bilevel Optimization (G-DSBO). 'Samples' represents the sample complexity of finding an ϵ-stationary point. $\tilde{\mathcal{O}}$ hides the $\log(\frac{1}{\epsilon})$ factor. 'Full Jacobian' refers to whether the algorithm requires computing full Hessian or Jacobian matrix. 'Mini-batch' refers to whether the algorithm requires their batch size depending on ϵ^{-1}.

However, they both require computing the full Jacobian or Hessian matrices, which is extremely time-consuming when q is large. In comparison, computing a Hessian-vector or Jacobian-vector product is more efficient in large-scale machine learning problems (Bottou et al., 2018), and is commonly used in vanilla bilevel optimization (Ghadimi and Wang, 2018; Ji et al., 2021; Chen et al., 2021) to avoid computing the Hessian inverse. In personalized DSBO 3, local computation is sufficient to approximate $\nabla f_i(x, y_i^*(x))$, and thus does not require computing the Hessian or Jacobian matrices and single-agent bilevel optimization methods can be directly incorporated in the distributed regime. In our paper we propose a novel algorithm that estimates the global hypergradient using only first-order oracle and matrix-vector products oracle. Based on this we further design our algorithm for solving DSBO that does not require to compute the full Jacobian or Hessian matrices. We summarize the results of aforementioned works and our results in Table 1 [1].

Notation. We denote by $\nabla f(x,y)$ and $\nabla^2 f(x,y)$ the gradient and Hessian matrix of f, respectively. We use $\nabla_x f(x,y)$ and $\nabla_y f(x,y)$ to represent the gradients of f with respect to x and y, respectively. Denote by $\nabla^2_{xy} f(x,y) \in \mathbb{R}^{p\times q}$ the Jacobian matrix of f and $\nabla^2_y f(x,y)$ the Hessian matrix of f with respect to y. $\|\cdot\|$ denotes the ℓ_2 norm for vectors and Frobenius norm for matrices, unless specified. $\mathbf{1}_n$ is the all one vector in $\mathbb{R}^n$, and $J_n = \mathbf{1}_n\mathbf{1}_n^\top$ is the $n\times n$ all one matrix. We use uppercase letters to represent the matrix that collecting all the variables (corresponding lowercase) as columns. For example $X_k = (x_{1,k}, ..., x_{n,k})$, $Y_k^{(t)} = \left(y_{1,k}^{(t)}, ..., y_{n,k}^{(t)}\right)$. We add an overbar to a letter to denote the average over all nodes. For example, $\bar{x}_k = \frac{1}{n}\sum_{i=1}^n x_{i,k}$, $\bar{y}_k^{(t)} = \frac{1}{n}\sum_{i=1}^n y_{i,k}^{(t)}$.

2 Preliminaries

The following assumptions are used throughout this paper. They are standard assumptions that are made in the literature on bilevel optimization (Chen et al., 2021; Ghadimi and Wang, 2018;

[1] Since Hessian matrix of a function f can be viewed as the Jacobian matrix of the ∇f, for simplicity we use 'Full Jacobian' in the fifth column to refer to both Full Hessian and Full Jacobian computation.

Hong et al., 2020; Ji et al., 2021; Huang et al., 2022) and decentralized optimization (Qu and Li, 2017; Nedic et al., 2017; Lian et al., 2017; Tang et al., 2018).

Assumption 2.1 (Smoothness). *There exist positive constants μ_g, $L_{f,0}$, $L_{f,1}$, $L_{g,1}$, $L_{g,2}$ such that for any i, functions f_i, ∇f_i, ∇g_i, $\nabla^2 g_i$ are $L_{f,0}$, $L_{f,1}$, $L_{g,1}$, $L_{g,2}$ Lipschitz continuous respectively, and function g_i is μ_g-strongly convex in y.*

Assumption 2.2 (Network topology). *The weight matrix $W = (w_{ij}) \in \mathbb{R}^{n\times n}$ is symmetric and doubly stochastic, i.e.:*

$$W = W^\top, \quad W\mathbf{1_n} = \mathbf{1_n}, \quad w_{ij} \geq 0, \forall i, j,$$

and its eigenvalues satisfy $1 = \lambda_1 > \lambda_2 \geq ... \geq \lambda_n$ and $\rho := \max\{|\lambda_2|, |\lambda_n|\} < 1$.

The weight matrix given in Assumption 2.2 characterizes the network topology by setting the weight parameter between agent i and agent j to be w_{ij}. The condition $\rho < 1$ is termed as 'spectral gap', and is used in distributed optimization to ensure the decay of the consensus error, i.e., $\frac{\mathbb{E}\left[\|X_k - \bar{x}_k \mathbf{1}_n^\top\|^2\right]}{n}$, among the agents (Lian et al., 2017), which eventually guarantees the consensus among agents.

Assumption 2.3 (Gradient heterogeneity). *There exists a constant $\delta \geq 0$ such that for all $1 \leq i \leq n, x \in \mathbb{R}^p, y \in \mathbb{R}^q$,*

$$\|\nabla_y g_i(x, y) - \frac{1}{n}\sum_{l=1}^{n} \nabla_y g_l(x, y)\| \leq \delta.$$

The above assumption is commonly used in distributed optimization literature (see, e.g., Lian et al. (2017)), and it indicates the level of similarity between the local gradient and the global gradient. Moreover, it is weaker than the Assumption 3.4 (iv) of Yang et al. (2022) which assumes that $\nabla_y g_i(x, y; \xi)$ has a bounded second moment. This is because the bounded second moment implies the boundedness of $\nabla_y g(x, y)$, as we have

$$\|\nabla_y g(x, y)\|^2 \leq \mathbb{E}\left[\|\nabla_y g(x, y) - \nabla_y g(x, y; \xi)\|^2\right] + \|\nabla_y g(x, y)\|^2 = \mathbb{E}\left[\|\nabla_y g(x, y; \xi)\|^2\right],$$

where the equality holds since we have $\mathbb{E}\left[\|X\|^2\right] = \mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] + \|\mathbb{E}[X]\|^2$ for any random vector X. It directly gives the inequality in Assumption 2.3. However Assumption 2.3 does not imply the boundedness of $\nabla_y g(x, y)$ (e.g., $g_i(x, y) = y^\top y$ for all i satisfies Assumption 2.3 but does not have bounded gradient.)

Assumption 2.4 (Bounded variance). *The stochastic derivatives, $\nabla f_i(x, y; \phi)$, $\nabla g_i(x, y; \xi)$, and $\nabla^2 g_i(x, y; \xi)$, are unbiased with bounded variances σ_f^2, $\sigma_{g,1}^2$, $\sigma_{g,2}^2$, respectively.*

Note that we do not make any assumptions on whether the data distributions are heterogeneous or identically distributed.

3 DSBO Algorithm with Improved Per-Iteration Complexity

We start with following standard result in the bilevel optimization literature (Ghadimi and Wang, 2018; Hong et al., 2020; Ji et al., 2020; Chen et al., 2021) that gives a closed form expression of the hypergradient $\nabla\Phi(x)$, which makes gradient-based bilevel optimization tractable in the first place.

Lemma 3.1. *Suppose Assumption 2.1 holds. The hypergradient $\nabla\Phi(x)$ of (2) takes the form:*

$$\nabla\Phi(x) = \frac{1}{n}\left(\sum_{i=1}^{n}\nabla_x f_i(\tilde{x})\right) - \nabla^2_{xy}g(\tilde{x})\left(\nabla^2_y g(\tilde{x})\right)^{-1}\left[\frac{1}{n}\left(\sum_{i=1}^{n}\nabla_y f_i(\tilde{x})\right)\right], \tag{4}$$

where $\tilde{x} = (x, y^*(x))$.

We also include smoothness properties of $\nabla\Phi(x)$ and $y^*(x)$ in Section A in the appendix.

3.1 Main challenge

As discussed in Chen et al. (2022b) and Yang et al. (2022), the main challenge in designing DSBO algorithms is to estimate the global hypergradient. This is challenging because of the data heterogeneity across agents, which leads to

$$\nabla^2_{xy}g(x, y^*(x))\left(\nabla^2_y g(x, y^*(x))\right)^{-1} \neq \frac{1}{n}\sum_{i=1}^{n}\nabla^2_{xy}g_i(x, y_i^*(x))\left(\nabla^2_y g_i(x, y_i^*(x))\right)^{-1}, \tag{5}$$

where $y_i^*(x) = \arg\min_{y\in\mathbb{R}^q} g_i(x, y)$. This shows that simply averaging the local hypergradients does not give a good approximation to the global hypergradient. A decentralized approach should be designed to estimate the global hypergradient $\nabla\Phi(x)$.

To this end, the JHIP oracle in Chen et al. (2022b) manages to estimate

$$\left(\sum_{i=1}^{n}\nabla^2_{xy}g_i(x, y^*(x))\right)\left(\sum_{i=1}^{n}\nabla^2_y g_i(x, y^*(x))\right)^{-1}$$

using decentralized optimization approach, and Yang et al. (2022) proposed to estimate the global Hessian-inverse, i.e.,

$$\left(\sum_{i=1}^{n}\nabla^2_y g_i(x, y^*(x))\right)^{-1}$$

via a Neumann series based approach. Instead of focusing on full matrices computation, we consider approximating

$$z = \left(\sum_{i=1}^{n}\nabla^2_y g_i(x, y^*(x))\right)^{-1}\left(\sum_{i=1}^{n}\nabla_y f_i(x, y^*(x))\right). \tag{6}$$

According to (4), the global hypergradient is given by

$$\nabla\Phi(x) = \frac{1}{n}\sum_{i=1}^{n}(\nabla_x f_i(x, y^*(x)) - \nabla^2_{xy}g_i(x, y^*(x))z). \tag{7}$$

From the above expression we know that as long as node i can have a good estimate of $\nabla_x f_i(x, y^*(x))$ and $\nabla^2_{xy}g_i(x, y^*(x))z$, then on average the update will be a good approximation to the global hypergradient. More importantly, the process of estimating z can avoid computing the full Hessian or Jacobian matrices.

3.2 Hessian-Inverse-Gradient-Product oracle

Solving (6) is essentially a decentralized optimization with a strongly convex quadratic objective function. Suppose each agent only has access to $H_i \in \mathbb{S}_{++}^{q\times q}$ and $b_i \in \mathbb{R}^q$, and all the agents collectively solve for

$$\sum_{i=1}^n H_i z = \sum_{i=1}^n b_i, \text{ or } z = \left(\sum_{i=1}^n H_i\right)^{-1}\left(\sum_{i=1}^n b_i\right). \tag{8}$$

From an optimization perspective, the above expression is the optimality condition of:

$$\min_{z\in\mathbb{R}^q} \frac{1}{n}\sum_{i=1}^n h_i(z), \text{ where } h_i(z) = \frac{1}{2}z^\top H_i z - b_i^\top z. \tag{9}$$

Hence we can design a decentralized algorithm to solve for z without the presence of a central server. Based on this observation and (7), we present our Hessian-Inverse-Gradient Product oracle in Algorithm 1.

Algorithm 1: Hessian-Inverse-Gradient Product oracle

Input: $(H_{i,t}^{(k)}, b_{i,t}^{(k)})$, for $0 \le t \le N$ accessible only to agent i. Stepsize γ, iteration number N, and $d_{i,0}^{(k)} = -b_{i,0}^{(k)}$, $s_{i,0}^{(k)} = -b_{i,0}^{(k)}$, and $z_{i,0}^{(k)} = 0$

1 **for** $t = 0, 1, ..., N-1$ **do**
2 **for** $i = 1, ..., n$ **do**
3 $z_{i,t+1}^{(k)} = \sum_{j=1}^n w_{ij} z_{j,t}^{(k)} - \gamma d_{i,t}^{(k)}$,
4 $s_{i,t+1}^{(k)} = H_{i,t+1}^{(k)} z_{i,t+1}^{(k)} - b_{i,t+1}^{(k)}$,
5 $d_{i,t+1}^{(k)} = \sum_{i=1}^n w_{ij} d_{j,t}^{(k)} + s_{i,t+1}^{(k)} - s_{i,t}^{(k)}$
6 **end**
7 **end**

Output: $z_{i,N}^{(k)}$ on each node.

It is known that vanilla decentralized gradient descent (DGD) with a constant stepsize only converges to a neighborhood of the optimal solution even under the deterministic setting (Yuan et al., 2016). Therefore, one must use diminishing stepsize in DGD, and this leads to the sublinear convergence rate even when the objective function is strongly convex. To resolve this issue, there are various decentralized algorithms with a fixed stepsize (Xu et al., 2015; Di Lorenzo and Scutari, 2016; Shi et al., 2015; Nedic et al., 2017; Qu and Li, 2017) achieving linear convergence on a strongly convex function in the deterministic setting. Among them, one widely used technique is the gradient tracking method (Xu et al., 2015; Qu and Li, 2017; Nedic et al., 2017; Pu and Nedić, 2021), which is also incorporated in our Algorithm 1. Instead of using the local stochastic gradient in line 3 of Algorithm 1, we maintain another set of variables $d_{i,t+1}^{(k)}$ in line 5 as the gradient tracking step. We will utilize the linear convergence property of gradient tracking in our convergence analysis.

Note that for simplicity we write $H_{i,t}^{(k)} = \nabla_y^2 g_i(x_{i,k}, y_{i,k}^{(T)}; \xi_{i,t})$ in line 1 of Algorithm 2, however, the real implementation only requires Hessian-vector products, as shown in Algorithm 1, and we do not need to compute the full Hessian.

Algorithm 2: Hypergradient Estimation

Input: Samples $\phi = (\phi_{i,0}, ..., \phi_{i,N})$, $\xi = (\xi_{i,0}, ..., \xi_{i,N})$ on node i.

Run Algorithm 1 with $H_{i,t}^{(k)} = \nabla_y^2 g_i(x_{i,k}, y_{i,k}^{(T)}; \xi_{i,t})$, $b_{i,t}^{(k)} = \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,t})$ to get $z_{i,N}^{(k)}$.

Set $u_{i,k} = \nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,0}) - \nabla_{xy}^2 g_i(x_{i,k}, y_{i,k}^{(T)}; \xi_{i,0}) z_{i,N}^{(k)}$.

Output: $u_{i,k}$ on node i.

3.3 Decentralized Stochastic Bilevel Optimization

Now we are ready to present our DSBO algorithm with the moving average technique, which we refer to as the MA-DSBO algorithm. In Algorithm 3 we adopt the basic structure of double-loop

Algorithm 3: MA-DSBO Algorithm

Input: $W, N, K, T, \alpha_k, \beta_k, x_i^{(0)}$.

1 **for** $k = 0, 1, ..., K-1$ **do**

2 $y_{i,k}^{(0)} = y_{i,k-1}^{(T)}$ with $y_{i,k}^{(0)} = 0$.

3 **for** $t = 0, 1, ..., T-1$ **do**

4 **for** $i = 1, ...,$ **do**

5 $y_{i,k}^{(t+1)} = \sum_{j=1}^n w_{ij} y_{j,k}^{(t)} - \beta_k v_{i,k}^{(t)}$ with $v_{i,k}^{(t)} = \nabla_y g_i(x_{i,k}, y_{i,k}^{(t)}; \tilde{\xi}_{i,k}^{(t)})$

6 **end**

7 **end**

8 Run Algorithm 2 and set the output as $u_{i,k}$.

9 **for** $i = 1, ..., n$ **do**

10 $x_{i,k+1} = \sum_{j=1}^n w_{ij} x_{j,k} - \alpha_k r_{i,k}$.

11 $r_{i,k+1} = (1 - \alpha_k) r_{i,k} + \alpha_k u_{i,k}$.

12 **end**

13 **end**

Output: $\bar{x}_K = \frac{1}{n} \sum_{i=1}^n x_{i,K}$.

bilevel optimization algorithm (Ghadimi and Wang, 2018; Ji et al., 2021; Chen et al., 2021) – we first run T-step inner loop (line 3-7) to obtain a good approximation of y^*. Next, we run Algorithm 2 to estimate the hypergradient. To reduce the order of the bias in hypergradient estimation error (see Section 3.4.1 for details), we introduce the moving average update to maintain another set of variables $r_{i,k}$ as the update direction of x. The using of the moving average update helps reduce the order of bias in the stochastic gradient estimate. It is worth noting that similar techniques have been used in the context of nested stochastic composition optimization in Ghadimi et al. (2020); Balasubramanian et al. (2022).

We now introduce our notion of convergence. Specifically, the ϵ-stationary point of (3) is defined as follows.

Definition 3.2. *For a sequence $\{\bar{x}_k\}_{k=0}^K$ generated by Algorithm 3, if $\min_{0 \le k \le K} \mathbb{E}\left[\|\nabla \Phi(\bar{x}_k)\|^2\right] \le \epsilon$ for some positive integer K, then we say that we find an ϵ-stationary point of (3).*

The above notion of stationary point is commonly used in decentralized non-convex stochastic optimization (Lian et al., 2017). When $\epsilon = 0$, it indicates that the hypergradient at some iterate $\bar{x}_k$ is zero. The convergence result of Algorithm 3 is given in Theorem 3.3.

Theorem 3.3. *Suppose Assumptions 2.1, 2.2, 2.3, and 2.4 hold. There exist constants* [2] $0 < c_1 < c_2$ *such that in Algorithm 3 if we set* $\gamma \in (c_1, c_2)$, $T \geq 2$, *and*

$$\alpha_k \equiv \Theta\left(\frac{1}{\sqrt{K}}\right), \ \beta_k \equiv \Theta\left(\frac{1}{\sqrt{K}}\right), \ N = \Theta\left(\log K\right),$$

then we have

$$\min_{0\leq k\leq K} \mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \ \min_{0\leq k\leq K} \frac{\mathbb{E}\left[\|X_k - \bar{x}_k\mathbf{1}_n^\top\|^2\right]}{n} = \mathcal{O}\left(\frac{1}{K}\right).$$

Note that this theorem indicates that the consensus error is of order $\mathcal{O}\left(\frac{1}{K}\right)$, and for any positive constant ϵ, the iteration complexity of Algorithm 3 for obtaining an ϵ-stationary point of (2) is $\mathcal{O}(\epsilon^{-2})$. Moreover, we have the following corollary that gives the sample complexity of our algorithm.

Corollary 3.1. *Suppose the conditions of Theorem 3.3 hold. For any* $\epsilon > 0$, *if we set* $K = \mathcal{O}\left(\epsilon^{-2}\right)$, $N = \Theta(\log \frac{1}{\epsilon})$, *and* $T = 2$, *then in Algorithm 3 the sample complexity to find an* ϵ*-stationary point is* $\mathcal{O}(\epsilon^{-2}\log(\frac{1}{\epsilon}))$.

It is worth noting that $T \geq 2$ in Theorem 3.3 implies, to some extent, that by setting a single timescale, more inner loop iterations will not help improve the convergence result in terms of K. This observation partially answers the decentralized version of the question 'Will Bilevel Optimizers Benefit from Loops?' mentioned in the title of Ji et al. (2022). The hypergradient estimation algorithms (i.e., HIGP oracle and Algorithm 2) provide an additional $\mathcal{O}(\log \frac{1}{\epsilon})$ factor in the sample complexity, which matches Chen et al. (2021). To further remove the log factor, Arbel and Mairal (2021) applies warm start to hypergradient estimation and uses mini-batch method to reduce this complexity and eventually obtain $\mathcal{O}(\epsilon^{-2})$. It would be interesting to study how to apply the warm start strategy to remove the log factor in our complexity bound without using mini-batch method. One restriction of Theorem 3.3 is that we do not obtain the convergence rate $\mathcal{O}(\frac{1}{\sqrt{nK}})$, i.e., the linear speedup in terms of the number of the agents. The recent work of Yang et al. (2022) achieves linear speedup. However, some of their assumptions are restrictive (see Section B for a detailed discussion). It would be interesting to study how to incorporate Jacobian-computing-free algorithm in DSBO under the mild assumptions without affecting linear speedup.

3.4 Proof sketch

In this section we briefly introduce a sketch of our proof for Theorem 3.3 as well as the ideas of the algorithm design. Throughout our analysis, we define the filtration as

$$\mathcal{F}_k = \sigma\left(\bigcup_{i=1}^n \{y_{i,0}^{(T)}, ..., y_{i,k}^{(T)}, x_{i,0}, ..., x_{i,k}, r_{i,0}, ..., r_{i,k}\}\right).$$

3.4.1 Moving average method

The moving average method used in line 11 of Algorithm 3 serves as a key step in setting up the convergence analysis framework. We focus on estimating

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

[2] The constants are independent of K and the details are given in the appendix.

which provides another optimality measure for finding the ϵ-stationary point since we know

$$\mathbb{E}\left[\|\bar{r}_k\|^2 + \|\bar{r}_k - \nabla\Phi(\bar{x}_k)\|^2\right] \geq \frac{1}{2}\mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right].$$

Then, it can be shown that by appropriately choosing parameters (see Lemma A.11 and A.12 for details), we can obtain

$$\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\left[\|\bar{r}_k\|^2 + \|\bar{r}_k - \nabla\Phi(\bar{x}_k)\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{K}} + \frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2\right]\right),$$

which implies that it suffices to bound the hypergradient estimation error, namely, the second term on the right hand side of the above equality. The moving average technique reduces the bias in the hypergradient estimate so that we can directly bound $\mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2\right]$ instead of $\mathbb{E}\left[\|\bar{u}_k - \nabla\Phi(\bar{x}_k)\|^2\right]$, and the former one makes use of the linear convergence property of the gradient tracking methods, which is elaborated in the next section.

3.4.2 Convergence of HIGP

Define

$$y_k^* = y^*(\bar{x}_k),\ z_*^{(k)} = \left(\sum_{i=1}^{n}\nabla_y^2 g_i(\bar{x}_k, y_k^*)\right)^{-1}\left(\sum_{i=1}^{n}\nabla_y f_i(\bar{x}_k, y_k^*)\right).$$

To bound the hypergradient estimation error, a rough analysis (see Lemma A.13) show that

$$\mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2\right] = \mathcal{O}\Bigg(\mathbb{E}\left[\|X_k - \bar{x}_k\mathbf{1}^\top\|^2 + \|Y_k^{(T)} - \bar{y}_k^{(T)}\mathbf{1}^\top\|^2 + \|\bar{y}_k^{(T)} - y_k^*\|^2\right] + E\left[\|\mathbb{E}\left[z_{i,N}^{(k)} - \bar{z}_N^{(k)}|\mathcal{F}_k\right]\|^2 + \|\mathbb{E}\left[\bar{z}_N^{(k)}|\mathcal{F}_k\right] - z_*^{(k)}\|^2\right]\Bigg),$$

where the first two terms on the right hand side denote the consensus error among agents, and can be bounded via techniques in decentralized optimization (Lemma A.7). The third term represents the inner loop estimation error, which can be bounded by considering its decrease as k increases (Lemma A.8). Our novelty lies in bounding the last two terms – the consensus and convergence analysis of the HIGP oracle. Observe that by setting

$$\dot{z}_{i,t}^{(k)} = \mathbb{E}\left[z_{i,t}^{(k)}|\mathcal{F}_k\right], \dot{d}_{j,t}^{(k)} = \mathbb{E}\left[d_{j,t}^{(k)}|\mathcal{F}_k\right], \dot{s}_{i,t}^{(k)} = \mathbb{E}\left[s_{i,t}^{(k)}|\mathcal{F}_k\right],$$

we know from Algorithm 1

$$\begin{aligned}
\dot{z}_{i,t+1}^{(k)} &= \sum_{j=1}^{n} w_{ij}\dot{z}_{j,t}^{(k)} - \gamma\dot{d}_{i,t}^{(k)},\ Z_0^{(k)} = 0,\\
\dot{d}_{i,t+1}^{(k)} &= \sum_{i=1}^{n} w_{ij}\dot{d}_{j,t}^{(k)} + \dot{s}_{i,t+1}^{(k)} - \dot{s}_{i,t}^{(k)},\\
\dot{s}_{i,t}^{(k)} &= \nabla_y^2 g_i(x_{i,k}, y_{i,k}^{(T)})\dot{z}_{i,t}^{(k)} - \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}),
\end{aligned}$$

which is exactly a deterministic gradient descent scheme with gradient tracking on a strongly convex and smooth quadratic function. Hence the linear convergence results in gradient tracking methods can be applied, and this also explains why γ can be chosen as a constant that is independent of K. Mathematically, in Lemmas A.9 and A.13 we explicitly characterize the error and eventually obtain the final convergence result in Theorem 3.3.

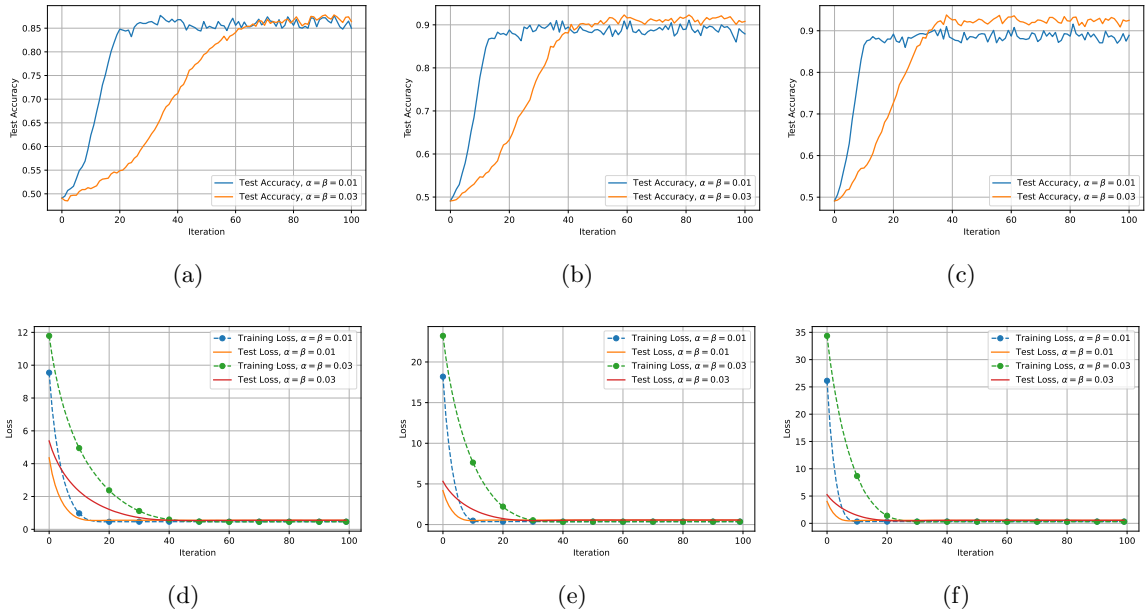

Figure 1: ℓ^2-regularized logistic regression on synthetic data.

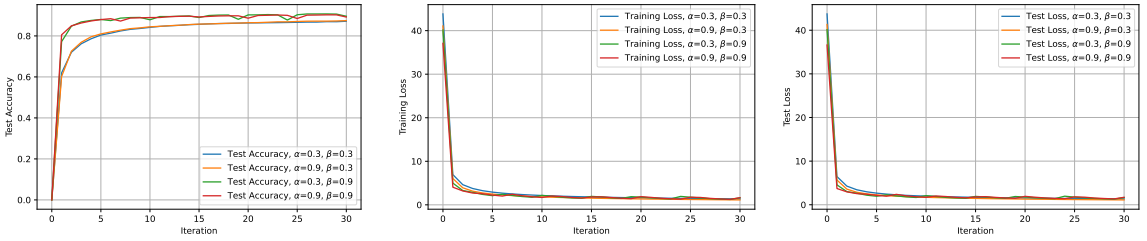

Figure 2: ℓ^2-regularized logistic regression on MNIST.

4 Numerical experiments

In this section we study the applications of Algorithm 3 on hyperparameter optimization:

$$\min_{\lambda \in \mathbb{R}^p} \quad \frac{1}{n} \sum_{i=1}^{n} f_i(\lambda, \omega^*(\lambda)),$$
$$\text{s.t.} \quad \omega^*(\lambda) = \arg\min_{w \in \mathbb{R}^q} \frac{1}{n} \sum_{i=1}^{n} g_i(\lambda, \omega),$$

where we aim at finding the optimal hyperparameter λ under the constraint that $\omega^*(\lambda)$ is the optimal model parameter given λ. We consider both the synthetic and real world data. Comparing to hypergradient estimation algorithms in Chen et al. (2022b) and Yang et al. (2022), our HIGP oracle (Algorithm 1) reduces both the per-iteration complexity and storage from $\mathcal{O}(q^2)$ to $\mathcal{O}(q)$. All the experiments are performed on a local device with 8 cores ($n = 8$) using mpi4py (Dalcin and Fang, 2021) for parallel computing and PyTorch (Paszke et al., 2019) for computing stochastic oracles. The network topology is set to be the ring topology with the weight matrix $W = (w_{ij})$

given by

$$w_{ii} = w,\ w_{i,i+1} = w_{i,i-1} = \frac{1-w}{2},\ \text{for some } w \in (0,1).$$

Here $w_{1,0} = w_{1,n}$ and $w_{n,n+1} = w_{n,1}$. In other words, the neighbors of agent i only include $i-1$ and $i+1$ for $i = 1, 2, ..., n$ with 0 and $n+1$ representing n and 1 respectively.

4.1 Heterogeneous and normally distributed data

Following Pedregosa (2016); Grazzi et al. (2020); Chen et al. (2022b), f_i and g_i are defined as:

$$f_i(\lambda, \omega) = \sum_{(x_e, y_e) \in \mathcal{D}_i'} \psi(y_e x_e^\top \omega),$$

$$g_i(\lambda, \omega) = \sum_{(x_e, y_e) \in \mathcal{D}_i} \psi(y_e x_e^\top \omega) + \frac{1}{2} \sum_{i=1}^{p} e^{\lambda_i} \omega_i^2,$$

where $\psi(x) = \log(1+e^{-x})$ and $p = 200$ denotes the dimension parameter. A ground truth vector w^* is generated in the beginning, and each $x_e \in \mathbb{R}^p$ is generated according to the normal distribution. To introduce heterogeneity, we set r as the heterogeneity rate, and the data distribution of x_e on node i is $\mathcal{N}(0, i^2 \cdot r^2)$. Then we set $y_e = x_e^\top w + \varepsilon \cdot z$, where $\varepsilon = 0.1$ denotes the noise rate and $z \in \mathbb{R}^p$ is the noise vector sampled from standard normal distribution. The task is to learn the optimal regularization parameter $\lambda \in \mathbb{R}^p$. In Figure 1(a), 1(b) and 1(c) (and similarly for 1(d), 1(e), and 1(f)) we set r as 0.5, 1.0, and 1.5 respectively. The accuracy and loss results demonstrate that our algorithm works well under different homogeneity rates.

4.2 MNIST

Now we consider hyperparameter optimization on MNIST dataset (LeCun et al., 1998). Following Grazzi et al. (2020), we have

$$f_i(\lambda, \omega) = \frac{1}{|\mathcal{D}_i'|} \sum_{(x_e, y_e) \in \mathcal{D}_i'} L(x_e^\top \omega, y_e),$$

$$g_i(\lambda, \omega) = \frac{1}{|D_i|} \sum_{(x_e, y_e) \in \mathcal{D}_i} L(x_e^\top \omega, y_e) + \frac{1}{cp} \sum_{i=1}^{c} \sum_{j=1}^{p} e^{\lambda_j} \omega_{ij}^2,$$

where $c = 10,\ p = 784$ denote the number of classes and the number of features, $\omega \in \mathbb{R}^{c \times p}$ is the model parameter, and L denotes the cross entropy loss. $\mathcal{D}_i$ and $\mathcal{D}_i'$ denote the training and validation set respectively. The batch size is 1000 in each stochastic oracle. We include the numerical results of different stepsize choices in Figure 2. Note that in previous algorithms (Chen et al., 2022b; Yang et al., 2022) one Hessian matrix of the lower level function requires $\mathcal{O}(c^2 p^2)$ storage, while in our algorithm a Hessian-vector product only requires $\mathcal{O}(cp)$ storage, which improves both the space and the communication complexity. The accuracy and the loss curves indicate that our DSBO Algorithm 3 has a considerably good performance on real world dataset.

5 Conclusion

In this paper, we propose a DSBO algorithm that does not require computing full Hessian and Jacobian matrices, thereby improving the per-iteration complexity of currently known DSBO

algorithms, under mild assumptions. Moreover, we prove that our algorithm achieves $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity, which matches the result in state-of-the-art single-agent bilevel optimization algorithms. We would like to point out that Assumption 2.3 (or bounded second moment condition in Yang et al. (2022)) requires certain types of upper bounds on $\|\nabla_y g(x,y)\|$, which may not hold in decentralized optimization (see, e.g., Pu and Nedić (2021)). It is interesting to study decentralized stochastic bilevel optimization without this type of conditions, and one promising direction is to apply variance reduction techniques like in Tang et al. (2018).

Acknowledgments

XC acknowledges the support by UC Davis Dean's Graduate Summer Fellowship. SM acknowledges the support by NSF grants DMS-1953210 and CCF-2007797, and a startup fund at Rice University. KB acknowledges the support by National Science Foundation (NSF) via the grant NSF DMS-2053918.

References

H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017. (Cited on page 2.)

M. Arbel and J. Mairal. Amortized implicit differentiation for stochastic bilevel optimization. *arXiv preprint arXiv:2111.14580*, 2021. (Cited on page 9.)

K. Balasubramanian, S. Ghadimi, and A. Nguyen. Stochastic multilevel composition optimization algorithms with level-independent convergence rates. *SIAM Journal on Optimization*, 32(2): 519–544, 2022. (Cited on page 8.)

L. Bertinetto, J. F. Henriques, P. H. Torr, and A. Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018. (Cited on pages 1 and 3.)

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018. (Cited on page 4.)

T. Chen, Y. Sun, and W. Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34, 2021. (Cited on pages 1, 2, 3, 4, 5, 8, and 9.)

T. Chen, Y. Sun, Q. Xiao, and W. Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2466–2488. PMLR, 2022a. (Cited on page 3.)

X. Chen, M. Huang, and S. Ma. Decentralized bilevel optimization. *arXiv preprint arXiv:2206.05670*, 2022b. (Cited on pages 2, 3, 4, 6, 11, and 12.)

L. Dalcin and Y.-L. L. Fang. mpi4py: Status update after 12 years of development. *Computing in Science & Engineering*, 23(4):47–54, 2021. (Cited on page 11.)

P. Di Lorenzo and G. Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016. (Cited on page 7.)

J. Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012. (Cited on page 3.)

L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018. (Cited on pages 1 and 3.)

H. Gao, B. Gu, and M. T. Thai. Stochastic bilevel distributed optimization over a network. *arXiv preprint arXiv:2206.15025*, 2022. (Cited on pages 2, 3, and 4.)

S. Ghadimi and M. Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018. (Cited on pages 1, 3, 4, 5, 8, and 17.)

S. Ghadimi, A. Ruszczynski, and M. Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020. (Cited on page 8.)

S. Gould, B. Fernando, A. Cherian, P. Anderson, R. S. Cruz, and E. Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016. (Cited on page 3.)

R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020. (Cited on pages 3 and 12.)

Z. Guo, Q. Hu, L. Zhang, and T. Yang. Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021. (Cited on page 3.)

M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020. (Cited on pages 1, 3, and 5.)

M. Huang, K. Ji, S. Ma, and L. Lai. Efficiently escaping saddle points in bilevel optimization. *arXiv preprint arXiv:2202.03684*, 2022. (Cited on page 5.)

K. Ji, J. D. Lee, Y. Liang, and H. V. Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. *Advances in Neural Information Processing Systems*, 33: 11490–11500, 2020. (Cited on pages 1 and 5.)

K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892. PMLR, 2021. (Cited on pages 3, 4, 5, and 8.)

K. Ji, M. Liu, Y. Liang, and L. Ying. Will bilevel optimizers benefit from loops. *arXiv preprint arXiv:2205.14224*, 2022. (Cited on page 9.)

M. Kayaalp, S. Vlaski, and A. H. Sayed. Dif-maml: Decentralized multi-agent meta-learning. *IEEE Open Journal of Signal Processing*, 3:71–93, 2022. (Cited on page 2.)

P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in Neural Information Processing Systems*, 34:30271–30283, 2021. (Cited on page 3.)

A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020. (Cited on page 3.)

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. (Cited on page 12.)

J. Li, F. Huang, and H. Huang. Local stochastic bilevel optimization with momentum-based variance reduction. *arXiv preprint arXiv: 2205.01608*, 2022. (Cited on page 3.)

X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017. (Cited on pages 2, 3, 5, and 8.)

S. Lu, X. Cui, M. S. Squillante, B. Kingsbury, and L. Horesh. Decentralized bilevel optimization for personalized client learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5543–5547. IEEE, 2022. (Cited on pages 2, 3, and 4.)

D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015. (Cited on page 3.)

A. Nedic, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017. (Cited on pages 5 and 7.)

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. (Cited on page 11.)

B. A. Pearlmutter. Fast exact multiplication by the hessian. *Neural computation*, 6(1):147–160, 1994. (Cited on page 2.)

F. Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pages 737–746. PMLR, 2016. (Cited on pages 1, 3, and 12.)

S. Pu and A. Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1):409–457, 2021. (Cited on pages 7, 13, and 26.)

G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017. (Cited on pages 5, 7, and 18.)

A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019. (Cited on pages 1 and 3.)

S. S. Ram, A. Nedić, and V. V. Veeravalli. Asynchronous gossip algorithms for stochastic optimization. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pages 3581–3586. IEEE, 2009. (Cited on page 3.)

W. Shi, Q. Ling, G. Wu, and W. Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015. (Cited on page 7.)

J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. (Cited on pages 1 and 3.)

H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu. d^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pages 4848–4856. PMLR, 2018. (Cited on pages 5 and 13.)

D. A. Tarzanagh, M. Li, C. Thrampoulidis, and S. Oymak. Fednest: Federated bilevel, minimax, and compositional optimization. *arXiv preprint arXiv:2205.02215*, 2022. (Cited on pages 3 and 4.)

T. Wu, K. Yuan, Q. Ling, W. Yin, and A. H. Sayed. Decentralized consensus optimization with asynchrony and delays. *IEEE Transactions on Signal and Information Processing over Networks*, 4(2):293–307, 2017. (Cited on page 3.)

J. Xu, S. Zhu, Y. C. Soh, and L. Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 2055–2060. IEEE, 2015. (Cited on page 7.)

F. Yan, S. Sundaram, S. Vishwanathan, and Y. Qi. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2483–2493, 2012. (Cited on page 3.)

J. Yang, K. Ji, and Y. Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34:13670–13682, 2021. (Cited on page 3.)

S. Yang, X. Zhang, and M. Wang. Decentralized gossip-based stochastic bilevel optimization over communication networks. *arXiv preprint arXiv:2206.10870*, 2022. (Cited on pages 2, 3, 4, 5, 6, 9, 11, 12, 13, and 37.)

K. Yuan, Q. Ling, and W. Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016. (Cited on page 7.)

X. S. Zhang, F. Tang, H. H. Dodge, J. Zhou, and F. Wang. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2487–2495, 2019. (Cited on page 2.)

A Analysis

Figure 3 represents the structure of the proof. For convenience we restate our notation convention here again:

- We use the first subscript (usually denoted as i) to represent the agent number, and the second subscript (usually denoted as k or t) to represent the iteration number. For example $x_{i,k}$ represents the x variable of agent i at k-th iteration. For the inner loop iterate like $y_{i,k}^{(t)}$, the superscript t represents the iteration number of the inner loop.
- We use uppercase letters to represent the matrix that collecting all the variables (corresponding lowercase) as columns. For example $X_k = (x_{1,k}, ..., x_{n,k})$, $Y_k^{(t)} = \left(y_{1,k}^{(t)}, ..., y_{n,k}^{(t)}\right)$.

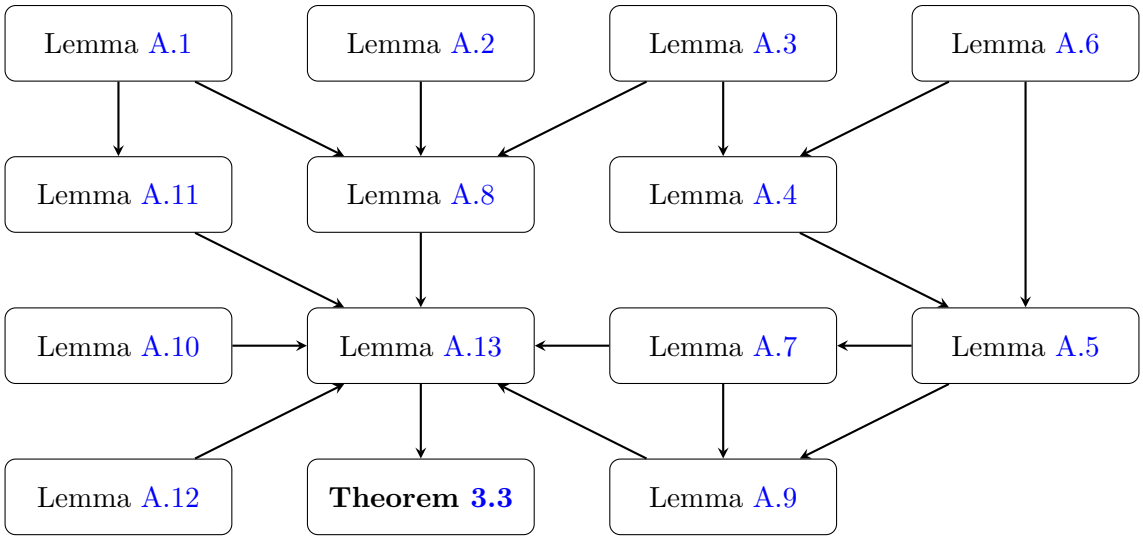

Figure 3: Structure of the proof

- We add an overbar to a letter to denote the average over all nodes. For example, $\bar{x}_k = \frac{1}{n}\sum_{i=1}^n x_{i,k}$, $\bar{y}_k^{(t)} = \frac{1}{n}\sum_{i=1}^n y_{i,k}^{(t)}$.
- The filtration is defined as

$$\mathcal{F}_k = \sigma\left(\bigcup_{i=1}^n \{y_{i,0}^{(T)}, ..., y_{i,k}^{(T)}, x_{i,0}, ..., x_{i,k}, r_{i,0}, ..., r_{i,k}\}\right).$$

We first state several well-known results in bilevel optimization literature (see, e.g., Lemma 2.2 in Ghadimi and Wang (2018).).

Lemma A.1. *Suppose Assumptions 2.1 and 2.4 hold, we know $\nabla\Phi(x)$ and $y^*(x)$ defined in (2) are L_Φ and L_{y^*}-Lipschitz continuous respectively with the constants given by*

$$L_\Phi = L_{f,1} + \frac{2L_{f,1}L_{g,1} + L_{g,2}L_{f,0}^2}{\mu_g} + \frac{2L_{g,1}L_{f,0}L_{g,2} + L_{g,1}^2 L_{f,1}}{\mu_g^2} + \frac{L_{g,2}L_{g,1}^2 L_{f,0}}{\mu_g^3},\ L_{y^*} = \frac{L_{g,1}}{\mu_g}. \quad (10)$$

The following inequality is a standard result and will be used in our later analysis. We prove it here for completeness.

Lemma A.2. *Suppose we are given two sequences $\{a_k\}$ and $\{b_k\}$ that satisfy*

$$a_{k+1} \le \delta a_k + b_k,\ a_k \ge 0,\ b_k \ge 0 \text{ for all } k \ge 0$$

for some $\delta \in (0,1)$. Then we have

$$a_{k+1} \le \delta^{k+1} a_0 + \sum_{i=0}^k b_i \delta^{k-i}.$$

Proof of Lemma A.2. Setting $c_i = \frac{a_i}{\delta^i}$, we know

$$c_{i+1} \le c_i + b_i \cdot \delta^{-i-1} \text{ for all } i \ge 0.$$

Taking summation on both sides (i from 0 to k) and multiplying δ^{k+1}, we know for $k \geq 0$,

$$a_{k+1} \leq \delta^{k+1} a_0 + \sum_{i=0}^{k} b_i \delta^{k-i},$$

which completes the proof. □

The following lemma is standard in stochastic optimization (see, e.g., Lemma 10 in Qu and Li (2017)).

Lemma A.3. *Suppose $f(x)$ is μ-strongly convex and $L-$smooth. For any x and $\eta < \frac{2}{\mu+L}$, define $x^+ = x - \eta \nabla f(x)$, $x^* = \arg\min f(x)$. Then we have*

$$\|x^+ - x^*\| \leq (1 - \eta\mu)\|x - x^*\|$$

Next, we characterize the bounded second moment of the HIGP oracle. Note that Algorithm 1 is essentially decentralized stochastic gradient descent with gradient tracking on a strongly convex quadratic function.

Lemma A.4. *Suppose we are given matrices A_i and vectors b_i such that there exist $0 < \mu < L$ such that $\mu I \preceq A_i \preceq LI$ for $1 \leq i \leq n$. $W = (w_{ij})$ satisfies Assumption 2.2. The sequences $\{x_{i,k}\}$, $\{s_{i,k}\}$ and $\{v_{i,k}\}$ satisfy for any $k \geq 0$ and $1 \leq i \leq n$,*

$$x_{i,k+1} = \sum_{j=1}^{n} w_{ij} x_{j,k} - \alpha s_{i,k},\ s_{i,k+1} = \sum_{j=1}^{n} w_{ij} s_{j,k} + v_{i,k+1} - v_{i,k},\ v_{i,k} = A_{i,k} x_{i,k} - b_{i,k},\ s_{i,0} = v_{i,0},$$
$$\mathbb{E}\left[A_{i,k}\right] = A_i,\ \mathbb{E}\left[b_{i,k}\right] = b_i,\ \mathbb{E}\left[\|A_{i,k} - A_i\|^2\right] \leq \sigma_1^2,\ \mathbb{E}\left[\|b_{i,k} - b_i\|^2\right] \leq \sigma_2^2.$$

Moreover, we assume $A_{i,k}, x_{j,k}, b_{i,k}$ are independent for any $i, j \in \{1, ..., n\}$, $\{A_{i,k}\}_{i=1}^n$ are independent and $\{b_{i,k}\}_{i=1}^n$ are independent. Define

$$\tilde{\sigma}_1^2 = \sigma_1^2 + L^2,\ \tilde{\sigma}_2^2 = \sigma_2^2 + \max_i \|b_i\|^2,\ x^* := \left(\frac{1}{n}\sum_{i=1}^{n} A_i\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} b_i\right),$$
$$C_1 = 9\sigma_1^2 + 6\alpha^2\tilde{\sigma}_1^2 + \frac{18\alpha^2\sigma_1^2\tilde{\sigma}_1^2}{n},\ C_2 = 12\tilde{\sigma}_1^2 + 9\sigma_1^2 + 12\alpha^2 L^2\tilde{\sigma}_1^2 + \frac{18\alpha^2\sigma_1^2\tilde{\sigma}_1^2}{n},$$
$$C_3 = 6\rho^2\tilde{\sigma}_1^2,\ C_4 = 2\sigma_2^2 + \frac{6\alpha^2\sigma_2^2\tilde{\sigma}_1^2}{n} + (9\sigma_1^2 + \frac{18\alpha^2\sigma_1^2\tilde{\sigma}_1^2}{n})\|x^*\|^2,$$
$$c = \left(\frac{\alpha^2}{n}(3\sigma_1^2\|x^*\|^2 + \sigma_2^2),\ 0,\ \frac{(1+\rho^2)}{1-\rho^2}C_4\right)^\top,\ M = \begin{pmatrix} M_{11} & M_{12} & 0 \\ 0 & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{pmatrix},$$
$$M_{11} = 1 - \alpha\mu,\ M_{12} = (\frac{2\alpha}{\mu} + 2\alpha^2)\tilde{\sigma}_1^2,\ M_{22} = \frac{1+\rho^2}{2},\ M_{23} = \alpha^2\frac{1+\rho^2}{1-\rho^2}$$
$$M_{31} = \frac{1+\rho^2}{1-\rho^2}C_1,\ M_{32} = \frac{1+\rho^2}{1-\rho^2}C_2,\ M_{33} = \frac{1+\rho^2}{2} + \frac{1+\rho^2}{1-\rho^2}C_3\alpha^2.$$

If α satisfies

$$\begin{aligned} &(1 + \frac{\alpha\mu}{2})(1-\alpha\mu)^2 + \frac{3\alpha^2\sigma_1^2}{n} < 1 - \alpha\mu,\ 0 < \alpha_1 \leq \alpha \leq \alpha_2 \text{ for some } 0 < \alpha_1 < \alpha_2, \\ &\rho(M) < 1 - \frac{2\alpha\mu}{3}, \quad \text{and } M \text{ has 3 different positive eigenvalues,} \end{aligned} \tag{11}$$

then we have

$$\begin{aligned}
\mathbb{E}\left[\|\bar{x}_{k+1}-x^*\|^2\right] &\le (1-\alpha\mu)\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]+(\frac{2\alpha}{\mu}+2\alpha^2)\frac{\tilde{\sigma}_1^2}{n}\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right]\\
&\quad+\frac{\alpha^2}{n}(3\sigma_1^2\|x^*\|^2+\sigma_2^2),\\
\|X_{k+1}-\bar{x}_{k+1}\mathbf{1}^\top\|^2 &\le \frac{(1+\rho^2)}{2}\|X_k-\bar{x}_k\mathbf{1}^\top\|^2+\alpha^2\frac{1+\rho^2}{1-\rho^2}\|S_k-\bar{s}_k\mathbf{1}^\top\|^2, \qquad (12)\\
\mathbb{E}\left[\frac{\|S_{k+1}-\bar{s}_{k+1}\mathbf{1}^\top\|^2}{n}\right] &\le \frac{1+\rho^2}{1-\rho^2}C_1\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]+\frac{1+\rho^2}{1-\rho^2}C_2\mathbb{E}\left[\frac{\|X_k-\bar{x}_k\mathbf{1}^\top\|^2}{n}\right]\\
&\quad+(\frac{1+\rho^2}{2}+\frac{1+\rho^2}{1-\rho^2}C_3\alpha^2)\mathbb{E}\left[\frac{\|S_k-\bar{s}_k\mathbf{1}^\top\|^2}{n}\right]+\frac{1+\rho^2}{1-\rho^2}C_4.
\end{aligned}$$

Moreover, we set P such that $M=P\cdot diag(\lambda_1,\lambda_2,\lambda_3)P^{-1}$ with $0<\lambda_3<\lambda_2<\lambda_1$ being eigenvalues and each column of P is a unit vector. Define $C_M:=\|P\|_2\|P^{-1}\|_2$, we have

$$\begin{aligned}
&\max\left(\frac{1}{n}\mathbb{E}\left[\|X_k-x^*\mathbf{1}^\top\|^2\right],\frac{1}{n}\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right]\right)\\
\le& 3C_M(1-\frac{2\alpha\mu}{3})^k(\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right]+\mathbb{E}\left[\frac{\|X_0\|^2+\|S_0\|^2}{n}\right])+\frac{5C_M\|c\|}{\alpha\mu}, \qquad (13)\\
&\frac{1}{n}\mathbb{E}\left[\|X_k\|^2\right]\le 6C_M(1-\frac{2\alpha\mu}{3})^k(\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right]+\mathbb{E}\left[\frac{\|X_0\|^2+\|S_0\|^2}{n}\right])+\frac{10C_M\|c\|}{\alpha\mu}+2\|x^*\|^2. \qquad (14)
\end{aligned}$$

Proof of Lemma A.4. Note that by definition of $\tilde{\sigma}_1^2$ and $\tilde{\sigma}_2^2$ we have

$$\begin{aligned}
\mathbb{E}\left[\|A_{i,k}\|^2\right]&=\mathbb{E}\left[\|A_{i,k}-A_i\|^2\right]+\|A_i\|_2^2\le\sigma_1^2+L^2=\tilde{\sigma}_1^2,\\
\mathbb{E}\left[\|b_{i,k}\|^2\right]&=\mathbb{E}\left[\|b_{i,k}-b_i\|^2\right]+\|b_i\|^2\le\sigma_2^2+\max_i\|b_i\|^2=\tilde{\sigma}_2^2.
\end{aligned} \qquad (15)$$

By $s_{i,0}=v_{i,0}$ we know $\bar{s}_0=\bar{v}_0$. From the recursion we know

$$\bar{s}_{k+1}=\bar{s}_k+\bar{v}_{k+1}-\bar{v}_k,$$

and hence $\bar{s}_k=\bar{v}_k$ by induction. For $\bar{x}_k$ we know

$$\begin{aligned}
&\bar{x}_{k+1}-x^*\\
=&\bar{x}_k-x^*-\frac{\alpha}{n}\sum_{i=1}^n(A_{i,k}x_{i,k}-b_{i,k})\\
=&\bar{x}_k-x^*-\frac{\alpha}{n}\sum_{i=1}^n(A_i\bar{x}_k-b_i)+\frac{\alpha}{n}\sum_{i=1}^n(A_i\bar{x}_k-b_i)-\frac{\alpha}{n}\sum_{i=1}^n(A_{i,k}x_{i,k}-b_{i,k})\\
=&(I-\frac{\alpha}{n}\sum_{i=1}^nA_i)(\bar{x}_k-x^*)+\frac{\alpha}{n}\sum_{i=1}^nA_{i,k}(\bar{x}_k-x_{i,k})+\frac{\alpha}{n}\sum_{i=1}^n((A_i-A_{i,k})\bar{x}_k+b_{i,k}-b_i).
\end{aligned}$$

Using the above equality, $\mathbb{E}\left[A_{i,k}\right]=A_i$ and $\mathbb{E}\left[b_{i,k}\right]=b_i$, we know

$$\mathbb{E}\left[\|\bar{x}_{k+1}-x^*\|^2\right]$$

$$
\begin{aligned}
=&\mathbb{E}\left[\|(I-\frac{\alpha}{n}\sum_{i=1}^{n}A_i)(\bar{x}_k-x^*)+\frac{\alpha}{n}\sum_{i=1}^{n}A_{i,k}(\bar{x}_k-x_{i,k})\|^2\right]+\frac{\alpha^2}{n^2}\mathbb{E}\left[\|\sum_{i=1}^{n}((A_i-A_{i,k})\bar{x}_k+b_{i,k}-b_i)\|^2\right]\\
+&\mathbb{E}\left[\langle(I-\frac{\alpha}{n}\sum_{i=1}^{n}A_i)(\bar{x}_k-x^*)+\frac{\alpha}{n}\sum_{i=1}^{n}A_{i,k}(\bar{x}_k-x_{i,k}),\frac{\alpha}{n}\sum_{i=1}^{n}((A_i-A_{i,k})\bar{x}_k+b_{i,k}-b_i)\rangle\right]\\
\leq&(1+\frac{\alpha\mu}{2})(1-\alpha\mu)^2\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]+(1+\frac{2}{\alpha\mu})\frac{\alpha^2\tilde{\sigma}_1^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\bar{x}_k-x_{i,k}\|^2\right]+\frac{\alpha^2}{n^2}(n\sigma_1^2\mathbb{E}\left[\|\bar{x}_k\|^2\right]+n\sigma_2^2)\\
+&\frac{\alpha^2}{2n^2}\sum_{i=1}^{n}\mathbb{E}\left[\sigma_1^2\|\bar{x}_k\|^2+\tilde{\sigma}_1^2\|\bar{x}_k-x_{i,k}\|^2\right]\\
=&(1+\frac{\alpha\mu}{2})(1-\alpha\mu)^2\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]+(\frac{2\alpha}{\mu}+\alpha^2+\frac{\alpha^2}{2n})\frac{\tilde{\sigma}_1^2}{n}\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right]+\frac{\alpha^2}{n}(\frac{3\sigma_1^2}{2}\mathbb{E}\left[\|\bar{x}_k\|^2\right]+\sigma_2^2)\\
\leq&\left[(1+\frac{\alpha\mu}{2})(1-\alpha\mu)^2+\frac{3\alpha^2\sigma_1^2}{n}\right]\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]+(\frac{2\alpha}{\mu}+2\alpha^2)\frac{\tilde{\sigma}_1^2}{n}\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right]+\frac{\alpha^2}{n}(3\sigma_1^2\|x^*\|^2+\sigma_2^2)\\
\leq&(1-\alpha\mu)\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]+(\frac{2\alpha}{\mu}+2\alpha^2)\frac{\tilde{\sigma}_1^2}{n}\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right]+\frac{\alpha^2}{n}(3\sigma_1^2\|x^*\|^2+\sigma_2^2).
\end{aligned}
$$

The first inequality holds because we have

$$
\begin{aligned}
&\mathbb{E}\left[\langle(I-\frac{\alpha}{n}\sum_{i=1}^{n}A_i)(\bar{x}_k-x^*)+\frac{\alpha}{n}\sum_{i=1}^{n}A_{i,k}(\bar{x}_k-x_{i,k}),\frac{\alpha}{n}\sum_{i=1}^{n}((A_i-A_{i,k})\bar{x}_k+b_{i,k}-b_i)\rangle\right]\\
=&\mathbb{E}\left[\langle\frac{\alpha}{n}\sum_{i=1}^{n}A_{i,k}(\bar{x}_k-x_{i,k}),\frac{\alpha}{n}\sum_{i=1}^{n}((A_i-A_{i,k})\bar{x}_k+b_{i,k}-b_i)\rangle\right]\\
=&\mathbb{E}\left[\langle\frac{\alpha}{n}\sum_{i=1}^{n}A_{i,k}(\bar{x}_k-x_{i,k}),\frac{\alpha}{n}\sum_{i=1}^{n}(A_i-A_{i,k})\bar{x}_k\rangle\right]\\
=&\frac{\alpha^2}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(\bar{x}_k-x_{i,k})^\top A_{i,k}^\top(A_i-A_{i,k})\bar{x}_k\right]\leq\frac{\alpha^2}{2n^2}\sum_{i=1}^{n}\mathbb{E}\left[\sigma_1^2\|\bar{x}_k\|^2+\tilde{\sigma}_1^2\|\bar{x}_k-x_{i,k}\|^2\right],
\end{aligned}
$$

the second inequality uses $\|\bar{x}_k\|^2\leq 2\|\bar{x}_k-x^*\|^2+2\|x^*\|^2$, and the third inequality uses (11). For $\|X_{k+1}-\bar{x}_{k+1}\mathbf{1}^\top\|^2$ we know

$$
\begin{aligned}
&\|X_{k+1}-\bar{x}_{k+1}\mathbf{1}^\top\|^2=\|X_kW-\bar{x}_k\mathbf{1}^\top-\alpha(S_k-\bar{s}_k\mathbf{1}^\top)\|^2\\
\leq&(1+\frac{1-\rho^2}{2\rho^2})\rho^2\|X_k-\bar{x}_k\mathbf{1}^\top\|^2+(1+\frac{2\rho^2}{1-\rho^2})\alpha^2\|S_k-\bar{s}_k\mathbf{1}^\top\|^2.
\end{aligned}
\tag{16}
$$

The inequality uses Cauchy-Schwarz inequality and the fact that

$$
\begin{aligned}
&\|X_kW-\bar{x}_k\mathbf{1}^\top\|=\|(X_k-\bar{x}_k\mathbf{1}^\top)(W-\frac{\mathbf{1}\mathbf{1}^\top}{n})\|=\|(W-\frac{\mathbf{1}\mathbf{1}^\top}{n})(X_k-\bar{x}_k\mathbf{1}^\top)^\top\|\\
\leq&\|W-\frac{\mathbf{1}\mathbf{1}^\top}{n}\|_2\|X_k-\bar{x}_k\mathbf{1}^\top\|\leq\rho\|X_k-\bar{x}_k\mathbf{1}^\top\|,
\end{aligned}
$$

where the last inequality uses Assumption 2.2. For $\|S_k - \bar{s}_k \mathbf{1}^\top\|^2$ we know

$$\begin{aligned}
&\|S_{k+1} - \bar{s}_{k+1}\mathbf{1}^\top\|^2 = \|S_k W - \bar{s}_k \mathbf{1}^\top + V_{k+1} - V_k - \bar{v}_{k+1}\mathbf{1}^\top + \bar{v}_k \mathbf{1}^\top\|^2 \\
\leq &(1 + \frac{1-\rho^2}{2\rho^2})\|S_k - \bar{s}_k\mathbf{1}^\top\|^2 + (1 + \frac{2\rho^2}{1-\rho^2})\|(V_{k+1} - V_k)(I - \frac{\mathbf{1}\mathbf{1}^\top}{n})\|^2 \\
= &\frac{1+\rho^2}{2}\|S_k - \bar{s}_k\mathbf{1}^\top\|^2 + \frac{1+\rho^2}{1-\rho^2}\|V_{k+1} - V_k\|^2.
\end{aligned} \tag{17}$$

For $V_{k+1} - V_k$ we have

$$\begin{aligned}
&\mathbb{E}\left[\|V_{k+1} - V_k\|^2\right] \\
= &\sum_{i=1}^n \mathbb{E}\left[\|A_{i,k+1}(x_{i,k+1} - x_{i,k}) + (A_{i,k+1} - A_i + A_i - A_{i,k})x_{i,k} + (b_{i,k} - b_i + b_i - b_{i,k+1})\|^2\right] \\
= &\sum_{i=1}^n \mathbb{E}\left[\|A_{i,k+1}(x_{i,k+1} - x_{i,k}) + (A_{i,k+1} - A_i)x_{i,k}\|^2 + \|(A_i - A_{i,k})x_{i,k}\|^2 + \|b_{i,k} - b_i\|^2 + \|b_i - b_{i,k+1}\|^2\right] \\
\leq &\sum_{i=1}^n \mathbb{E}\left[2\|A_{i,k+1}(x_{i,k+1} - x_{i,k})\|^2 + 2\|(A_{i,k+1} - A_i)x_{i,k}\|^2 + \|(A_i - A_{i,k})x_{i,k}\|^2 + \|b_{i,k} - b_i\|^2 + \|b_i - b_{i,k+1}\|^2\right] \\
\leq &2\tilde{\sigma}_1^2 \mathbb{E}\left[\|X_{k+1} - X_k\|^2\right] + 3\sigma_1^2 \mathbb{E}\left[\|X_k\|^2\right] + 2n\sigma_2^2.
\end{aligned}$$

For $\|X_{k+1} - X_k\|$ we know

$$\begin{aligned}
&\mathbb{E}\left[\|X_{k+1} - X_k\|^2\right] = \mathbb{E}\left[\|X_k W - X_k - \alpha S_k W\|^2\right] \\
= &\mathbb{E}\left[\|(X_k - \bar{x}_k\mathbf{1}^\top)(W - I) - \alpha(S_k W - \bar{s}_k\mathbf{1}^\top) - \alpha\bar{s}_k\mathbf{1}^\top\|^2\right] \\
\leq &3\|W - I\|_2^2 \mathbb{E}\left[\|X_k - \bar{x}_k\mathbf{1}^\top\|^2\right] + 3\alpha^2\rho^2\mathbb{E}\left[\|S_k - \bar{s}_k\mathbf{1}^\top\|^2\right] + 3n\alpha^2\mathbb{E}\left[\|\bar{s}_k\|^2\right] \\
\leq &6\mathbb{E}\left[\|X_k - \bar{x}_k\mathbf{1}^\top\|^2\right] + 3\alpha^2\rho^2\mathbb{E}\left[\|S_k - \bar{s}_k\mathbf{1}^\top\|^2\right] \\
&\qquad + 3\alpha^2(\frac{\sigma_1^2}{n}\mathbb{E}\left[\|X_k\|^2\right] + \sigma_2^2 + 2L^2\mathbb{E}\left[\|X_k - \bar{x}_k\mathbf{1}^\top\|^2 + n\|\bar{x}_k - x^*\|^2\right]) \\
= &(6 + 6\alpha^2L^2)\mathbb{E}\left[\|X_k - \bar{x}_k\mathbf{1}^\top\|^2\right] + 3\alpha^2\rho^2\mathbb{E}\left[\|S_k - \bar{s}_k\mathbf{1}^\top\|^2\right] + \frac{3\alpha^2\sigma_1^2}{n}\mathbb{E}\left[\|X_k\|^2\right] \\
&\qquad + 3n\alpha^2\mathbb{E}\left[\|\bar{x}_k - x^*\|^2\right] + 3\alpha^2\sigma_2^2,
\end{aligned}$$

where the second inequality holds since

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{s}_k\|^2\right] = \mathbb{E}\left[\|\frac{1}{n}\sum_{i=1}^n (A_{i,k}x_{i,k} - b_{i,k})\|^2\right] \\
= &\mathbb{E}\left[\|\frac{1}{n}\sum_{i=1}^n ((A_{i,k} - A_i)x_{i,k} - (b_{i,k} - b_i)) + \frac{1}{n}\sum_{i=1}^n (A_i x_{i,k} - A_i\bar{x}_k) + \frac{1}{n}\sum_{i=1}^n A_i(\bar{x}_k - x^*)\|^2\right] \\
= &\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\|(A_{i,k} - A_i)x_{i,k}\|^2 + \|b_{i,k} - b_i\|^2\right] + \frac{1}{n^2}\mathbb{E}\left[\|\sum_{i=1}^n ((A_i x_{i,k} - A_i\bar{x}_k) + A_i(\bar{x}_k - x^*))\|^2\right] \\
\leq &\frac{\sigma_1^2}{n^2}\mathbb{E}\left[\|X_k\|^2\right] + \frac{\sigma_2^2}{n} + \frac{2L^2}{n}\mathbb{E}\left[\|X_k - \bar{x}_k\mathbf{1}^\top\|^2 + n\|\bar{x}_k - x^*\|^2\right].
\end{aligned}$$

Hence we know

$$\begin{aligned}
&\mathbb{E}\left[\|V_{k+1}-V_k\|^2\right] \le 2\tilde{\sigma}_1^2\mathbb{E}\left[\|X_{k+1}-X_k\|^2\right] + 3\sigma_1^2\mathbb{E}\left[\|X_k\|^2\right] + 2n\sigma_2^2\\
\le& 2\tilde{\sigma}_1^2\left\{(6+6\alpha^2L^2)\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right] + 3\alpha^2\rho^2\mathbb{E}\left[\|S_k-\bar{s}_k\mathbf{1}^\top\|^2\right] + 3n\alpha^2\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]\right\}\\
&+(3\sigma_1^2+\frac{6\alpha^2\sigma_1^2\tilde{\sigma}_1^2}{n})\mathbb{E}\left[\|X_k\|^2\right] + (2n\sigma_2^2+6\alpha^2\sigma_2^2\tilde{\sigma}_1^2)\\
\le& nC_1\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right] + C_2\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right] + \alpha^2C_3\mathbb{E}\left[\|S_k-\bar{s}_k\mathbf{1}^\top\|^2\right] + nC_4,
\end{aligned}$$

where the second inequality uses

$$\|X_k\|^2 \le 3\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2 + n\|\bar{x}_k-x^*\|^2 + n\|x^*\|^2\right].$$

The above inequalities and (17) imply

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}\left[\|S_{k+1}-\bar{s}_{k+1}\mathbf{1}^\top\|^2\right]\\
\le&\frac{1+\rho^2}{2n}\|S_k-\bar{s}_k\mathbf{1}^\top\|^2 + \frac{1+\rho^2}{1-\rho^2}(C_1\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right] + C_2\mathbb{E}\left[\frac{\|X_k-\bar{x}_k\mathbf{1}^\top\|^2}{n}\right] + \alpha^2C_3\mathbb{E}\left[\frac{\|S_k-\bar{s}_k\mathbf{1}^\top\|^2}{n}\right] + C_4)\\
\le&\frac{1+\rho^2}{1-\rho^2}C_1\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right] + \frac{1+\rho^2}{1-\rho^2}C_2\mathbb{E}\left[\frac{\|X_k-\bar{x}_k\mathbf{1}^\top\|^2}{n}\right]\\
&\quad+(\frac{1+\rho^2}{2}+\frac{1+\rho^2}{1-\rho^2}C_3\alpha^2)\mathbb{E}\left[\frac{\|S_k-\bar{s}_k\mathbf{1}^\top\|^2}{n}\right] + \frac{1+\rho^2}{1-\rho^2}C_4.
\end{aligned} \tag{18}$$

Now if we define

$$\begin{aligned}
&\Gamma_k = \left(\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right],\ \mathbb{E}\left[\frac{\|X_k-\bar{x}_k\mathbf{1}^\top\|^2}{n}\right],\ \mathbb{E}\left[\frac{\|S_k-\bar{s}_k\mathbf{1}^\top\|^2}{n}\right]\right)^\top,\\
&c = \left(\frac{\alpha^2}{n}(3\sigma_1^2\|x^*\|^2+\sigma_2^2),\ 0,\ \frac{(1+\rho^2)}{1-\rho^2}C_4\right)^\top,\ M = \begin{pmatrix} M_{11} & M_{12} & 0\\ 0 & M_{22} & M_{23}\\ M_{31} & M_{32} & M_{33}\end{pmatrix},\\
&M_{11} = 1-\alpha\mu,\ M_{12} = (\frac{2\alpha}{\mu}+2\alpha^2)\tilde{\sigma}_1^2,\ M_{22} = \frac{1+\rho^2}{2},\ M_{23} = \alpha^2\frac{1+\rho^2}{1-\rho^2}\\
&M_{31} = \frac{1+\rho^2}{1-\rho^2}C_1,\ M_{32} = \frac{1+\rho^2}{1-\rho^2}C_2,\ M_{33} = \frac{1+\rho^2}{2}+\frac{1+\rho^2}{1-\rho^2}C_3\alpha^2,
\end{aligned}$$

then by (12) we know $\Gamma_{i+1} \le M\Gamma_i + c$ for any i, and thus

$$\Gamma_{k+1} \le M\Gamma_k + c \le ... \le M^{k+1}\Gamma_0 + \sum_{i=0}^{k} M^i c,$$

where all the inequalities are element-wise. By (11) we know there exists an invertible matrix $P \in \mathbb{R}^{3\times3}$ such that $M = P\cdot\mathrm{diag}(\lambda_1,\lambda_2,\lambda_3)P^{-1}$, and $0<\lambda_3<\lambda_2<\lambda_1<1-\frac{2\alpha\mu}{3}$. Without loss of generality we may assume each column of P (as an eigenvector) is a unit vector. Hence we know

$$\|M^k\|_2 = \|P\cdot\mathrm{diag}(\lambda_1^k,\lambda_2^k,\lambda_3^k)P^{-1}\|_2 \le (1-\frac{2\alpha\mu}{3})^k\|P\|_2\|P^{-1}\|_2 = C_M\cdot(1-\frac{2\alpha\mu}{3})^k, \tag{19}$$

where we define $C_M := \|P\|_2\|P^{-1}\|_2$ in the last equality. Note that since we choose P such that each column is a unit vector and $M = P\cdot\mathrm{diag}(\lambda_1,\lambda_2,\lambda_3)P^{-1}$, P is uniquely determined and C_M

is a continuous function of α and other constants ($\sigma_1,\ \sigma_2,\ \mu,\ L,\ \max_i \|b_i\|,\ \|x^*\|,\ n,\ \rho$). On the other hand, observe that

$$\begin{aligned}\|\sum_{i=0}^{k} M^i\|_2 &= \|\sum_{i=0}^{k} P\cdot \mathrm{diag}(\lambda_1^i,\lambda_2^i,\lambda_3^i)P^{-1}\|_2 = \|P\cdot \mathrm{diag}(\sum_{i=0}^{k}\lambda_1^i,\sum_{i=0}^{k}\lambda_2^i,\sum_{i=0}^{k}\lambda_3^i)P^{-1}\|_2 \\ &\le C_M\cdot \max_i \frac{1}{1-\lambda_i} < \frac{3C_M}{2\alpha\mu},\end{aligned} \tag{20}$$

where the last inequality uses the upper bound of the eigenvalues. For (13) we have

$$\begin{aligned}&\max\left(\frac{1}{n}\mathbb{E}\left[\|X_k - x^*\mathbf{1}^\top\|^2\right], \frac{1}{n}\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right]\right) \le \frac{2}{n}\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2 + n\|\bar{x}_k - x^*\|^2\right] \le 2\sqrt{2}\|\Gamma_k\| \\ \le& 2\sqrt{2}\|M^k\Gamma_0 + \sum_{i=0}^{k-1} M^i c\| \le 2\sqrt{2}(\|M^k\|_2\|\Gamma_0\| + \|\sum_{i=1}^{k-1}M^i\|_2\|c\|) \\ \le& 2\sqrt{2}C_M(1-\frac{2\alpha\mu}{3})^k\|\Gamma_0\| + 2\sqrt{2}\cdot\frac{3C_M}{2\alpha\mu}\|c\| \\ \le& 2\sqrt{2}C_M(1-\frac{2\alpha\mu}{3})^k(\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right] + \mathbb{E}\left[\frac{\|X_0-\bar{x}_0\mathbf{1}^\top\|^2}{n}\right] + \mathbb{E}\left[\frac{\|S_0-\bar{s}_0\mathbf{1}^\top\|^2}{n}\right]) + \frac{3\sqrt{2}C_M\|c\|}{\alpha\mu} \\ \le& 3C_M(1-\frac{2\alpha\mu}{3})^k(\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right] + \mathbb{E}\left[\frac{\|X_0\|^2+\|S_0\|^2}{n}\right]) + \frac{5C_M\|c\|}{\alpha\mu},\end{aligned}$$

where the fifth inequality uses (19) and (20), and the seventh inequality uses the fact that $\|X_0 - \bar{x}_0\mathbf{1}^\top\| = \|X_0(I-\frac{\mathbf{1}\mathbf{1}^\top}{n})\| \le \|X_0\|$ (same for S_0). (14) can be viewed as a corollary of the above inequality by noticing that

$$\begin{aligned}&\frac{1}{n}\mathbb{E}\left[\|X_k\|^2\right] \le \frac{2}{n}\mathbb{E}\left[\|X_k - x^*\mathbf{1}^\top\|^2 + n\|x^*\|^2\right] \\ \le& 6C_M(1-\frac{2\alpha\mu}{3})^k(\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right] + \mathbb{E}\left[\frac{\|X_0\|^2+\|S_0\|^2}{n}\right]) + \frac{10C_M\|c\|}{\alpha\mu} + 2\|x^*\|^2.\end{aligned}$$

□

Remark:

- Lemma A.4 characterizes convergence results of decentralized stochastic gradient descent with gradient tracking on strongly convex quadratic functions. Moreover, it also indicates that the second moment of X_k can be bounded by using (14), which will be used in proving the boundedness of second moment of $Z_t^{(k)}$ of our HIGP oracle.
- If we consider the same updates under the deterministic setting, then $\sigma_1 = \sigma_2 = 0$ and thus $\|c\| = 0$ by definition, which indicates the constant term in (13) vanishes (i.e., linear convergence). We will utilize this important observation in the next lemma.

Lemma A.5. *Suppose Assumptions 2.1, 2.2 and 2.4 hold. In Algorithm 1 define*

$$\begin{aligned}C_1 &= 9\sigma_{g,2}^2 + 6\gamma^2(\sigma_{g,2}^2+L_{g,1}^2) + \frac{18\gamma^2\sigma_{g,2}^2(\sigma_{g,2}^2+L_{g,1}^2)}{n}, \\ C_2 &= 12(\sigma_{g,2}^2+L_{g,1}^2) + 9\sigma_{g,2}^2 + 12\gamma^2L_{g,1}^2(\sigma_{g,2}^2+L_{g,1}^2) + \frac{18\gamma^2\sigma_{g,2}^2(\sigma_{g,2}^2+L_{g,1}^2)}{n},\end{aligned}$$

$$C_3 = 6\rho^2(\sigma_{g,2}^2 + L_{g,1}^2),\ C_4 = 2\sigma_f^2 + \frac{6\gamma^2\sigma_f^2(\sigma_{g,2}^2 + L_{g,1}^2)}{n} + (9\sigma_{g,2}^2 + \frac{18\gamma^2\sigma_{g,2}^2(\sigma_{g,2}^2 + L_{g,1}^2)}{n})\|x^*\|^2,$$
$$c = \left(\frac{\gamma^2}{n}(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2} + \sigma_f^2),\ 0,\ \frac{(1+\rho^2)}{1-\rho^2}C_4\right)^\top,\ M = \begin{pmatrix} M_{11} & M_{12} & 0 \\ 0 & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{pmatrix},$$
$$M_{11} = 1-\gamma\mu_g,\ M_{12} = (\frac{2\gamma}{\mu_g} + 2\gamma^2)(\sigma_{g,2}^2 + L_{g,1}^2),\ M_{22} = \frac{1+\rho^2}{2},\ M_{23} = \gamma^2\frac{1+\rho^2}{1-\rho^2},$$
$$M_{31} = \frac{1+\rho^2}{1-\rho^2}C_1,\ M_{32} = \frac{1+\rho^2}{1-\rho^2}C_2,\ M_{33} = \frac{1+\rho^2}{2} + \frac{1+\rho^2}{1-\rho^2}C_3\gamma^2.$$

Define $\tilde{M}$ to be matrix M and $C_{\tilde{M}}$ to be C_M when $\sigma_{g,2} = \sigma_f = 0$. If γ satisfies

$$\begin{aligned} &(1+\frac{\gamma\mu_g}{2})(1-\gamma\mu_g)^2 + \frac{3\gamma^2\sigma_{g,2}^2}{n} < 1-\gamma\mu_g,\ 0<\gamma_1\le\gamma\le\gamma_2 \text{ for some } 0<\gamma_1<\gamma_2, \\ &\max\left(\rho(\tilde{M}), \rho(M)\right) < 1 - \frac{2\gamma\mu_g}{3},\ \text{both } M \text{ and } \tilde{M} \text{have 3 different positive eigenvalues,} \end{aligned} \tag{21}$$

then for any $0 \le t \le N$ we have

$$\mathbb{E}\left[\|\bar{z}_t^{(k)}\|^2|\mathcal{F}_k\right] \le \frac{1}{n}\mathbb{E}\left[\|Z_t^{(k)}\|^2|\mathcal{F}_k\right] \le \sigma_z^2 := 6C_M(\frac{L_{f,0}^2}{\mu_g^2} + \sigma_f^2 + L_{f,0}^2) + \frac{10C_M\|c\|}{\gamma\mu_g} + \frac{2L_{f,0}^2}{\mu_g^2}, \tag{22}$$
$$\frac{1}{n}\|\mathbb{E}\left[Z_t^{(k)} - \bar{z}_t^{(k)}\mathbf{1}^\top|\mathcal{F}_k\right]\|^2 \le 3C_{\tilde{M}}(1-\frac{2\gamma\mu_g}{3})^t(\frac{L_{f,0}^2}{\mu_g^2} + L_{f,0}^2). \tag{23}$$

Proof of Lemma A.5. Note that (22) is a direct results of Lemma A.4 by noticing that

$$\begin{aligned} &z_{i,t+1}^{(k)} = \sum_{j=1}^n w_{ij}z_{j,t}^{(k)} - \gamma d_{i,t}^{(k)},\ Z_0^{(k)} = 0, \\ &d_{i,t+1}^{(k)} = \sum_{i=1}^n w_{ij}d_{j,t}^{(k)} + s_{i,t+1}^{(k)} - s_{i,t}^{(k)},\ s_{i,t}^{(k)} = H_{i,t}^{(k)}z_{i,t}^{(k)} - b_{i,t}^{(k)}, \\ &\mathbb{E}\left[H_{i,t}^{(k)}|\mathcal{F}_k\right] = \nabla_y^2 g_i(x_{i,k}, y_{i,k}^{(T)}),\ \mathbb{E}\left[\|H_{i,t}^{(k)} - \nabla_y^2 g_i(x_{i,k}, y_{i,k}^{(T)})\|^2|\mathcal{F}_k\right] \le \sigma_{g,2}^2, \\ &\mathbb{E}\left[b_{i,t}^{(k)}|\mathcal{F}_k\right] = \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}),\ \mathbb{E}\left[\|b_{i,t}^{(k)} - \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)})\|^2|\mathcal{F}_k\right] \le \sigma_f^2, \end{aligned}$$

for any $k \ge 0, 1 \le i \le n$, and $t \ge 0$. Assumption 2.1 also indicates that

$$\mu_g I \preceq \nabla_y^2 g_i(x_{i,k}, y_{i,k}^{(T)}) \preceq L_{g,1}I,\ \|\nabla_y f_i(x_{i,k}, y_{i,k}^{(T)})\| \le L_{f,0}.$$

Hence we know by (14),

$$\mathbb{E}\left[\|\bar{z}_t^{(k)}\|^2|\mathcal{F}_k\right] \le \frac{1}{n}\mathbb{E}\left[\|Z_t^{(k)}\|^2|\mathcal{F}_k\right] \le 6C_M(1-\frac{2\gamma\mu_g}{3})^k(\frac{L_{f,0}^2}{\mu_g^2} + \sigma_f^2 + L_{f,0}^2) + \frac{10C_M\|c\|}{\gamma\mu_g} + \frac{2L_{f,0}^2}{\mu_g^2} \le \sigma_z^2,$$

which proves (22). To prove (23), we notice that in expectation, the updates can be written as

$$\mathbb{E}\left[z_{i,t+1}^{(k)}|\mathcal{F}_k\right] = \sum_{j=1}^n w_{ij}\mathbb{E}\left[z_{j,t}^{(k)}|\mathcal{F}_k\right] - \gamma\mathbb{E}\left[d_{i,t}^{(k)}|\mathcal{F}_k\right],\ Z_0^{(k)} = 0,$$

$$\mathbb{E}\left[d_{i,t+1}^{(k)}|\mathcal{F}_k\right] = \sum_{i=1}^{n} w_{ij}\mathbb{E}\left[d_{j,t}^{(k)}|\mathcal{F}_k\right] + \mathbb{E}\left[s_{i,t+1}^{(k)}|\mathcal{F}_k\right] - \mathbb{E}\left[s_{i,t}^{(k)}|\mathcal{F}_k\right],$$
$$\mathbb{E}\left[s_{i,t}^{(k)}|\mathcal{F}_k\right] = \nabla_y^2 g_i(x_{i,k}, y_{i,k}^{(T)})\mathbb{E}\left[z_{i,t}^{(k)}|\mathcal{F}_k\right] - \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}).$$

The updates of $\mathbb{E}\left[z_{i,t}^{(k)}|\mathcal{F}_k\right]$ can be viewed as a noiseless case (i.e., $\sigma_{g,2} = \sigma_f = 0$) of Lemma A.4. Using this observation, (13), and the definition of $\|c\|$ and $\tilde{M}$, we know (23) holds. □

Now we provide a technical lemma that guarantees (11) and (21). For simplicity we can just consider (11).

Lemma A.6. *Let M be the matrix defined in Lemma A.4. There exist $0 < \alpha_1 < \alpha_2$ such that $\alpha \in (\alpha_1, \alpha_2)$ and*

$$(1 + \frac{\alpha\mu}{2})(1 - \alpha\mu)^2 + \frac{3\alpha^2\sigma_1^2}{n} < 1 - \alpha\mu, \tag{24}$$
$$\rho(M) < 1 - \frac{2\alpha\mu}{3}, \quad \textit{and} \quad M \textit{ has 3 different positive eigenvalues.} \tag{25}$$

Proof of Lemma A.6. Note that (24) is equivalent to

$$\mu^3\alpha^2 + \frac{6\alpha\sigma_1^2}{n} - \mu < 0,$$

which implies any $\alpha_1,\ \alpha_2$ satisfying

$$0 < \alpha_1 < \alpha_2 < \frac{\sqrt{9\sigma_1^4 + n^2\mu^4} - 3\sigma_1^2}{n\mu^3} \tag{26}$$

will ensure (24). Next we consider (25). Define

$$\varphi(\lambda) := \det(\lambda I - M) = \prod_{i=1}^{3}(\lambda - M_{ii}) - M_{23}M_{32}(\lambda - M_{11}) - M_{12}M_{23}M_{31}.$$

We know that a sufficient condition to guarantee (25) is

$$\varphi(1 - \frac{2\alpha\mu}{3}) > 0,\ \varphi(M_{11}) < 0,\ \varphi(M_{22}) > 0,\ \varphi(0) < 0,\ M_{11} > M_{22}, \tag{27}$$

since this implies $0 < M_{22} < M_{11} = 1 - \alpha\mu < 1 - \frac{2\alpha\mu}{3}$ and

$$\varphi(1 - \frac{2\alpha\mu}{3}) \cdot \varphi(M_{11}) < 0,\ \varphi(M_{11}) \cdot \varphi(M_{22}) < 0,\ \varphi(M_{22}) \cdot \varphi(0) < 0,$$

which together with continuity of φ indicate the roots of $\varphi(\lambda) = 0$ (i.e., the eigenvalues of M, denoted as $\lambda_1, \lambda_2,\ \lambda_3$ in descending order) satisfy

$$0 < \lambda_3 < M_{22} < \lambda_2 < M_{11} < \lambda_1 < 1 - \frac{2\alpha\mu}{3}.$$

The condition $\varphi(M_{11}) < 0$ is automatically true by definition of φ and M, and for the rest of the conditions in (27) we have

$$\varphi(1 - \frac{2\alpha\mu}{3}) > 0$$

$$\Leftrightarrow \alpha \cdot \varphi_1(\alpha)$$
$$:= \frac{\alpha\mu}{3}\left[(\frac{1-\rho^2}{2} - \frac{2\alpha\mu}{3})(\frac{1-\rho^2}{2} - \frac{2\alpha\mu}{3} - \frac{1+\rho^2}{1-\rho^2}C_3\alpha^2) - (\frac{1+\rho^2}{1-\rho^2})^2 C_2\alpha^2\right] - (\frac{1+\rho^2}{1-\rho^2})^2 C_1\alpha^2(\frac{2\alpha}{\mu} + 2\alpha^2)\tilde{\sigma}_1^2 > 0,$$
$$\varphi(M_{22}) > 0 \Leftrightarrow M_{23}((M_{11} - M_{22})M_{32} - M_{12}M_{31}) > 0 \Leftrightarrow (M_{11} - M_{22})M_{32} - M_{12}M_{31} > 0$$
$$\Leftrightarrow \varphi_2(\alpha) := (\frac{1-\rho^2}{2} - \alpha\mu)\frac{1+\rho^2}{1-\rho^2}C_2 - \frac{1+\rho^2}{1-\rho^2}C_1(\frac{2\alpha}{\mu} + 2\alpha^2)\tilde{\sigma}_1^2 > 0, \text{ (by definition of } C_2, C_2 > 0 \text{ when } \alpha = 0)$$
$$\varphi(0) < 0 \Leftrightarrow -M_{11}(M_{22}M_{33} - M_{23}M_{32}) - M_{12}M_{23}M_{31} < 0 \Leftarrow M_{22}M_{33} - M_{23}M_{32} > 0$$
$$\Leftrightarrow \varphi_3(\alpha) := \frac{1+\rho^2}{2}(\frac{1+\rho^2}{2} + \frac{1+\rho^2}{1-\rho^2}C_3\alpha^2) - (\frac{1+\rho^2}{1-\rho^2})^2 C_2\alpha^2 > 0,$$
$$M_{11} > M_{22} \Leftrightarrow \alpha < \frac{1-\rho^2}{2\mu}.$$

Hence a sufficient condition for (27) is

$$\varphi_1(\alpha) > 0,\ \varphi_2(\alpha) > 0,\ \varphi_3(\alpha) > 0,\ \alpha < \frac{1-\rho^2}{2\mu}.$$

Given the expressions of $\varphi_i(\alpha)$ above, we know they satisfy $\varphi_i(0) > 0$. Hence we can define β to be the minimum positive constant such that $\varphi_1(\beta)\varphi_2(\beta)\varphi_3(\beta) = 0$, and

$$\alpha_2 = \min\left(\frac{\sqrt{9\sigma_1^4 + n^2\mu^4} - 3\sigma_1^2}{n\mu^3}, \frac{1-\rho^2}{2\mu}, \beta\right),\ \alpha_1 = \text{ any constant in } (0, \alpha_2),$$

which implies that for any $\alpha \in (\alpha_1, \alpha_2)$, we always have

$$\varphi_1(\alpha) > 0,\ \varphi_2(\alpha) > 0,\ \varphi_3(\alpha) > 0,\ \alpha < \frac{\sqrt{9\sigma_1^4 + n^2\mu^4} - 3\sigma_1^2}{n\mu^3},\ \alpha < \frac{1-\rho^2}{2\mu},$$

because of the definition of β, and $\varphi_i(0) = 0$ for all $1 \le i \le 3$. (26). The above expression implies (26) and (27), and hence (24) and (25) are satisfied. □

Remark:

- One can follow the proof of Corollary 1 in Pu and Nedić (2021) to obtain an explicit dependence between α_1, α_2 and other parameters, which is purely technical and we omit it in this lemma.
- Define $\tilde{\alpha}_2$ to be the constant α_2 when $\sigma_1 = \sigma_2 = 0$ in the above lemma. We can check that the proof is still valid and thus for any $\alpha \in (\frac{\min(\alpha_2, \tilde{\alpha}_2)}{2}, \min(\alpha_2, \tilde{\alpha}_2))$ we have

  $$(1 + \frac{\alpha\mu}{2})(1 - \alpha\mu)^2 + \frac{3\alpha^2\sigma_1^2}{n} < 1 - \alpha\mu,$$
  $$\max\left(\rho(\tilde{M}), \rho(M)\right) < 1 - \frac{2\alpha\mu}{3},\ \text{both } M \text{ and } \tilde{M} \text{ have 3 different positive eigenvalues,}$$

  and thus the existence of γ_1 and γ_2 in (21) is also guaranteed.

Using Lemma A.5 we could directly bound $\|X_k - \bar{x}_k\mathbf{1}^\top\|^2$ and $\|Y_k^{(t+1)} - \bar{y}_k^{(t+1)}\mathbf{1}^\top\|^2$.

Lemma A.7. *Suppose Assumptions 2.1, 2.2, 2.3, and 2.4 hold. Define*

$$\sigma_u^2 = 2(L_{f,0}^2 + \sigma_f^2) + 2(L_{g,1}^2 + \sigma_{g,2}^2)\sigma_z^2,\ \sigma_x^2 = \frac{1+\rho^2}{1-\rho^2}\cdot\sigma_u^2,\ \tilde{\alpha}_{k+1}^2 = \sum_{i=0}^{k}\alpha_i^2\left(\frac{1+\rho^2}{2}\right)^{k-i},$$
$$\tilde{\beta}_{k+1}^2 = \frac{1+\rho^2}{1-\rho^2}\sum_{i=0}^{k}\beta_i^2(2\sigma_{g,1}^2 + 6L_{g,1}^2\sigma_x^2\tilde{\alpha}_i^2 + 3\delta^2)\left(\frac{3+\rho^2}{4}\right)^{k-i},\ \tilde{\alpha}_0 = \tilde{\beta}_0 = 0.$$

If β_k satisfy

$$\frac{(1+\rho^2)}{2} + \beta_k^2\frac{1+\rho^2}{1-\rho^2}\cdot 6L_{g,1}^2 \le \frac{3+\rho^2}{4} < 1, \tag{28}$$

then in Algorithm 3, for any $k \ge 0$ and $0 \le t \le T-1$ we have

$$\begin{aligned}&\mathbb{E}\left[\|U_k\|^2\right] \le n\sigma_u^2,\ \mathbb{E}\left[\|X_k - \bar{x}_k\mathbf{1}^\top\|^2\right] \le n\sigma_x^2\tilde{\alpha}_k^2,\\ &\frac{1}{n}\mathbb{E}\left[\|Y_k^{(t)} - \bar{y}_k^{(t)}\mathbf{1}^\top\|^2\right] \le \left[\left(\frac{3+\rho^2}{4}\right)^t T - t\left(\frac{3+\rho^2}{4}\right)\right]\tilde{\beta}_k^2 + t\tilde{\beta}_{k+1}^2.\end{aligned} \tag{29}$$

Proof of Lemma A.7. Note that the inner and outer loop updates satisfy

$$\begin{aligned}&\bar{x}_{k+1} = \bar{x}_k - \alpha_k\bar{u}_k,\ X_{k+1} - \bar{x}_{k+1}\mathbf{1}^\top = X_kW - \bar{x}_k\mathbf{1}^\top - \alpha_k(R_k - \bar{r}_k\mathbf{1}^\top),\\ &\bar{y}_k^{(t+1)} = \bar{y}_k^{(t)} - \beta_k\bar{v}_k^{(t)},\ Y_k^{(t+1)} - \bar{y}_k^{(t+1)} = Y_k^{(t)}W - \bar{y}_k^{(t)}\mathbf{1}^\top - \beta_k(V_k^{(t)} - \bar{v}_k^{(t)}\mathbf{1}^\top),\end{aligned}$$

which gives

$$\|X_{k+1} - \bar{x}_{k+1}\mathbf{1}^\top\|^2 \le \frac{(1+\rho^2)}{2}\|X_k - \bar{x}_k\mathbf{1}^\top\|^2 + \alpha_k^2\frac{1+\rho^2}{1-\rho^2}\|R_k - \bar{r}_k\mathbf{1}^\top\|^2, \tag{30}$$
$$\|Y_k^{(t+1)} - \bar{y}_k^{(t+1)}\mathbf{1}^\top\|^2 \le \frac{(1+\rho^2)}{2}\|Y_k^{(t)} - \bar{y}_k^{(t)}\mathbf{1}^\top\|^2 + \beta_k^2\frac{1+\rho^2}{1-\rho^2}\|V_k^{(t)} - \bar{v}_k^{(t)}\mathbf{1}^\top\|^2. \tag{31}$$

The inequalities hold similarly as the inequality in (16). Notice that we have

$$\begin{aligned}&\|R_k - \bar{r}_k\mathbf{1}^\top\| = \|R_k(I - \frac{\mathbf{1}\mathbf{1}^\top}{n})\| = \|\left[(1-\alpha_k)R_{k-1} + \alpha_kU_{k-1}\right](I - \frac{\mathbf{1}\mathbf{1}^\top}{n})\|\\ \le &\max\left(\|R_{k-1}(I - \frac{\mathbf{1}\mathbf{1}^\top}{n})\|,\ \|U_{k-1}(I - \frac{\mathbf{1}\mathbf{1}^\top}{n})\|\right) \le \max_{0\le i\le k-1}\left(\|U_i(I - \frac{\mathbf{1}\mathbf{1}^\top}{n})\|\right).\end{aligned}$$

The second inequality holds by repeating the first inequality multiple times. For each $\|U_k - \bar{u}_k\mathbf{1}^\top\|$ we have

$$\begin{aligned}&\mathbb{E}\left[\|U_k - \bar{u}_k\mathbf{1}^\top\|^2\right] = \mathbb{E}\left[\|U_k(I - \frac{\mathbf{1}\mathbf{1}^\top}{n})\|^2\right] \le \mathbb{E}\left[\|U_k\|^2\right] = \sum_{i=1}^{n}\mathbb{E}\left[\|u_{i,k}\|^2\right]\\ \le &2\sum_{i=1}^{n}\left(\mathbb{E}\left[\|\nabla_x f_i(x_{i,k}, y_{i,k}^{(T)};\phi_{i,0})\|^2\right] + \mathbb{E}\left[\|\nabla_{xy}^2 g_i(x_{i,k}, y_{i,k}^{(T)};\xi_{i,0})z_{i,N}^{(k)}\|^2\right]\right)\\ \le &2\sum_{i=1}^{n}\left(L_{f,0}^2 + \sigma_f^2 + (L_{g,1}^2 + \sigma_{g,2}^2)\mathbb{E}\left[\|z_{i,N}^{(k)}\|^2\right]\right) \le 2n(L_{f,0}^2 + \sigma_f^2) + 2n(L_{g,1}^2 + \sigma_{g,2}^2)\sigma_z^2 = n\sigma_u^2.\end{aligned}$$

The fourth inequality uses (22). Using the above two inequaities in (30) we know

$$\|X_{k+1} - \bar{x}_{k+1}\mathbf{1}^\top\|^2 \leq \frac{(1+\rho^2)}{2}\|X_k - \bar{x}_k\mathbf{1}^\top\|^2 + n\alpha_k^2\sigma_x^2.$$

Using Lemma A.2 and $X_0 = 0$, we can obtain the first two results of (29). To analyze $\|V_k^{(t)} - \bar{v}_k^{(t)}\mathbf{1}^\top\|$, we first notice that

$$\begin{aligned} v_{i,k}^{(t)} - \bar{v}_k^{(t)} = v_{i,k}^{(t)} - \nabla_y g_i(x_{i,k}, y_{i,k}^{(t)}) - (\bar{v}_k^{(t)} - \frac{1}{n}\sum_{l=1}^n \nabla_y g_l(x_{l,k}, y_{l,k}^{(t)})) + \nabla_y g_i(x_{i,k}, y_{i,k}^{(t)}) - \nabla_y g_i(\bar{x}_k, \bar{y}_k^{(t)}) \\ - \frac{1}{n}\sum_{l=1}^n (\nabla_y g_l(x_{l,k}, y_{l,k}^{(t)}) - \nabla_y g_l(\bar{x}_k, \bar{y}_k^{(t)})) + \nabla_y g_i(\bar{x}_k, \bar{y}_k^{(t)}) - \frac{1}{n}\sum_{l=1}^n \nabla_y g_l(\bar{x}_k, \bar{y}_k^{(t)}). \end{aligned}$$

Hence we know

$$\begin{aligned} &\mathbb{E}\left[\|V_k^{(t)} - \bar{v}_k^{(t)}\mathbf{1}^\top\|^2\right] = \sum_{i=1}^n \mathbb{E}\left[\|v_{i,k}^{(t)} - \bar{v}_k^{(t)}\|^2\right] \\ \leq &(n+1)\sigma_{g,1}^2 + 3\sum_{i=1}^n \mathbb{E}\left[L_{g,1}^2(\|x_{i,k} - \bar{x}_k\|^2 + \|y_{i,k}^{(t)} - \bar{y}_k^{(t)}\|^2) + \frac{L_{g,1}^2}{n}\sum_{l=1}^n(\|x_{l,k} - \bar{x}_k\|^2 + \|y_{l,k}^{(t)} - \bar{y}_k^{(t)}\|^2) + \delta^2\right] \\ = &(n+1)\sigma_{g,1}^2 + 6L_{g,1}^2\mathbb{E}\left[\|X_k - \bar{x}_k\mathbf{1}^\top\|^2 + \|Y_k - \bar{y}_k^{(t)}\mathbf{1}^\top\|^2\right] + 3n\delta^2 \\ \leq &6L_{g,1}^2\mathbb{E}\left[\|Y_k^{(t)} - \bar{y}_k^{(t)}\mathbf{1}^\top\|^2\right] + 2n\sigma_{g,1}^2 + 6nL_{g,1}^2\sigma_x^2\tilde{\alpha}_k^2 + 3n\delta^2, \end{aligned}$$

where the second inequality uses the first result of (29). The above inequality together with (31) imply

$$\begin{aligned} &\frac{1}{n}\mathbb{E}\left[\|Y_k^{(t+1)} - \bar{y}_k^{(t+1)}\mathbf{1}^\top\|^2\right] \\ \leq &\left[\frac{(1+\rho^2)}{2} + \beta_k^2\frac{1+\rho^2}{1-\rho^2}\cdot 6L_{g,1}^2\right]\cdot\frac{1}{n}\mathbb{E}\left[\|Y_k^{(t)} - \bar{y}_k^{(t)}\mathbf{1}^\top\|^2\right] + \beta_k^2\frac{1+\rho^2}{1-\rho^2}(2\sigma_{g,1}^2 + 6L_{g,1}^2\sigma_x^2\tilde{\alpha}_k^2 + 3\delta^2) \\ \leq &\left(\frac{3+\rho^2}{4}\right)^{t+1}\cdot\frac{1}{n}\mathbb{E}\left[\|Y_k^{(0)} - \bar{y}_k^{(0)}\mathbf{1}^\top\|^2\right] + \beta_k^2\frac{1+\rho^2}{1-\rho^2}(2\sigma_{g,1}^2 + 6L_{g,1}^2\sigma_x^2\tilde{\alpha}_k^2 + 3\delta^2)\sum_{l=0}^t\left(\frac{3+\rho^2}{4}\right)^i \\ \leq &\left(\frac{3+\rho^2}{4}\right)^{t+1}\cdot\frac{1}{n}\mathbb{E}\left[\|Y_k^{(0)} - \bar{y}_k^{(0)}\mathbf{1}^\top\|^2\right] + (t+1)\beta_k^2\frac{1+\rho^2}{1-\rho^2}(2\sigma_{g,1}^2 + 6L_{g,1}^2\sigma_x^2\tilde{\alpha}_k^2 + 3\delta^2), \end{aligned} \tag{32}$$

where the second inequality uses Lemma A.2 and (28). Notice that we use warm-start strategy (i.e., $Y_{k+1}^{(0)} = Y_k^{(T)}$), hence we know

$$\begin{aligned} &\frac{1}{n}\mathbb{E}\left[\|Y_{k+1}^{(0)} - \bar{y}_{k+1}^{(0)}\mathbf{1}^\top\|^2\right] = \frac{1}{n}\mathbb{E}\left[\|Y_k^{(T)} - \bar{y}_k^{(T)}\mathbf{1}^\top\|^2\right] \\ \leq &\left(\frac{3+\rho^2}{4}\right)^T\cdot\frac{1}{n}\mathbb{E}\left[\|Y_k^{(0)} - \bar{y}_k^{(0)}\mathbf{1}^\top\|^2\right] + T\beta_k^2\frac{1+\rho^2}{1-\rho^2}(2\sigma_{g,1}^2 + 6L_{g,1}^2\sigma_x^2\tilde{\alpha}_k^2 + 3\delta^2) \\ \leq &T\frac{1+\rho^2}{1-\rho^2}\sum_{i=0}^k\beta_i^2(2\sigma_{g,1}^2 + 6L_{g,1}^2\sigma_x^2\tilde{\alpha}_i^2 + 3\delta^2)\left(\frac{3+\rho^2}{4}\right)^{k-i} = T\tilde{\beta}_{k+1}^2, \end{aligned}$$

where the second inequality uses Lemma A.2. Using the above estimation in (32), we know

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}\left[\|Y_k^{(t+1)}-\bar{y}_k^{(t+1)}\mathbf{1}^\top\|^2\right]\\
\leq&\left(\frac{3+\rho^2}{4}\right)^{t+1}\cdot\frac{1}{n}\mathbb{E}\left[\|Y_k^{(0)}-\bar{y}_k^{(0)}\mathbf{1}^\top\|^2\right]+(t+1)\beta_k^2\frac{1+\rho^2}{1-\rho^2}(2\sigma_{g,1}^2+6L_{g,1}^2\sigma_x^2\tilde{\alpha}_k^2+3\delta^2)\\
\leq&\left(\frac{3+\rho^2}{4}\right)^{t+1}T\tilde{\beta}_k^2+(t+1)(\tilde{\beta}_{k+1}^2-\left(\frac{3+\rho^2}{4}\right)\tilde{\beta}_k^2),
\end{aligned}$$

and thus the proof is complete by rearranging the terms. □

Now we are ready to analyze the convergence of the inner loop of Algorithm 3.

Lemma A.8. *Suppose Assumptions 2.1 and 2.4 hold. For any $0\leq t\leq T-1$ define*

$$C_{k,t+1}=\sum_{l=0}^{t}\left[(\frac{\beta_k}{\mu_g}+\beta_k^2)L_{g,1}^2\left(\sigma_x^2\tilde{\alpha}_k^2+\left[\left(\frac{3+\rho^2}{4}\right)^l T-l\left(\frac{3+\rho^2}{4}\right)\right]\tilde{\beta}_k^2+l\tilde{\beta}_{k+1}^2\right)+\frac{\beta_k^2\sigma_{g,1}^2}{n}\right]. \quad (33)$$

If $T\geq 2$ and $0<\beta_k\leq\min\{1,\frac{1}{\mu_g}\}$, then in Algorithm 3, we have

$$\mu_g\sum_{k=1}^{K}\beta_k\mathbb{E}\left[\|\bar{y}_k^{(0)}-y_{k-1}^*\|^2\right]\leq\mathbb{E}\left[\|\bar{y}_1^{(0)}-y_0^*\|^2\right]+L_{y^*}^2\sum_{k=1}^{K}(\frac{\alpha_{k-1}^2}{\beta_k\mu_g}+\alpha_{k-1}^2)\mathbb{E}\left[\|\bar{r}_k\|^2\right]+\sum_{k=1}^{K}C_{k,T}, \quad (34)$$

where $y_k^*=y^*(\bar{x}_k)=\arg\min_y\sum_{i=1}^n g_i(\bar{x}_k,y)$

Proof of Lemma A.8. For any $k\geq 0,\ 1\leq t\leq T-1$, define

$$\mathcal{G}_t^{(k)}=\sigma(\bigcup_{i=1}^{n}\{y_{i,0}^{(T)},...,y_{i,k-1}^{(T)},y_{i,k}^{(t)},x_{i,0},...,x_{i,k},r_{i,0},...,r_{i,k}\}).$$

We know

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{y}_k^{(t+1)}-y_k^*\|^2|\mathcal{G}_t\right]\\
=&\mathbb{E}\left[\|\bar{y}_k^{(t)}-\beta_k\nabla_y g(\bar{x}_k,\bar{y}_k^{(t)})-y_k^*-\beta_k(\bar{v}_k^{(t)}-\mathbb{E}\left[\bar{v}_k^{(t)}|\mathcal{G}_t\right])-\beta_k(\mathbb{E}\left[\bar{v}_k^{(t)}|\mathcal{G}_t\right]-\nabla_y g(\bar{x}_k,\bar{y}_k^{(t)}))\|^2|\mathcal{G}_t\right]\\
=&\mathbb{E}\left[\|\bar{y}_k^{(t)}-\beta_k\nabla_y g(\bar{x}_k,\bar{y}_k^{(t)})-y_k^*-\beta_k(\mathbb{E}\left[\bar{v}_k^{(t)}|\mathcal{G}_t\right]-\nabla_y g(\bar{x}_k,\bar{y}_k^{(t)}))\|^2|\mathcal{G}_t\right]+\frac{\beta_k^2\sigma_{g,1}^2}{n}\\
\leq&(1+\beta_k\mu_g)\|\bar{y}_k^{(t)}-\beta_k\nabla_y g(\bar{x}_k,\bar{y}_k^{(t)})-y_k^*\|^2+(1+\frac{1}{\beta_k\mu_g})\beta_k^2\mathbb{E}\left[\|\mathbb{E}\left[\bar{v}_k^{(t)}|\mathcal{G}_t\right]-\nabla_y g(\bar{x}_k,\bar{y}_k^{(t)})\|^2|\mathcal{G}_t\right]+\frac{\beta_k^2\sigma_{g,1}^2}{n}\\
\leq&(1+\beta_k\mu_g)(1-\beta_k\mu_g)^2\|\bar{y}_k^{(t)}-y_k^*\|^2+(\frac{\beta_k}{\mu_g}+\beta_k^2)\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla_y g_i(x_{i,k},y_{i,k}^{(t)})-\nabla_y g_i(\bar{x}_k,\bar{y}_k^{(t)})\right)\|^2+\frac{\beta_k^2\sigma_{g,1}^2}{n}\\
\leq&(1-\beta_k\mu_g)\|\bar{y}_k^{(t)}-y_k^*\|^2+\frac{(\frac{\beta_k}{\mu_g}+\beta_k^2)L_{g,1}^2}{n}\left(\|X_k-\bar{x}_k\mathbf{1}^\top\|^2+\|Y_k^{(t)}-\bar{y}_k^{(t)}\mathbf{1}^\top\|^2\right)+\frac{\beta_k^2\sigma_{g,1}^2}{n},
\end{aligned} \quad (35)$$

where the second equality holds since $\bar{v}_k^{(t)}-\mathbb{E}\left[\bar{v}_k^{(t)}|\mathcal{G}_t\right]$ has expectation 0 and

$$\mathbb{E}\left[\|\bar{v}_k^{(t)}-\mathbb{E}\left[\bar{v}_k^{(t)}|\mathcal{G}_t\right]\|^2|\mathcal{G}_t\right]=\mathbb{E}\left[\|\frac{1}{n}\sum_{i=1}^{n}\left(v_{i,k}^{(t)}-\mathbb{E}\left[v_{i,k}^{(t)}|\mathcal{G}_t\right]\right)\|^2|\mathcal{G}_t\right]\leq\frac{\sigma_{g,1}^2}{n},$$

due to independence, the second inequality holds due to Lemma A.3 and $\beta_k \leq 1$, and the third inequality holds due to Lipschitz continuity of $\nabla_y g$. Taking expectation on both sides and using (29) we know

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{y}_k^{(t+1)} - y_k^*\|^2\right] \\
\leq&(1-\beta_k\mu_g)\mathbb{E}\left[\|\bar{y}_k^{(t)} - y_k^*\|^2\right] + (\frac{\beta_k}{\mu_g} + \beta_k^2)L_{g,1}^2\left(\sigma_x^2\tilde{\alpha}_k^2 + \left[\left(\frac{3+\rho^2}{4}\right)^t T - t\left(\frac{3+\rho^2}{4}\right)\right]\tilde{\beta}_k^2 + t\tilde{\beta}_{k+1}^2\right) + \frac{\beta_k^2\sigma_{g,1}^2}{n} \\
\leq&(1-\beta_k\mu_g)^{t+1}\mathbb{E}\left[\|\bar{y}_k^{(0)} - y_k^*\|^2\right] + C_{k,t+1},
\end{aligned}$$

where the second inequality uses Lemma A.2. Observe that we also have

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{y}_{k+1}^{(0)} - y_k^*\|^2\right] = \mathbb{E}\left[\|\bar{y}_k^{(T)} - y_k^*\|^2\right] \leq (1-\beta_k\mu_g)^T\mathbb{E}\left[\|\bar{y}_k^{(0)} - y_k^*\|^2\right] + C_{k,T} \\
\leq&(1-\beta_k\mu_g)^T\mathbb{E}\left[(1+\beta_k\mu_g)\|\bar{y}_k^{(0)} - y_{k-1}^*\|^2 + (1+\frac{1}{\beta_k\mu_g})\|y_{k-1}^* - y_k^*\|^2\right] + C_{k,T} \\
\leq&(1+\beta_k\mu_g)(1-\beta_k\mu_g)^T\mathbb{E}\left[\|\bar{y}_k^{(0)} - y_{k-1}^*\|^2\right] + (\frac{\alpha_{k-1}^2}{\beta_k\mu_g} + \alpha_{k-1}^2)L_{y^*}^2\mathbb{E}\left[\|\bar{r}_{k-1}\|^2\right] + C_{k,T} \\
\leq&(1-\beta_k\mu_g)\mathbb{E}\left[\|\bar{y}_k^{(0)} - y_{k-1}^*\|^2\right] + (\frac{\alpha_{k-1}^2}{\beta_k\mu_g} + \alpha_{k-1}^2)L_{y^*}^2\mathbb{E}\left[\|\bar{r}_k\|^2\right] + C_{k,T},
\end{aligned} \tag{36}$$

where the third inequality holds since $T \geq 2$ and $y^*(x)$ is L_{y^*}-smooth. This implies

$$\beta_k\mu_g\mathbb{E}\left[\|\bar{y}_k^{(0)} - y_{k-1}^*\|^2\right] \leq \mathbb{E}\left[\|\bar{y}_k^{(0)} - y_{k-1}^*\|^2\right] - \mathbb{E}\left[\|\bar{y}_{k+1}^{(0)} - y_k^*\|^2\right] + (\frac{\alpha_{k-1}^2}{\beta_k\mu_g} + \alpha_{k-1}^2)L_{y^*}^2\mathbb{E}\left[\|\bar{r}_k\|^2\right] + C_{k,T}.$$

Taking summation on both sides, we have

$$\mu_g\sum_{k=1}^K \beta_k\mathbb{E}\left[\|\bar{y}_k^{(0)} - y_{k-1}^*\|^2\right] \leq \mathbb{E}\left[\|\bar{y}_1^{(0)} - y_0^*\|^2\right] + L_{y^*}^2\sum_{k=1}^K(\frac{\alpha_{k-1}^2}{\beta_k\mu_g} + \alpha_{k-1}^2)\mathbb{E}\left[\|\bar{r}_k\|^2\right] + \sum_{k=1}^K C_{k,T}.$$

□

Lemma A.9. *Suppose Assumptions 2.1, 2.2, 2.3, and 2.4 hold. In Algorithm 1 define*

$$\begin{aligned}
H^{(k)} &:= \frac{1}{n}\sum_{i=1}^n \nabla_y^2 g_i(\bar{x}_k, y_k^*),\ b^{(k)} := \frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\bar{x}_k, y_k^*), \\
z_*^{(k)} &:= \left(H^{(k)}\right)^{-1}\cdot b^{(k)} = \left(\sum_{i=1}^n \nabla_y^2 g_i(\bar{x}_k, y_k^*)\right)^{-1}\left(\sum_{i=1}^n \nabla_y f_i(\bar{x}_k, y_k^*)\right),
\end{aligned}$$

If γ satisfies (21), then we have

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbb{E}\left[\bar{z}_t^{(k)}|\mathcal{F}_k\right] - z_*^{(k)}\|^2\right] \\
\leq&(1-\gamma\mu_g)^N\cdot\frac{L_{f,0}^2}{\mu_g^2} + 5(\frac{1}{\mu_g^2} + \frac{\gamma}{\mu_g})(L_{g,2}^2\sigma_z^2 + L_{f,1}^2)\left(\mathbb{E}\left[\|\bar{y}_{k+1}^{(0)} - y_k^*\|^2\right] + \sigma_x^2\tilde{\alpha}_k^2 + T\tilde{\beta}_{k+1}^2\right) \\
&+90C_{\tilde{M}}L_{g,1}^2(\frac{1}{\mu_g^2} + \frac{\gamma}{\mu_g})(\frac{L_{f,0}^2}{\mu_g^2} + L_{f,0}^2)(1 - \frac{2\gamma\mu_g}{3})^{N-1}.
\end{aligned} \tag{37}$$

Proof of Lemma A.9. Define

$$\dot{z}_{t,k} := \mathbb{E}\left[\bar{z}_t^{(k)}|\mathcal{F}_k\right],\ \dot{s}_{t,k} := \mathbb{E}\left[\bar{s}_t^{(k)}|\mathcal{F}_k\right].$$

We know

$$\begin{aligned}
&\dot{z}_{t+1,k} - z_*^{(k)} = \dot{z}_{t+1,k} - z_*^{(k)} = \mathbb{E}\left[\bar{z}_t^{(k)}|\mathcal{F}_k\right] - \gamma\mathbb{E}\left[\bar{s}_t^{(k)}|\mathcal{F}_k\right] - z_*^{(k)}\\
=&\mathbb{E}\left[\bar{z}_t^{(k)}|\mathcal{F}_k\right] - \gamma\left(H^{(k)}\mathbb{E}\left[\bar{z}_t^{(k)}|\mathcal{F}_k\right] - b^{(k)}\right) - z_*^{(k)} - \gamma\left(\mathbb{E}\left[\bar{s}_t^{(k)}|\mathcal{F}_k\right] - \left(H^{(k)}\mathbb{E}\left[\bar{z}_t^{(k)}|\mathcal{F}_k\right] - b^{(k)}\right)\right)\\
=&\dot{z}_{t,k} - \gamma\left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right) - z_*^{(k)} - \gamma\left(\dot{s}_{t,k} - \left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right)\right).
\end{aligned}$$

Hence we know

$$\begin{aligned}
&\|\dot{z}_{t+1,k} - z_*^{(k)}\|^2\\
=&\|\dot{z}_{t,k} - \gamma\left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right) - z_*^{(k)} - \gamma\left(\dot{s}_{t,k} - \left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right)\right)\|^2\\
\le&(1+\gamma\mu_g)\|\dot{z}_{t,k} - \gamma\left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right) - z_*^{(k)}\|^2 + (1+\frac{1}{\gamma\mu_g})\gamma^2\|\dot{s}_{t,k} - \left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right)\|^2\\
\le&(1+\gamma\mu_g)(1-\gamma\mu_g)^2\|\dot{z}_{t,k} - z_*^{(k)}\|^2 + (\frac{\gamma}{\mu_g}+\gamma^2)\|\dot{s}_{t,k} - \left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right)\|^2\\
\le&(1-\gamma\mu_g)\|\dot{z}_{t,k} - z_*^{(k)}\|^2 + (\frac{\gamma}{\mu_g}+\gamma^2)\|\dot{s}_{t,k} - \left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right)\|^2,
\end{aligned} \tag{38}$$

where the second inequality uses Lemma A.3. For $\dot{s}_{t,k} - \left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right)$ we have

$$\begin{aligned}
&\|\dot{s}_{t,k} - \left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right)\|^2\\
=&\frac{1}{n^2}\|\sum_{i=1}^n\left(\nabla_y^2 g_i(x_{i,k}, y_{i,k}^{(T)})\mathbb{E}\left[z_{i,t}^{(k)}|\mathcal{F}_k\right] - \nabla_y^2 g_i(\bar{x}_k, y_k^*)\mathbb{E}\left[\bar{z}_t^{(k)}|\mathcal{F}_k\right] + \nabla_y f_i(\bar{x}_k, y_k^*) - \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)})\right)\|^2\\
=&\frac{1}{n^2}\|\sum_{i=1}^n\left[\nabla_y^2 g_i(x_{i,k}, y_{i,k}^{(T)})\mathbb{E}\left[z_{i,t}^{(k)} - \bar{z}_t^{(k)}|\mathcal{F}_k\right] - \left(\nabla_y^2 g_i(x_{i,k}, y_{i,k}^{(T)}) - \nabla_y^2 g_i(\bar{x}_k, \bar{y}_k^{(T)})\right)\dot{z}_{t,k}\right]\\
+&\sum_{i=1}^n\left[\left(\nabla_y^2 g_i(\bar{x}_k, \bar{y}_k^{(T)}) - \nabla_y^2 g_i(\bar{x}_k, y_k^*)\right)\dot{z}_{t,k} + \nabla_y f_i(\bar{x}_k, y_k^*) - \nabla_y f_i(\bar{x}_k, \bar{y}_k^{(T)}) + \nabla_y f_i(\bar{x}_k, \bar{y}_k^{(T)}) - \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)})\right]\|^2\\
\le&\frac{5}{n}\sum_{i=1}^n\left[L_{g,1}^2\|\mathbb{E}\left[z_{i,t}^{(k)} - \bar{z}_t^{(k)}|\mathcal{F}_k\right]\|^2 + (\|x_{i,k} - \bar{x}_k\|^2 + \|y_{i,k}^{(T)} - \bar{y}_k^{(T)}\|^2 + \|\bar{y}_k^{(T)} - y_k^*\|^2)(L_{g,2}^2\|\dot{z}_{t,k}\|^2 + L_{f,1}^2)\right]\\
=&\frac{5L_{g,1}^2}{n}\|\mathbb{E}\left[Z_t^{(k)} - \bar{z}_t^{(k)}\mathbf{1}^\top|\mathcal{F}_k\right]\|^2 + \frac{5(L_{g,2}^2\sigma_z^2 + L_{f,1}^2)}{n}\left(n\|\bar{y}_k^{(T)} - y_k^*\|^2 + \|X_k - \bar{x}_k\mathbf{1}^\top\|^2 + \|Y_k^{(T)} - \bar{y}_k^{(T)}\mathbf{1}^\top\|^2\right).
\end{aligned}$$

The above inequality and (38) imply

$$
\begin{aligned}
\mathbb{E}\left[\|\dot{z}_{N,k}-z_*^{(k)}\|^2\right] \leq & (1-\gamma\mu_g)\mathbb{E}\left[\|\dot{z}_{N-1,k}-z_*^{(k)}\|^2\right]+(\frac{\gamma}{\mu_g}+\gamma^2)\mathbb{E}\left[\|\dot{s}_{N-1,k}-\left(H^{(k)}\dot{z}_{N-1,k}-b^{(k)}\right)\|^2\right] \\
\leq & (1-\gamma\mu_g)^N\mathbb{E}\left[\|z_*^{(k)}\|^2\right]+\frac{5L_{g,1}^2(\frac{\gamma}{\mu_g}+\gamma^2)}{n}\sum_{t=0}^{N-1}(1-\gamma\mu_g)^{N-1-t}\mathbb{E}\left[\|\mathbb{E}\left[Z_t^{(k)}-\bar{z}_t^{(k)}\mathbf{1}^\top|\mathcal{F}_k\right]\|^2\right] \\
& +\frac{\frac{\gamma}{\mu_g}+\gamma^2}{1-(1-\gamma\mu_g)}\cdot\frac{5(L_{g,2}^2\sigma_z^2+L_{f,1}^2)}{n}\mathbb{E}\left[n\|\bar{y}_k^{(T)}-y_k^*\|^2+\|X_k-\bar{x}_k\mathbf{1}^\top\|^2+\|Y_k^{(T)}-\bar{y}_k^{(T)}\mathbf{1}^\top\|^2\right] \\
\leq & (1-\gamma\mu_g)^N\cdot\frac{L_{f,0}^2}{\mu_g^2}+5(\frac{1}{\mu_g^2}+\frac{\gamma}{\mu_g})(L_{g,2}^2\sigma_z^2+L_{f,1}^2)\left(\mathbb{E}\left[\|\bar{y}_k^{(T)}-y_k^*\|^2\right]+\sigma_x^2\tilde{\alpha}_k^2+T\tilde{\beta}_{k+1}^2\right) \\
& +5L_{g,1}^2(\frac{\gamma}{\mu_g}+\gamma^2)\sum_{t=0}^{N-1}(1-\frac{\gamma\mu_g}{2})^{N-1-t}(3C_{\tilde{M}}(1-\frac{2\gamma\mu_g}{3})^t(\frac{L_{f,0}^2}{\mu_g^2}+L_{f,0}^2)) \\
\leq & (1-\gamma\mu_g)^N\cdot\frac{L_{f,0}^2}{\mu_g^2}+5(\frac{1}{\mu_g^2}+\frac{\gamma}{\mu_g})(L_{g,2}^2\sigma_z^2+L_{f,1}^2)\left(\mathbb{E}\left[\|\bar{y}_k^{(T)}-y_k^*\|^2\right]+\sigma_x^2\tilde{\alpha}_k^2+T\tilde{\beta}_{k+1}^2\right) \\
& +90C_{\tilde{M}}L_{g,1}^2(\frac{1}{\mu_g^2}+\frac{\gamma}{\mu_g})(\frac{L_{f,0}^2}{\mu_g^2}+L_{f,0}^2)(1-\frac{2\gamma\mu_g}{3})^{N-1},
\end{aligned}
\tag{39}
$$

where the second inequality uses Lemma A.2, the third inequality uses (23), and the fourth inequality holds since

$$
\sum_{t=0}^{N-1-t}(1-\frac{\gamma\mu_g}{2})^{N-1-t}(1-\frac{2\gamma\mu_g}{3})^t=(1-\frac{2\gamma\mu_g}{3})^{N-1}\sum_{t=0}^{N-1}\frac{1-\frac{2\gamma\mu_g}{3}}{1-\frac{\gamma\mu_g}{2}}<(1-\frac{2\gamma\mu_g}{3})^{N-1}\cdot\frac{6}{\gamma\mu_g}.
$$

□

Lemma A.10. *If $0<\beta_k\leq 1$ and $\alpha_k>0$ for any $k\geq 0$, then the parameters $\tilde{\alpha}_k$, $\tilde{\beta}_k$, and $C_{k,T}$ defined in Lemmas A.7 and A.8 satisfy*

$$
\begin{aligned}
&\sum_{k=0}^{K}\tilde{\alpha}_k^2\leq\frac{2}{1-\rho^2}\sum_{i=0}^{K-1}\alpha_i^2=\mathcal{O}\left(\sum_{k=0}^{K}\alpha_k^2\right) \\
&\sum_{k=0}^{K}\tilde{\beta}_{k+1}^2\leq\frac{4(1+\rho^2)}{(1-\rho^2)^2}\left[(10\sigma_{g,1}^2+5\delta^2)\sum_{i=0}^{K}\beta_i^2+\frac{20L_{g,1}^2\sigma_x^2}{1-\rho^2}\sum_{i=0}^{K-1}\alpha_i^2\right]=\mathcal{O}\left(\sum_{k=0}^{K}(\alpha_k^2+\beta_k^2)\right) \\
&\sum_{k=1}^{K}C_{k,T}\leq(\frac{1}{\mu_g}+1)L_{g,1}^2\left[T\sigma_x^2\sum_{k=1}^{K}\tilde{\alpha}_k^2+2T^2\sum_{k=0}^{K}\tilde{\beta}_{k+1}^2\right]+\frac{T\sigma_{g,1}^2}{n}\sum_{k=1}^{K}\beta_k^2=\mathcal{O}\left(\sum_{k=0}^{K}(\alpha_k^2+\beta_k^2)\right).
\end{aligned}
$$

Proof of Lemma A.10. The first inequality holds due to $\tilde{\alpha}_0=0$ and

$$
\sum_{k=0}^{K-1}\tilde{\alpha}_{k+1}^2=\sum_{k=0}^{K-1}\sum_{i=0}^{k}\alpha_i^2\left(\frac{1+\rho^2}{2}\right)^{k-i}=\sum_{i=0}^{K-1}\sum_{k=i}^{K-1}\alpha_i^2\left(\frac{1+\rho^2}{2}\right)^{k-i}\leq\frac{2}{1-\rho^2}\sum_{i=0}^{K-1}\alpha_i^2.
$$

Similarly, we have

$$
\sum_{k=0}^{K}\tilde{\beta}_{k+1}^2=\frac{1+\rho^2}{1-\rho^2}\sum_{k=0}^{K}\sum_{i=0}^{k}\beta_i^2(10\sigma_{g,1}^2+10L_{g,1}^2\sigma_x^2\tilde{\alpha}_i^2+5\delta^2)\left(\frac{3+\rho^2}{4}\right)^{k-i}
$$

$$\leq \frac{4(1+\rho^2)}{(1-\rho^2)^2}\sum_{i=0}^{K}\beta_i^2(10\sigma_{g,1}^2+10L_{g,1}^2\sigma_x^2\tilde{\alpha}_i^2+5\delta^2) \leq \frac{4(1+\rho^2)}{(1-\rho^2)^2}\left[(10\sigma_{g,1}^2+5\delta^2)\sum_{i=0}^{K}\beta_i^2+\frac{20L_{g,1}^2\sigma_x^2}{1-\rho^2}\sum_{i=0}^{K-1}\alpha_i^2\right].$$

Lastly, we know

$$\begin{aligned}\sum_{k=1}^{K}C_{k,T} =& \sum_{k=1}^{K}\sum_{l=0}^{T-1}\left[(\frac{\beta_k}{\mu_g}+\beta_k^2)L_{g,1}^2\left(\sigma_x^2\tilde{\alpha}_k^2+\left[\left(\frac{3+\rho^2}{4}\right)^l T-l\left(\frac{3+\rho^2}{4}\right)\right]\tilde{\beta}_k^2+l\tilde{\beta}_{k+1}^2\right)+\frac{\beta_k^2\sigma_{g,1}^2}{n}\right]\\ \leq& \sum_{k=1}^{K}(\frac{\beta_k}{\mu_g}+\beta_k^2)L_{g,1}^2\left(T\sigma_x^2\tilde{\alpha}_k^2+T^2\tilde{\beta}_k^2+T^2\tilde{\beta}_{k+1}^2\right)+\sum_{k=1}^{K}T\frac{\beta_k^2\sigma_{g,1}^2}{n}\\ \leq& (\frac{1}{\mu_g}+1)L_{g,1}^2\left[T\sigma_x^2\sum_{k=1}^{K}\tilde{\alpha}_k^2+2T^2\sum_{k=0}^{K}\tilde{\beta}_{k+1}^2\right]+\frac{T\sigma_{g,1}^2}{n}\sum_{k=1}^{K}\beta_k^2,\end{aligned}$$

where the last inequality uses $0<\beta_k\leq 1$. □

Now we are ready to give the proof of Theorem 3.3.

Lemma A.11. *Suppose Assumptions 2.1, 2.2, 2.3, and 2.4 hold. For Algorithm 3 we have*

$$\sum_{k=0}^{K}(\frac{\alpha_k}{2}-\frac{L_\Phi\alpha_k^2}{2})\mathbb{E}\left[\|\bar{r}_k\|^2\right] \leq \frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2\right]+2\sigma_u^2\sum_{k=0}^{K}\alpha_k^2+\Phi(0)-\inf_x\Phi(x)+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_0\|^2\right]. \tag{40}$$

Proof of Lemma A.11. The L_Φ-smoothness of Φ indicates that

$$\Phi(\bar{x}_{k+1})-\Phi(\bar{x}_k) \leq \nabla\Phi(\bar{x}_k)^\top(-\alpha_k\bar{r}_k)+\frac{L_\Phi\alpha_k^2}{2}\|\bar{r}_k\|^2. \tag{41}$$

Notice that we also have

$$\frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}\|^2|\mathcal{F}_k\right]-\frac{1}{2}\|\bar{r}_k\|^2 = -\alpha_k\|\bar{r}_k\|^2+\alpha_k\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]^\top\bar{r}_k+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}-\bar{r}_k\|^2|\mathcal{F}_k\right]. \tag{42}$$

Hence we know

$$\begin{aligned}&\Phi(\bar{x}_{k+1})-\Phi(\bar{x}_k)+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}\|^2|\mathcal{F}_k\right]-\frac{1}{2}\|\bar{r}_k\|^2\\ \leq&\alpha_k(\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k))^\top\bar{r}_k+(\frac{L_\Phi\alpha_k^2}{2}-\alpha_k)\|\bar{r}_k\|^2+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}-\bar{r}_k\|^2|\mathcal{F}_k\right]\\ \leq&\frac{\alpha_k}{2}\left(\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2+\|\bar{r}_k\|^2\right)+(\frac{L_\Phi\alpha_k^2}{2}-\alpha_k)\|\bar{r}_k\|^2+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}-\bar{r}_k\|^2|\mathcal{F}_k\right],\end{aligned}$$

which implies

$$\begin{aligned}&(\frac{\alpha_k}{2}-\frac{L_\Phi\alpha_k^2}{2})\mathbb{E}\left[\|\bar{r}_k\|^2\right]\\ \leq&\frac{\alpha_k}{2}\mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2\right]+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}-\bar{r}_k\|^2\right]+\mathbb{E}\left[\Phi(\bar{x}_k)-\Phi(\bar{x}_{k+1})\right]+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_k\|^2\right]-\frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}\|^2\right]\\ \leq&\frac{\alpha_k}{2}\mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2\right]+2\alpha_k^2\sigma_u^2+\mathbb{E}\left[\Phi(\bar{x}_k)-\Phi(\bar{x}_{k+1})\right]+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_k\|^2\right]-\frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}\|^2\right],\end{aligned} \tag{43}$$

where the second inequality holds since we know

$$\mathbb{E}\left[\|\bar{r}_k\|^2\right] \leq \max\left(\mathbb{E}\left[\|\bar{r}_{k-1}\|^2\right], \mathbb{E}\left[\|\bar{u}_k\|^2\right]\right) \leq \max_{0\leq i\leq k} \mathbb{E}\left[\|\bar{u}_i\|^2\right] \leq \sigma_u^2,$$
$$\mathbb{E}\left[\|\bar{r}_{k+1} - \bar{r}_k\|^2\right] = \alpha_k^2 \mathbb{E}\left[\|\bar{r}_k - \bar{u}_k\|^2\right] \leq 2\alpha_k^2 \mathbb{E}\left[\|\bar{r}_k\|^2 + \|\bar{u}_k\|^2\right] \leq 4\sigma_u^2.$$

In these two conclusions $\mathbb{E}\left[\|\bar{u}_i\|^2\right] \leq \sigma_u^2$ is due to the first inequality in (29). Taking summation on both sides of (43), we have (40). □

Lemma A.12. *For Algorithm 3 we have*

$$\sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\|\bar{r}_k - \nabla\Phi(\bar{x}_k)\|^2\right] \leq \mathbb{E}\left[\|\bar{r}_0 - \nabla\Phi(0)\|^2\right] + 2\sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2\right] + 2\sum_{k=0}^{K} \alpha_k^2 \mathbb{E}\left[\|\bar{r}_k\|^2\right] + \sigma_u^2 \sum_{k=0}^{K} \alpha_k^2. \tag{44}$$

Proof of Lemma A.12. Recall that in Algorithm 3 we know

$$\bar{r}_{k+1} = (1-\alpha_k)\bar{r}_k + \alpha_k \bar{u}_k,$$

which implies

$$\begin{aligned}
&\|\bar{r}_{k+1} - \nabla\Phi(\bar{x}_{k+1})\| \\
=&\|(1-\alpha_k)(\bar{r}_k - \nabla\Phi(\bar{x}_k)) + \alpha_k(\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)) + \nabla\Phi(\bar{x}_k) - \nabla\Phi(\bar{x}_{k+1}) + \alpha_k(\bar{u}_k - \mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right])\|.
\end{aligned}$$

Hence we know

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{r}_{k+1} - \nabla\Phi(\bar{x}_{k+1})\|^2\right] \\
=&\mathbb{E}\left[\|(1-\alpha_k)(\bar{r}_k - \nabla\Phi(\bar{x}_k)) + \alpha_k(\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)) + \nabla\Phi(\bar{x}_k) - \nabla\Phi(\bar{x}_{k+1})\|^2\right] + \alpha_k^2 \mathbb{E}\left[\|\bar{u}_k - \mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]\|^2\right] \\
\leq&(1-\alpha_k)\mathbb{E}\left[\|\bar{r}_k - \nabla\Phi(\bar{x}_k)\|^2\right] + \alpha_k \mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k) + \frac{1}{\alpha_k}(\nabla\Phi(\bar{x}_k) - \nabla\Phi(\bar{x}_{k+1}))\|^2\right] + \alpha_k^2\sigma_u^2 \\
\leq&(1-\alpha_k)\mathbb{E}\left[\|\bar{r}_k - \nabla\Phi(\bar{x}_k)\|^2\right] + 2\alpha_k \mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2 + \alpha_k\|\bar{r}_k\|^2\right] + \alpha_k^2\sigma_u^2.
\end{aligned}$$

Taking summation on both sides, we obtain (44). □

The next lemma characterizes $\|\nabla\Phi(\bar{x}_k) - \mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]\|^2$, which together with previous lemmas prove Theorem 3.3.

Lemma A.13. *In Algorithm 3 if we define*

$$\alpha_k = \frac{\mu_g^4}{3L_{g,1}^2 C_y} \cdot \beta_k \equiv \frac{1}{\sqrt{K}},\ \gamma \text{ such that (21) holds},\ N = \Theta(\log K),\ T \geq 2,$$
$$C_y = 5(L_{f,1}^2 + \frac{L_{g,2}^2 L_{f,0}^2}{\mu_g^2}) + 50L_{g,1}^2(\frac{1}{\mu_g^2} + \frac{\gamma}{\mu_g})(L_{g,2}^2\sigma_z^2 + L_{f,1}^2).$$

we have

$$\sum_{k=0}^{K} \alpha_k \|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2 = C_y \sum_{k=0}^{K} \alpha_k \|\bar{y}_k^{(T)} - y_k^*\|^2 + \mathcal{O}\left(1 + (1-\frac{\gamma\mu_g}{2})^N \sum_{k=0}^{K} \alpha_k\right),$$
$$\frac{1}{K}\sum_{k=0}^{K} \mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

Proof of Lemma A.13. Notice that we have

$$\begin{aligned}
\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] &= \frac{1}{n}\sum_{i=1}^n \nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}) - \frac{1}{n}\sum_{i=1}^n \nabla_{xy}^2 g_i(x_{i,k}, y_{i,k}^{(T)})\mathbb{E}\left[z_{i,N}^{(k)}|\mathcal{F}_k\right], \\
\nabla\Phi(\bar{x}_k) &= \frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\bar{x}_k, y_k^*) - \left(\frac{1}{n}\sum_{i=1}^n \nabla_{xy}^2 g_i(\bar{x}_k, y_k^*)\right)\left(\frac{1}{n}\sum_{i=1}^n \nabla_y^2 g_i(\bar{x}_k, y_k^*)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \nabla_y f_i(\bar{x}_k, y_k^*)\right), \\
&= \frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\bar{x}_k, y_k^*) - \frac{1}{n}\left(\sum_{i=1}^n \nabla_{xy}^2 g_i(\bar{x}_k, y_k^*)\right)\left(\sum_{i=1}^n \nabla_y^2 g_i(\bar{x}_k, y_k^*)\right)^{-1}\left(\sum_{i=1}^n \nabla_y f_i(\bar{x}_k, y_k^*)\right) \\
&= \frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\bar{x}_k, y_k^*) - \frac{1}{n}\left(\sum_{i=1}^n \nabla_{xy}^2 g_i(\bar{x}_k, y_k^*)\right) z_*^{(k)}.
\end{aligned}$$

Hence we know

$$\begin{aligned}
&\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\| \\
=&\frac{1}{n}\sum_{i=1}^n \left(\|\nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}) - \nabla_x f_i(\bar{x}_k, \bar{y}_k^{(T)})\| + \|\nabla_x f_i(\bar{x}_k, \bar{y}_k^{(T)}) - \nabla_x f_i(\bar{x}_k, y_k^*)\|\right) \\
+&\frac{1}{n}\sum_{i=1}^n \left(\|\nabla_{xy}^2 g_i(x_{i,k}, y_{i,k}^{(T)})(\mathbb{E}\left[z_{i,N}^{(k)}|\mathcal{F}_k\right] - z_*^{(k)})\| + \|(\nabla_{xy}^2 g_i(x_{i,k}, y_{i,k}^{(T)}) - \nabla_{xy}^2 g_i(\bar{x}_k, \bar{y}_k^{(T)}))z_*^{(k)}\|\right) \\
+&\frac{1}{n}\sum_{i=1}^n \|(\nabla_{xy}^2 g_i(\bar{x}_k, \bar{y}_k^{(T)}) - \nabla_{xy}^2 g_i(\bar{x}_k, y_k^*))z_*^{(k)}\|,
\end{aligned}$$

which implies

$$\begin{aligned}
&\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2 \\
\leq&\frac{5}{n}\sum_{i=1}^n \left[L_{f,1}^2(\|x_{i,k} - \bar{x}_k\|^2 + \|y_{i,k}^{(T)} - \bar{y}_k^{(T)}\|^2 + \|\bar{y}_k^{(T)} - y_k^*\|^2) + L_{g,1}^2\|\mathbb{E}\left[z_{i,N}^{(k)}|\mathcal{F}_k\right] - z_*^{(k)}\|^2\right] \\
+&\frac{5}{n}\sum_{i=1}^n \left[\frac{L_{g,2}^2 L_{f,0}^2}{\mu_g^2}(\|x_{i,k} - \bar{x}_k\|^2 + \|y_{i,k}^{(T)} - \bar{y}_k^{(T)}\|^2 + \|\bar{y}_k^{(T)} - y_k^*\|^2)\right] \\
\leq&5(L_{f,1}^2 + \frac{L_{g,2}^2 L_{f,0}^2}{\mu_g^2})\cdot\frac{1}{n}\left(\|X_k - \bar{x}_k\mathbf{1}^\top\|^2 + \|Y_k^{(T)} - \bar{y}_k^{(T)}\mathbf{1}^\top\|^2 + n\|\bar{y}_k^{(T)} - y_k^*\|^2\right) \\
+&10L_{g,1}^2\cdot\frac{1}{n}\left(\|\mathbb{E}\left[Z_N^{(k)} - \bar{z}_N^{(k)}\mathbf{1}^\top|\mathcal{F}_k\right]\|^2\right) + 10L_{g,1}^2\|\mathbb{E}\left[\bar{z}_N^{(k)}|\mathcal{F}_k\right] - z_*^{(k)}\|^2 \\
\leq&\left[5(L_{f,1}^2 + \frac{L_{g,2}^2 L_{f,0}^2}{\mu_g^2}) + 50L_{g,1}^2(\frac{1}{\mu_g^2} + \frac{\gamma}{\mu_g})(L_{g,2}^2\sigma_z^2 + L_{f,1}^2)\right]\cdot\left(\|\bar{y}_k^{(T)} - y_k^*\|^2 + \sigma_x^2\tilde{\alpha}_k^2 + T\tilde{\beta}_{k+1}^2\right) \\
+&30L_{g,1}^2(1 - \frac{\gamma\mu_g}{2})^N(\frac{L_{f,0}^2}{\mu_g^2} + L_{f,0}^2) \\
+&10L_{g,1}^2\left[(1-\gamma\mu_g)^N\cdot\frac{L_{f,0}^2}{\mu_g^2} + 90C_{\tilde{M}}L_{g,1}^2(\frac{1}{\mu_g^2} + \frac{\gamma}{\mu_g})(\frac{L_{f,0}^2}{\mu_g^2} + L_{f,0}^2)(1 - \frac{2\gamma\mu_g}{3})^{N-1}\right],
\end{aligned}$$

where the third inequality uses (29), (23) and (37). Taking summation on both sides, we have

$$\sum_{k=0}^{K}\alpha_k\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2=C_y\sum_{k=0}^{K}\alpha_k\|\bar{y}_k^{(T)}-y_k^*\|^2+\mathcal{O}\left(\sum_{k=0}^{K}\alpha_k(\tilde{\alpha}_k^2+\tilde{\beta}_k^2)+(1-\frac{\gamma\mu_g}{2})^{N-1}\sum_{k=0}^{K}\alpha_k\right). \tag{45}$$

Setting

$$\alpha_k=C_{\alpha,\beta}\cdot\beta_k\equiv\frac{1}{\sqrt{K}},\ C_{\alpha,\beta}=\frac{\mu_g}{\sqrt{3C_yL_{y^*}}},$$

and using (34) and Lemma A.10, we know

$$\begin{aligned}
&\frac{1}{\sqrt{K}}\sum_{k=0}^{K}\mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2\right]=C_yC_{\alpha,\beta}\sum_{k=0}^{K}\beta_k\|\bar{y}_k^{(T)}-y_k^*\|^2+\mathcal{O}\left(\frac{1}{\sqrt{K}}+\sqrt{K}(1-\frac{\gamma\mu_g}{2})^{N-1}\right)\\
=&C_yC_{\alpha,\beta}L_{y^*}^2\sum_{k=1}^{K}(\frac{C_{\alpha,\beta}}{\sqrt{K}\mu_g^2}+\frac{1}{K\mu_g})\mathbb{E}\left[\|\bar{r}_k\|^2\right]+\mathcal{O}\left(1+\sqrt{K}(1-\frac{\gamma\mu_g}{2})^{N-1}\right)\\
=&\sum_{k=1}^{K}(\frac{1}{3\sqrt{K}}+\frac{C_yC_{\alpha,\beta}L_{y^*}^2}{K\mu_g})\mathbb{E}\left[\|\bar{r}_k\|^2\right]+\mathcal{O}\left(1+\sqrt{K}(1-\frac{\gamma\mu_g}{2})^{N-1}\right),
\end{aligned} \tag{46}$$

which together with (40) and (10) imply

$$\begin{aligned}
&(\frac{1}{2\sqrt{K}}-\frac{L_\Phi}{2K})\sum_{k=0}^{K}\mathbb{E}\left[\|\bar{r}_k\|^2\right]\\
\leq&\frac{1}{2\sqrt{K}}\sum_{k=0}^{K}\mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2\right]+2\sigma_u^2\sum_{k=0}^{K}\frac{1}{K}+\Phi(0)-\inf_x\Phi(x)+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_0\|^2\right]\\
\leq&\frac{1}{2\sqrt{K}}\sum_{k=1}^{K}(\frac{1}{3}+\frac{C_yC_{\alpha,\beta}L_{y^*}^2}{\sqrt{K}\mu_g})\mathbb{E}\left[\|\bar{r}_k\|^2\right]+\mathcal{O}\left(1+\sqrt{K}(1-\frac{\gamma\mu_g}{2})^{N-1}\right).
\end{aligned}$$

Hence we know

$$(\frac{1}{3\sqrt{K}}-\frac{L_\Phi}{2K}-\frac{C_yC_{\alpha,\beta}L_{y^*}^2}{2K\mu_g})\sum_{k=0}^{K}\mathbb{E}\left[\|\bar{r}_k\|^2\right]=\mathcal{O}\left(1+\sqrt{K}(1-\frac{\gamma\mu_g}{2})^{N-1}\right).$$

Using the above expression, (46) and Lemma A.12, we know

$$\frac{1}{\sqrt{K}}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right]\leq\frac{2}{\sqrt{K}}\sum_{k=0}^{K}\mathbb{E}\left[\|\bar{r}_k\|^2+\|\bar{r}_k-\nabla\Phi(\bar{x}_k)\|^2\right]=\mathcal{O}\left(1+\sqrt{K}(1-\frac{\gamma\mu_g}{2})^{N-1}\right).$$

Note that γ is in a constant interval by (21), hence $(1-\frac{\gamma\mu_g}{2})$ is a constant that is independent of K. Picking $N=\Theta(\log K)$ such that $(1-\frac{\gamma\mu_g}{2})^{N-1}=\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$, we know

$$\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right]=\mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

Moreover, from (29) we know:

$$\frac{1}{K}\sum_{k=0}^{K}\frac{\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right]}{n}=\mathcal{O}\left(\frac{1}{K}\sum_{k=0}^{K}\tilde{\alpha}_k^2\right)=\mathcal{O}\left(\frac{1}{K}\right),$$

where the second equality holds due to Lemma A.10 The above two equalities prove Theorem 3.3. To find an ϵ-stationary point, we may set $K=\Theta(\epsilon^{-2})$ and we know from $T\geq 2,\ N=\log K$ that the sample complexity will be $\tilde{\mathcal{O}}(\epsilon^{-2})$. □

B Discussion

We briefly discuss Assumption 3.4 (iv) and (v) in Yang et al. (2022) in this section.

- Assumption 3.4 (iv) assumes bounded second moment of $\nabla_y g_i(x,y;\xi)$. It is stronger than our Assumption 2.3 as discussed right after Assumption 2.3.
- Assumption 3.4 (v) assumes each $I-\frac{1}{L_g}\nabla_y^2 g_i(x,y;\xi)$ has bounded second moment such that

$$\mathbb{E}\left[\|I-\frac{1}{L_g}\nabla_y^2 g_i(x,y;\xi)\|_2^2\right]\leq(1-\kappa_g)^2,$$

  for some constant $\kappa_g\in(0,\frac{\mu_g}{L_g})$, where $L_g=\sqrt{L_{g,2}^2+\sigma_{g,2}^2}$. It serves as a key role in proving the linear convergence of the Hessian matrix inverse estimator (see Lemma A.2, A.3 and the definition of b right under section B of the Supplementary Material). However, it is restrictive under certain cases. For any given $0<\mu_g<L_g$, consider $X\in\mathbb{R}^{2\times 2}$ to be a random matrix and

$$X=\begin{pmatrix}2L_g & 0\\ 0 & 0\end{pmatrix}\text{ or }\begin{pmatrix}0 & 0\\ 0 & 2\mu_g\end{pmatrix}\text{ with equal probability,}$$

  then it is easy to verify that X has bounded variance and in expectation equals $\mathrm{diag}(L,\mu)$, but

$$\mathbb{E}\left[\|I-\frac{1}{L_g}X\|_2^2\right]=1,$$

  and thus their Assumption 3.4 (v) does not hold in this example.