Stochastic Optimization for Machine Learning: Investigations on Bilevel Optimization and Large Learning Rates

By

XUXING CHEN

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Applied Mathematics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Krishnakumar Balasubramanian, Chair

Jesús De Loera

Mina Karzand

Shiqian Ma

Committee in Charge

2024

To my family

Contents

Abstract v

Acknowledgments vi

Chapter 1. Introduction 1

1.1. Preliminaries 1

1.2. Outline of the Dissertation 7

Chapter 2. Stochastic Bilevel Optimization 9

2.1. Introduction 9

2.2. Proposed Framework: the MA-SOBA Algorithm 14

2.3. Theoretical Analysis 16

2.4. Min-Max Bilevel Optimization 19

2.5. Experiments 22

2.6. Conclusion 32

Chapter 3. Decentralized Stochastic Bilevel Optimization 33

3.1. Introduction 33

3.2. Preliminaries 37

3.3. DSBO Algorithm with Improved Per-Iteration Complexity 39

3.4. Numerical experiments 46

3.5. Conclusion 48

Chapter 4. Training Dynamics of Gradient Descent for Quadratic Regression 49

4.1. Introduction 49

4.2. Analyzing a discrete dynamical system with cubic map 54

4.3. Applications to quadratic regression models 60

4.4. Experimental investigations 64

4.5. Conclusion 66

Appendix A. Additional Experiments, Proofs, and Discussions 69

A.1. Proofs of Theorems in Chapter 2 69

A.2. Discussions on the Prior Works Related to Chapter 2 97

A.3. Additional Experiments on Heterogeneous Data of Chapter 3 99

A.4. Proofs of Theorems in Chapter 3 100

A.5. Discussions on the Prior Works Related to Chapter 3 127

A.6. Experimental Investigations of Chapter 4 129

A.7. Proofs of Theorems in Chapter 4 135

A.8. Auxiliary Results in Chapter 4 152

Bibliography 154

Abstract

Stochastic optimization is fundamental to modern machine learning and deep learning problems. It provides various algorithmic frameworks, such as stochastic gradient descent (SGD), adaptive gradient algorithm (ADAGRAD) and adaptive moment estimation (ADAM), to efficiently minimize loss functions constructed from large-scale datasets. In this dissertation, we explore the theoretical properties and empirical performance of bilevel optimization algorithms and the phenomenon of large learning rates in machine learning. First, we introduce a novel algorithm, the Moving-Average Stochastic Bilevel Algorithm (MA-SOBA), designed for solving stochastic bilevel optimization under standard smoothness assumptions. Next, we extend the scope of bilevel optimization algorithms from single-agent training to a multi-agent context, i.e., distributed training, by proposing the Moving-Average Decentralized Stochastic Bilevel Optimization (MA-DSBO) algorithm. This approach improves the per-iteration complexity of previous methods, reducing the quadratic dependency on dimensionality to linear dependency. Lastly, inspired by the Edge of Stability (EoS) phenomenon observed in modern deep learning, we examine the training dynamics of gradient descent in a class of quadratic regression models with large learning rates —- a scenario that classical optimization theory struggles to explain.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisors, Prof. Krishnakumar Balasubramanian and Prof. Shiqian Ma. Their patience and professionalism provided me with invaluable support and guidance, allowing me to freely pursue my interests during my Ph.D. studies. Through their mentorship, I have gained significant insights into optimization and developed essential soft skills for conducting scientific research. I am especially thankful that they recognized the value of my mathematical background and accepted me as their student, even when I had no prior publications.

I am also deeply grateful to Prof. Naoki Saito, Prof. Lifeng Lai, Prof. Xiaodong Li, Prof. Mina Karzand, and Prof. Jesús De Loera for serving as the committee members of my Qualifying Exam and Dissertation. Their professional suggestions and comments have greatly improved my work.

My life at Davis would have been incomplete without my amazing friends. I enjoy every chat with Jiaxiang Li, whose knowledge extends far beyond mathematics into many other areas like history and music. I am thankful to Minhui Huang for his encouragement and assistance while I was writing my first manuscript. Working with Tesi Xiao is a great pleasure, and our fruitful discussions always lead to high-quality work. I also cherish the time spent with my roommates, Qi Yu and Cheng Li. I am extremely fortunate to have all of them at Davis, and I want to thank them for enriching my Ph.D. life with their companionship.

I am also grateful to other collaborators, including Saeed Ghadimi, Kaiyi Ji, Promit Ghosal, Bhavya Agrawalla, Yifan Hu, Abhishek Roy, Xiaoyu Wang, and many others. I have learned so much from their expertise in various domains.

In addition to my academic experiences, I was also fortunate to have the opportunity to conduct research in the industry. I am particularly thankful to my mentor, Yun He, and my manager, Xiaoyi (Leo) Liu, during my internship at Meta. They provided me with great support to freely explore how to bridge the gap between theory and practice. I also thank Prof. Rong Jin for our insightful discussions on the theory and experiments of optimization algorithms in both academia and industry. I am grateful to my Metamates, including Jiayi Xu, Fei Tian, Xiaohan Wei, Xue Feng, Boyang Liu, Yang Yang, Chiyu Zhang, Tan Wang, and many others.

Special thanks go to my old friends, Jiwei Li, Zihan Chen, Panke Jing, and many others. Although we were in different time zones during my Ph.D. studies, they were always there willing to offer emotional support whenever I was going through difficult times.

I also want to extend a heartfelt thank you to my favorite singer-songwriter Aimyon, whose beautiful songs guided me through the darkest moments of my Ph.D. journey.

I would also like to acknowledge my high school Mathematical Olympiad coaches, Jia Zhang and Rui Zhang. They gave me the confidence and courage to enjoy the art of problem solving. Thanks to their rigorous training, I never doubted my ability to overcome the technical challenges of math problems encountered in my Ph.D. life, not even for a second.

Finally, I want to thank my parents, Mr. Chen and Mrs. Xu, for their unconditional love and support. Without them, I could never have achieved so much in my life.

CHAPTER 1

Introduction

In this Section, we first briefly introduce the basic setup of stochastic optimization theory as well as the background of bilevel optimization, decentralized optimization, and large learning rates phenomenon in deep learning. In Section 1.1, we present the preliminaries of the main topics and include commonly-used definitions and assumptions in optimization literature. In Section 1.2, we provide an overview of the main contents in this dissertation.

Notation. We use $\|\cdot\|$ for ℓ^2 norm of a vector and Frobenius norm of a matrix. $\mathbf{1}_n$ denotes the all-one vector in $\mathbb{R}^n$. $\Delta_n = \{\lambda \mid \lambda_i \geq 0, \sum_{i=1}^n \lambda_i = 1\}$ denotes the probability simplex. For a convex compact set $\mathcal{X}$, $\Pi_{\mathcal{X}}(\cdot)$ denotes the Euclidean projection onto it. $\langle \cdot, \cdot \rangle$ denotes the inner product.

1.1. Preliminaries

1.1.1. Preliminaries of Stochastic Optimization. Many machine learning problems can be thought of as stochastic optimization problems as

$$\min_{\theta \in \mathcal{X}} \quad f(\theta) := \mathbb{E}_{\xi} \left[F(\theta; \xi) \right], \tag{1.1}$$

where $\mathcal{X}$ is a convex compact set, and ξ is a random variable that denotes the sampling process to generate the objective function f. For example, ξ can refer to a mini-batch of data points, and $F(\theta; \xi)$ represents the loss function evaluated based on ξ.

Remark. Suppose we are given a dataset $\mathcal{D}$ with N datapoints, i.e., $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^N$. A typical choice of training objective is

$$f(\theta) = \frac{1}{N} \sum_{i=1}^{N} F(\theta; X_i, y_i) \tag{1.2}$$

where $F(\theta; X_i, y_i)$ represents the loss function evaluated on a single data point (X_i, y_i). We note that (1.2) can fit into the expectation formulation in (1.1). To see this, we denote by ξ the random variable to generate $\mathcal{B}_\xi$, a batch of data with a fixed size B uniformly sampled from the dataset. We

further define

$$F(\theta;\xi)=\frac{1}{B}\sum_{(X,y)\in\mathcal{B}_\xi}F(\theta;X,y). \tag{1.3}$$

We notice that

$$\begin{aligned}\mathbb{E}_\xi[F(\theta;\xi)]&=\sum_{\xi:\mathcal{B}_\xi\subseteq\mathcal{D}}\frac{1}{\binom{N}{B}}F(\theta;\xi)\\&=\sum_{\xi:\mathcal{B}_\xi\subseteq\mathcal{D}}\left(\sum_{(X,y)\in\mathcal{B}_\xi}\frac{1}{\binom{N}{B}}\cdot\frac{1}{B}F(\theta;X,y)\right)\\&=\sum_{(X,y)\in\mathcal{D}}\left(\sum_{\xi:(X,y)\in\mathcal{B}_\xi}\frac{1}{\binom{N}{B}}\cdot\frac{1}{B}F(\theta;X,y)\right)\\&=\sum_{(X,y)\in\mathcal{D}}\frac{\binom{N-1}{B-1}}{\binom{N}{B}}\cdot\frac{1}{B}F(\theta;X,y)\\&=\frac{1}{N}\sum_{(X,y)\in\mathcal{D}}F(\theta;X,y)=f(\theta).\end{aligned} \tag{1.4}$$

Here the third equality holds since we can exchange the summation order of a mini-batch and a particular data point, and the fourth equality holds since for a fixed datapoint $(X,y)\in\mathcal{D}$, we have in total $\binom{N-1}{B-1}$ batches with size B containing (X,y). This indicates that the average of the loss functions evaluated on the whole dataset can be seen as the expectation of the loss functions evaluated on a mini-batch of data points uniformly sampled from the whole dataset.

Unless specified, we will assume the functions of interest are continuously differentiable. We define of the notion of Lipschitz continuity, smoothness, and convexity assumptions as follows.

Definition 1.1.1. *Suppose we are given a differentiable function $f:\mathbb{R}^d\to\mathbb{R}$. For any constants $\ell_0,\ell_1>0$, we say f is ℓ_0-Lipschitz continuous, when $|f(\theta_1)-f(\theta_2)|\le\ell_0\|\theta_1-\theta_2\|$ for any θ_1,θ_2. f is ℓ_1-smooth, when $\|\nabla f(\theta_1)-\nabla f(\theta_2)\|\le\ell_1\|\theta_1-\theta_2\|$ for any θ_1,θ_2. We say f is convex when there exists a constant $\mu\ge0$ such that $f(\theta_2)\ge f(\theta_1)+\langle\nabla f(\theta_1),\theta_2-\theta_1\rangle+\frac{\mu}{2}\|\theta_1-\theta_2\|^2$ for any θ_1,θ_2. In particular, we say f is μ-strongly convex when $\mu>0$.*

REMARK. For any symmetric positive semi-definite matrix A, we can verify that $f(\theta) = \frac{1}{2}\langle A\theta, \theta\rangle$ is smooth and convex, since we have

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| = \|A\theta_1 - A\theta_2\| \leq \|A\|_2 \|\theta_1 - \theta_2\|,$$

and

$$\begin{aligned}
& f(\theta_2) - f(\theta_1) - \langle \nabla f(\theta_1), \theta_2 - \theta_1\rangle \\
=& \frac{1}{2}(\langle A\theta_2, \theta_2\rangle - \langle A\theta_1, \theta_1\rangle - 2\langle A\theta_1, \theta_2 - \theta_1\rangle) \\
=& \frac{1}{2}\langle A(\theta_2 - \theta_1), (\theta_2 - \theta_1)\rangle \\
\geq& \frac{\lambda_{\min}(A)}{2}\|\theta_2 - \theta_1\|^2,
\end{aligned}$$

where $\lambda_{\min}(A)$ denotes the smallest eigenvalue of A. Hence we know $f(\theta)$ is $\|A\|_2$-smooth and convex.

To develop stochastic optimization algorithms, we assume that we access to unbiased estimates of derivatives of different orders with bounded variance.

ASSUMPTION 1. *Suppose we are given a function $f(\theta) = \mathbb{E}_\xi[F(\theta;\xi)]$ defined on a convex compact set $\mathcal{X}$. We assume access to unbiased stochastic function values, first-order derivatives and second-order derivatives with bounded variance, i.e., there exist constants $\sigma_1, \sigma_2 > 0$ such that for any $\theta \in \mathcal{X}$,*

$$\begin{aligned}
&\mathbb{E}_\xi[\nabla F(\theta;\xi)] = \nabla f(\theta),\ \mathbb{E}_\xi\left[\|\nabla F(\theta;\xi) - \nabla f(\theta)\|^2\right] \leq \sigma_1^2, \\
&\mathbb{E}_\xi\left[\nabla^2 F(\theta;\xi)\right] = \nabla^2 f(\theta),\ \mathbb{E}_\xi\left[\left\|\nabla^2 F(\theta;\xi) - \nabla^2 f(\theta)\right\|^2\right] \leq \sigma_2^2.
\end{aligned}$$

REMARK. To see the Assumptions hold in practice, we take a linear regression problem as an example. Consider training a linear model under the setup in that of (1.2) and (1.3), the loss functions are defined as

$$F(\theta; X, y) = \frac{1}{2}(\langle X, \theta\rangle - y)^2, \quad F(\theta;\xi) = \frac{1}{B}\sum_{(X,y)\in\mathcal{B}_\xi} F(\theta; X, y).$$

We can follow (1.4) to similarly verify that the unbiasedness of $\nabla F(\theta;\xi)$ and $\nabla^2 F(\theta;\xi)$. To see the variances are bounded, we notice that $\nabla F(\theta;X,y) = XX^\top\theta - yX, \nabla^2 F(\theta;X,y) = XX^\top$ and thus

(1.5)

$$
\begin{aligned}
&\mathbb{E}_\xi\left[\|\nabla F(\theta;\xi) - \nabla f(\theta)\|^2\right] \\
=&\mathbb{E}_\xi\left[\left\|\left(\frac{1}{B}\sum_{(X_i,y_i)\in\mathcal{B}_\xi} X_iX_i^\top - \frac{1}{N}\sum_{j=1}^N X_jX_j^\top\right)\theta - \left(\frac{1}{B}\sum_{(X_i,y_i)\in\mathcal{B}_\xi} y_iX_i - \frac{1}{N}\sum_{j=1}^N y_jX_j\right)\right\|^2\right] \\
\le& 2\mathbb{E}_\xi\left[\left\|\frac{1}{B}\sum_{(X_i,y_i)\in\mathcal{B}_\xi} X_iX_i^\top - \frac{1}{N}\sum_{j=1}^N X_jX_j^\top\right\|^2\right]\|\theta\|^2 + 2\mathbb{E}_\xi\left[\left\|\frac{1}{B}\sum_{(X_i,y_i)\in\mathcal{B}_\xi} y_iX_i - \frac{1}{N}\sum_{j=1}^N y_jX_j\right\|^2\right] \\
<&+\infty
\end{aligned}
$$

where the first inequality holds by Cauchy-Schwarz inequality, and the second inequality holds since θ belongs to a compact set. Also we have

$$
\mathbb{E}_\xi\left[\left\|\nabla^2 F(\theta;\xi) - \nabla^2 f(\theta)\right\|^2\right] = \mathbb{E}_\xi\left[\left\|\frac{1}{B}\sum_{(X_i,y_i)\in\mathcal{B}_\xi} X_iX_i^\top - \frac{1}{N}\sum_{j=1}^N X_jX_j^\top\right\|^2\right] < +\infty.
$$

It is worth noting that the boundedness of $\mathcal{X}$ plays a crucial role in the upper bound of the variance of $\nabla F(\theta;\xi)$. When $\mathcal{X}$ is unbounded and (1.5) does not hold, then we may need to analyze the variance of the gradients evaluted at the iterates of the algorithms. See Theorem 1 in Chen et al. [2024] and its proof techniques for an example.

We call $\nabla F(\theta;\xi), \nabla^2 F(\theta;\xi)$ first-order, and second-order stochastic oracles respectively. In practice, for example, we may sample a mini-batch of data points (represented by ξ), and then evaluate oracles of different orders according to the algorithmic design.

We next discuss convergence criteria of optimization algorithms. For bounded continuously differentiable convex functions, it can be shown that there exists θ^* such that $f(\theta^*)$ equals to the minimum value of f. In deep learning, however, most loss functions constructed from deep neural networks are highly non-convex, and finding a global minimum point is usually infeasible. We thus consider finding first-order stationary points defined as follows.

DEFINITION 1.1.2. *Suppose we are given a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. For a positive constant $\epsilon > 0$, we say $\theta \in \mathbb{R}^d$ is an ϵ-stationary point of f, when $\|\nabla f(\theta)\|^2 \leq \epsilon$. Moreover, we say a stochastic optimization algorithm is able to find an ϵ-stationary point, when there exists a positive integer K such that the iterates $\{\theta_k\}_{k=0}^K$ generated by the algorithm satisfy* $\min_{0\leq k\leq K} \mathbb{E}\left[\|\nabla f(\theta_k)\|^2\right] \leq \epsilon$.

Note that for non-convex functions, a 0-stationary point can be a local minima, a local maxima, or a saddle point. In non-convex optimization, they represent critical points of interest because they can potentially offer good enough solutions in the absence of global guarantees. Additionally, many algorithms designed for non-convex optimization are built around the goal of efficiently converging to these stationary points, making them a central concept in the analysis and application of non-convex optimization algorithms like SGD [Robbins and Monro, 1951] and Adam [Kingma, 2014]. For a particular stochastic optimization algorithm, we are interested in analyzing the relationship between K and ϵ in Definition 1.1.2, which reveals the rate of convergence or the number of samples needed to find such a solution.

1.1.2. Preliminaries of (Decentralized) Bilevel Optimization. Despite that most machine learning problems can be written as the optimization formulation in (1.1), which can then be solved effectively by algorithms like SGD, there is a broad class of problems that cannot be formulated as (1.1). We will take hyperparameter optimization as an example. In classical wisdom of machine learning, the hyperparameter tuning process aims at finding the best hyperparameters based on the validation dataset after model training. Note that the training and tuning are based on different datasets and thus the optimization problems involved are different. We can write the hperparameter optimization problem as follows.

$$\begin{aligned} \min_{\lambda} \quad & \mathcal{L}_{\text{val}}(\lambda, \theta^*(\lambda)), \\ \text{s.t.} \quad & \theta^*(\lambda) = \arg\min_{\theta} \mathcal{L}_{\text{train}}(\lambda, \theta). \end{aligned}$$

Here, we denote by λ the hyperparameters and θ the trainable model parameters. We use $\mathcal{L}_{\text{val}}$ and $\mathcal{L}_{\text{train}}$ to represent validation loss and training loss respectively. Note that the problem considered here has a bilevel structure, in which we can evaluate the gradients in both levels to solve this problem through some gradient-based algorithms. The study of gradient-based hypergradient optimization can be dated back to Bengio [2000]. In general, bilevel optimization aims at solving problems with a

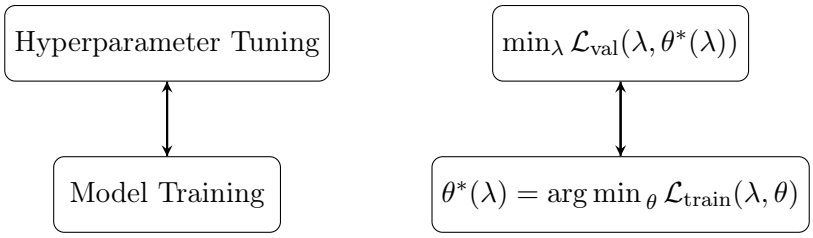


FIGURE 1.1. Hyperparameter optimization as a bilevel optimization problem.

bilevel structure, and can be written as follows.

$$\begin{aligned} \min_{\lambda} \quad & \Phi(\lambda) := f(\lambda, \theta^*(\lambda)), \\ \text{s.t.} \quad & \theta^*(\lambda) = \arg\min_{\theta} g(\lambda, \theta). \end{aligned}$$

It has attracted a lot of attention in recent years due to its capabilities of handling machine learning problems in different domains like hyperparameter optimization [Bengio, 2000, Franceschi et al., 2018, Bertrand et al., 2020], reinforcement learning [Hong et al., 2020, Chakraborty et al., 2024], meta-learning [Bertinetto et al., 2019, Franceschi et al., 2018, Rajeswaran et al., 2019, Ji et al., 2020], etc. We study the convergence rates and sample complexity of BO algorithms in Chapter 2.

Another important line of research is extending the single-agent training to decentralized setting [Lian et al., 2017], in which multiple agents, such as different devices distributed in heterogeneous environments [Yuan et al., 2022], work collaboratively to solve the problem. The decentralized version of Problem 1.1 can be formulated as

$$\min_{\theta} \quad f(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta).$$

Similarly, decentralized bilevel optimization can be written as

$$\begin{aligned} \min_{\lambda} \quad & \Phi(\lambda) := \frac{1}{n} \sum_{i=1}^{n} f_i(\lambda, \theta^*(\lambda)), \\ \text{s.t.} \quad & \theta^*(\lambda) = \arg\min_{\theta} g(\lambda, \theta) := \frac{1}{n} \sum_{i=1}^{n} g_i(\lambda, \theta), \end{aligned}$$

where n represents the number of agents for solving the problem. The information related to functions f_i and g_i (e.g., the data points used to generate them), is termed as local information only available to agent i. Typically, each agent in the network sends (receives) certain iterates to (from)

their neighbors, through a given communication network, to collect the global information. The network can be represented as a graph with vertices representing agents and edges representing the neighboring relations between pairs of agents. See, for example, Figure 1.2 for a visualization of different communication networks. The number of communication rounds between different agents is denoted as the communication complexity. We will elaborate how to design the communication process to achieve convergence for decentralized bilevel optimization in Chapter 3.

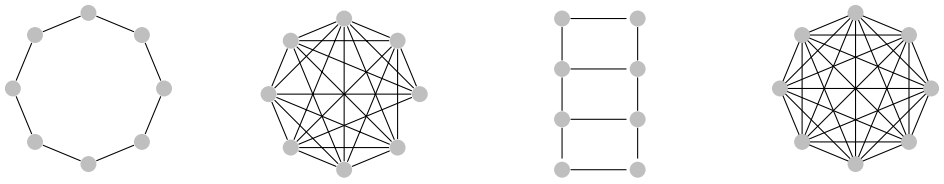


FIGURE 1.2. Different communication networks for $n = 8$. The four graphs represent the ring, (an instance of) the random, the ladder and the complete graph. The figure is adopted from Li et al. [2024].

1.1.3. Beyond Small Learning Rates in Optimization. Finally, we note that, from a theoretical perspective, classical optimization theory may require the learning rates in algorithms to be sufficiently small to achieve convergence guarantee for finding the stationary points. In practice, however, most empirical selection of the learning rates are larger than what the theory predicts, and usually the training process can be greatly accelerated as compared to the settings with conservative learning rates. See, for example, the Edge of Stability (EoS) [Cohen et al., 2021] phenomenon. Hence there is a clear gap between theory and practice. We aim at providing a partial answer to this open question in Chapter 4.

1.2. Outline of the Dissertation

Now that we have introduced the preliminaries of stochastic optimization with its applications in machine learning, we present the outline of the dissertation in this section.

In Chapter 2, we propose a fully single-loop algorithm for solving stochastic bilevel optimization problems. Our theoretical analysis reveals that the sample complexity of our algorithm for finding an ϵ-stationary point matches the optimal lower bound under standard assumptions, and thus closes the gap between the lower bound and upper bound. Furthermore, we show that by a slight modification of our approach, our algorithm can handle a more general multi-objective robust bilevel optimization

problem. For this case, we obtain the state-of-the-art oracle complexity results demonstrating the generality of both the proposed algorithmic and analytic frameworks. Numerical experiments demonstrate the performance gain of the proposed algorithms over existing ones.

In Chapter 3, we extend the single-agent training of bilevel problems to the decentralized setting. We provide a novel algorithm that successfully improves the per-iteration computational and communication complexity from quadratic dependence on dimension parameter to linear dependence, as compared to prior works on this topic. Numerical experiments showcase the efficiency of our algorithms on both synthetic datasets and real-world datasets.

In Chapter 4, we study the training dynamics of gradient descent with large learning rates for a class of quadratic regression problems. Through the lens of discrete dynamical systems, we show that the dynamics of gradient descent can exhibit five distinct phases based on different choices of learning rates.

CHAPTER 2

Stochastic Bilevel Optimization

2.1. Introduction

Bilevel optimization is gaining increasing popularity within the machine learning community due to its extensive range of applications, including meta-learning [Bertinetto et al., 2019, Franceschi et al., 2018, Rajeswaran et al., 2019, Ji et al., 2020], hyperparameter optimization [Bengio, 2000, Franceschi et al., 2018, Bertrand et al., 2020], data augmentation [Cubuk et al., 2019, Rommel et al., 2022], and neural architecture search [Liu et al., 2019, Zhang et al., 2022b]. The objective of bilevel optimization is to minimize a function that is dependent on the solution of another optimization problem. Formally, we have

$$\min_{x\in\mathcal{X}\subseteq\mathbb{R}^{d_x}} \Phi(x) := f(x, y^*(x)) \qquad \text{s.t.} \quad y^*(x) = \arg\min_{y\in\mathbb{R}^{d_y}} g(x, y) \tag{2.1}$$

where the upper-level (UL) function f (a.k.a. *outer function*) and the lower-level (LL) function g (a.k.a. *inner function*) are two real-valued functions defined on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. The set $\mathcal{X}$ is either $\mathbb{R}^{d_x}$ or a closed convex set in $\mathbb{R}^{d_x}$, and the LL function g is strongly convex. We call x the *outer variable* and y the *inner variable*. The objective function $\Phi(x)$ is called the *value function*. In this paper, we consider the stochastic setting in which the f and g are expressed in the form of expectations, i.e., $f(x, y) = \mathbb{E}_{\xi\sim\mathcal{D}_f}\left[F(x, y; \xi)\right]$, $g(x, y) = \mathbb{E}_{\phi\sim\mathcal{D}_g}\left[G(x, y; \phi)\right]$. Stochastic bilevel optimization can be considered as an extension of bilevel empirical risk minimization [Dagréou et al., 2023], allowing to effectively handle streaming data (ξ, ϕ).

In many instances, the analytical expression of $y^*(x)$ is unknown and can only be approximated using an optimization algorithm. This adds to the complexity of problem (2.1) compared to its single-level counterpart. Under regular conditions such that Φ is differentiable, the *hypergradient* $\nabla\Phi(x)$ derived by chain rule and implicit function theorem is given by

$$\nabla\Phi(x) = \nabla_1 f(x, y^*(x)) - \nabla^2_{12} g(x, y^*(x)) z^*(x), \tag{2.2}$$

where $z^*(x) \in \mathbb{R}^{d_y}$ is the solution of a linear system:

$$z^*(x) = \left[\nabla^2_{22} g(x, y^*(x))\right]^{-1} \nabla_2 f(x, y^*(x)). \tag{2.3}$$

Solving (2.1) using only stochastic oracles poses significant challenges since there is no direct unbiased estimator available for $\left[\nabla^2_{22} g(x, y^*(x))\right]^{-1}$ and also for $\nabla\Phi(x)$ as a consequence.

To mitigate the estimation bias, many existing methods [Ghadimi and Wang, 2018, Ji et al., 2021, Yang et al., 2021, Hong et al., 2023, Guo et al., 2021b, Khanduri et al., 2021, Chen et al., 2021a, Akhtar et al., 2022] employ a Hessian Inverse Approximation (`HIA`) subroutine, which involves drawing a mini-batch of stochastic Hessian matrices and computing a truncated Neumann series [Stewart, 1998]. However, this subroutine comes with an increased computational burden and introduces an additional factor of $\log(\epsilon^{-1})$ in the sample complexity. Alternative methods proposed by Chen et al. [2022a] and Guo et al. [2021a] calculate the explicit inverse of the stochastic Hessian matrix with momentum updates. To circumvent the need for explicit Hessian inversion and the `HIA` subroutine, Arbel and Mairal [2022] and Dagréou et al. [2022] propose running Stochastic Gradient Descent (`SGD`) steps to approximate the solution $z^*(x)$ of the linear system (2.3). In particular, the state-of-the-art Stochastic Bilevel Algorithm (`SOBA`) *only* utilizes SGD steps to simultaneously update three variables: the inner variable y, the outer variable x, and the auxiliary variable z. It was claimed that `SOBA` achieves the same complexity lower bound of its single-level counterpart ($\Phi \in \mathcal{C}^{1,1}_L$ [‡‡]) in the non-convex setting [Arjevani et al., 2023].

Despite the superior computational and sample efficiency of `SOBA`, there is crucial shortcoming as the current theoretical framework assumes high-order smoothness for the `UL` function f and the `LL` function g such that $z^*(x)$ has Lipschitz gradient. Specifically, unlike the typical assumptions in stochastic bilevel optimization that state $f \in \mathcal{C}^{1,1}_L$ and $g \in \mathcal{C}^{2,2}_L$ (`A1`), the current theory of `SOBA` requires $f \in \mathcal{C}^{2,2}_L$ and $g \in \mathcal{C}^{3,3}_L$ (`A2`). The necessity of (`A2`) is counter-intuitive as the partial gradients of x, y, z utilized in constructing SGD steps are already Lipschitz continuous under (`A1`). Furthermore, assuming g is strongly convex and the partial gradient of the `UL` function with respect to the inner variable y is bounded for all pairs of $(x, y^*(x))$, (i.e., $\|\nabla_2 f(x, y^*(x))\| \leq L_f$ for all $x \in \mathcal{X}$), there exists a subset relation among three function classes as indicated by Ghadimi and

[‡‡] $\mathcal{C}^{p,p}_L$ denotes p-times differentiability with Lipschitz k-th order derivatives for $0 < k \leq p$.

Wang [2018, Lemma 2.2] that

$$(\texttt{A2})\ \{f \in \mathcal{C}_L^{2,2}, g \in \mathcal{C}_L^{3,3}\} \subset\ (\texttt{A1})\ \{f \in \mathcal{C}_L^{1,1}, g \in \mathcal{C}_L^{2,2}\} \subset \{\Phi \in \mathcal{C}_L^{1,1}\}.$$

In light of this, it can be concluded that (A1) is sufficient to ensure the first-order Lipschitzness of Φ, which is the standard assumption in the single-level setting. On the other hand, it is worth noting that under (A2) it can be shown that $\Phi \in \mathcal{C}_L^{2,2}$, i.e., $\nabla\Phi(x)$ and $\nabla^2\Phi(x)$ are both Lipschitz continuous. It is known that higher order smoothness (e.g., Lipschitz continuity of $\nabla^2\Phi(x)$) will lead to better sample complexity [Carmon et al., 2017, Arjevani et al., 2020]. This indicates that the sample complexity $\mathcal{O}(\epsilon^{-2})$ obtained in Dagréou et al. [2022] may not be optimal under the set of assumptions made in their work.

Therefore, a natural question follows: *Is it possible to develop a* fully single-loop *and* Hessian-inversion-free *algorithm for solving stochastic bilevel optimization problems that achieves an* optimal *sample complexity of* $\mathcal{O}(\epsilon^{-2})$ *under* standard smoothness assumptions $\{f \in \mathcal{C}_L^{1,1}, g \in \mathcal{C}_L^{2,2}\}$[‡‡]*?* In this paper, we provide an affirmative answer to the aforementioned question. Our **contributions** can be summarized as follows.

- We propose a class of fully single-loop and Hession-inversion-free algorithm, named Moving-Average SOBA (MA-SOBA), which builds upon the SOBA algorithm by incorporating an additional sequence of average hypergradients. Unlike SOBA, MA-SOBA achieves an optimal sample complexity of $\mathcal{O}(\epsilon^{-2})$ under standard smoothness assumptions, without relying on high-order smoothness. In particular, in Section A.1.1.1 we explain how the introduced MA updates help reduce the order of bias in hypergradient estimation, and avoid higher order Taylor expansion (which requires higher-order smoothness of f and g) used in Dagréou et al. [2022]. Moreover, the introduced sequence of average hypergradients converges to $\nabla\Phi(x)$, thus offering a reliable termination criterion in the stochastic setting.
- We expand the scope of MA-SOBA to tackle a broader class of problems, specifically the min-max multi-objective bilevel optimization problem with significant applications in robust machine learning. We introduce MORMA-SOBA, an algorithm that can find an ϵ-first-order stationary point of the μ_λ-strongly-concave regularized formulation while achieving a sample complexity of $\mathcal{O}(n^5\mu_\lambda^{-4}\epsilon^{-2})$, which fills a gap (in terms of the order of ϵ-dependency) in the existing literature.

Method (double-loop)	Sample Complexity	(UL) f [‡]	(LL) g	Hession Inversion	Inner Loop	Batch Size
BSA [Ghadimi and Wang, 2018]	$\tilde{\mathcal{O}}(\epsilon^{-3})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	Neumann approx.	SGD on inner	$\tilde{\mathcal{O}}(1)$
stocBiO [Ji et al., 2021]	$\tilde{\mathcal{O}}(\epsilon^{-2})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	Neumann approx.	SGD on inner	$\tilde{\mathcal{O}}(\epsilon^{-1})$
[¶]ALSET [Chen et al., 2021a]	$\tilde{\mathcal{O}}(\epsilon^{-2})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	Neumann approx.	SGD on inner	$\tilde{\mathcal{O}}(1)$
AmIGO [Arbel and Mairal, 2022]	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	SGD	SGD on inner	$\mathcal{O}(\epsilon^{-1})$
Method (single-loop)	Sample Complexity	(UL) f [‡]	(LL) g	Hession Inversion	Inner Step	Batch Size
[¶]TTSA [Hong et al., 2023]	$\tilde{\mathcal{O}}(\epsilon^{-2.5})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	Neumann approx.	SGD	$\tilde{\mathcal{O}}(1)$
STABLE [Chen et al., 2022a]	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	Direct	SGD	$\mathcal{O}(1)$
SOBA [Dagréou et al., 2022]	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{3,3}$	SGD	SGD	$\mathcal{O}(1)$
MA-SOBA (Alg. 1)	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	SGD	SGD	$\mathcal{O}(1)$

TABLE 2.1. Comparison of the stochastic bilevel optimization solvers in the nonconvex-strongly-convex setting under smoothness assumptions [‡‡] on f and g. We omit the comparison with variance reduction-based methods (VRBO, MRBO [Yang et al., 2021]; SUSTAIN [Khanduri et al., 2021]; SABA [Yang et al., 2021]; SRBA [Dagréou et al., 2023]; SVRB [Guo et al., 2021a]; FLSA [Li et al., 2022a]; SBFW [Akhtar et al., 2022]) that may achieve $\mathcal{O}(\epsilon^{-1.5})$ sample complexity and under mean-squared smoothness assumptions on stochastic functions F_ξ and G_ϕ, and SBMA [Guo et al., 2021b] that achieves $\mathcal{O}(\epsilon^{-4})$ sample complexity. The sample complexity corresponds to the number of calls to stochastic gradients and Hessian(Jacobian)-vector products to get an ϵ-stationary point. The $\tilde{\mathcal{O}}$ notation hides a factor of $\log(\epsilon^{-1})$. "SC" means "strongly-convex".

- We conduct experiments on several machine learning problems. Our numerical results show the efficiency and superiority of our algorithms.

Related Work. The concept of bilevel optimization was initially introduced in the work of Bracken and McGill [1973]. Since then, numerous gradient-based bilevel optimization algorithms have been proposed, broadly categorized into two groups: ITerative Differentiation (ITD) based methods [Domke, 2012, Maclaurin et al., 2015, Franceschi et al., 2018, Grazzi et al., 2020, Ji et al., 2021] and Approximate Implicit Differentiation (AID) based methods [Domke, 2012, Pedregosa, 2016, Gould et al., 2016, Ghadimi and Wang, 2018, Grazzi et al., 2020, Ji et al., 2021, Arbel and Mairal, 2022, Grazzi et al., 2023]. The ITD-based algorithms typically involve approximating the solution of the inner problem using an iterative algorithm and then computing an approximate hypergradient through automatic differentiation. However, a major drawback of this approach is

[‡] All methods also assume $\|\nabla_2 f(x, y^*(x))\| \leq L_f < \infty$ for all $x \in \mathcal{X}$.

[¶] ALSET can achieve convergence without the need for double loops, but it comes at the cost of a worse dependence on κ in sample complexity. The mechanisms of single-loop ALSET and TTSA are essentially the same, except that ALSET employs single time-scale stepsizes while TTSA employs two time-scales.

the necessity of storing each iterate of the inner optimization algorithm in memory. The AID-based algorithms leverage the implicit gradient given by (2.2), which requires the solution of a linear system characterized by (2.3). Extensive research has been conducted on designing and analyzing deterministic bilevel optimization algorithms with strongly-convex LL functions; see Ji et al. [2021] and references therein.

In recent years, there has been a growing interest in stochastic bilevel optimization, especially in the setting of a non-convex UL function and a strongly-convex LL function. To address estimation bias, one set of methods uses SGD iterations for the inner problem and employs truncated stochastic Neumann series to approximate the inverse of the Hessian matrix in $z^*(x)$ [Ghadimi and Wang, 2018, Ji et al., 2021, Yang et al., 2021, Hong et al., 2023, Guo et al., 2021b, Khanduri et al., 2021, Chen et al., 2021a, Akhtar et al., 2022]. The analysis of such methods was refined by [Chen et al., 2021a] to achieve convergence rates similar to those of SGD. However, Neumann approximation subroutine introduces an additional factor of $\log(\epsilon^{-1})$ in the sample complexity. Some alternative approaches [Arbel and Mairal, 2022, Chen et al., 2022a, Guo et al., 2021a] calculate the explicit inverse of the stochastic Hessian matrix with momentum updates. Nevertheless, these methods encounter challenges related to computational complexity in matrix inversion and numerical stability.

To avoid the need for explicit Hessian inversion and Neumann approximation, recent algorithms [Arbel and Mairal, 2022, Dagréou et al., 2022] propose running SGD steps to approximate the solution $z^*(x)$ of the linear system (2.3). One such algorithm called AmIGO [Arbel and Mairal, 2022] employs a double-loop approach with warm-start strategy and achieves an optimal sample complexity of $\mathcal{O}(\epsilon^{-2})$ under regular assumptions. However, AmIGO requires a growing batch size inversely proportional to ϵ. Following AmIGO, Grazzi et al. [2023] proposes BSGM, which avoids using large batch size in the LL problem and warm-start strategy, but still requires double-loop framework and large batch sizes in the UL problem. On the other hand, the single-loop algorithm SOBA [Dagréou et al., 2022] achieves the same complexity lower bound but with constant batch size. Unfortunately, the current analysis of SOBA relies on the assumption of higher-order smoothness for the UL and LL functions. In this work, we introduce a novel algorithm framework that differs slightly from SOBA but can achieve optimal sample complexity in theory without higher-order smoothness assumptions. A summary of our results and comparison to prior work is provided in Table 2.1.

In addition, there exist several variance reduction-based methods following the line of research by Yang et al. [2021], Khanduri et al. [2021], Yang et al. [2021], Dagréou et al. [2023], Guo et al. [2021a], Li et al. [2022a]. Some of these methods achieve an improved sample complexity of $\mathcal{O}(\epsilon^{-1.5})$ and match the lower bounds of their single-level counterparts when stochastic functions F_ξ and G_ϕ satisfy mean-squared smoothness assumptions and the algorithm is allowed simultaneous queries at the same random seed [Arjevani et al., 2023]. However, since we are specifically considering smoothness assumptions on f and g, we will not delve into the comparison with these methods.

The most recent advancements in (stochastic) bilevel optimization focus on several new ideas: (i) addressing constrained lower-level problems [Shen and Chen, 2023, Xiao et al., 2023, Tsaknakis et al., 2022, Giovannelli et al., 2021], (ii) handling lower-level problems that lack strong convexity [Chen et al., 2023a, Huang, 2023, Liu et al., 2023, 2021, Sow et al., 2022a, Jiang et al., 2023], (iii) developing fully first-order (Hessian-free) algorithms [Liu et al., 2022, Kwon et al., 2023, Sow et al., 2022b], (iv) establishing convergence to the second-order stationary point [Huang et al., 2023], and (v) expanding the framework to encompass multi-objective optimization problems [Giovannelli et al., 2023, Gu et al., 2023, Hu et al., 2022]. It is promising to apply some of these advancements to our specific framework. Moreover, in this work, we also contribute to multi-objective bilevel problems with a slight modification of our approach. Other directions are left as future work.

2.2. Proposed Framework: the MA-SOBA Algorithm

Similar to Dagréou et al. [2022], Arbel and Mairal [2022], our algorithm initiates with inexact hypergradient descent techniques and seeks to offer an alternative in the stochastic setting. To provide a clear illustration, let us initially consider the deterministic setting. The `SOBA` framework keeps track of three sequences, denoted as $\{x^k, y^k, z^k\}$, and updates them using D_x, D_y, D_z as follows:

$$\text{(2.4)}\quad \textbf{(inner)}\quad y^{k+1} = y^k - \beta_k \boxed{\nabla_2 g(x^k, y^k)} = y^k - \beta_k D_y(x^k, y^k, z^k)$$

$$\textbf{(aux)}\quad z^{k+1} = z^k - \gamma_k \left\{\nabla^2_{22} g(x^k, y^*(x^k))z^k - \nabla_2 f(x^k, y^*(x^k))\right\}$$

$$\text{(2.5)}\quad \textbf{bias} \to \approx z^k - \gamma_k \boxed{\left\{\nabla^2_{22} g(x^k, y^k)z^k - \nabla_2 f(x^k, y^k)\right\}} = z^k - \gamma_k D_z(x^k, y^k, z^k)$$

$$\textbf{(outer)}\quad x^{k+1} = x^k - \alpha_k \left\{\nabla_1 f(x^k, y^*(x^k)) - \nabla^2_{12} g(x^k, y^*(x^k))z^*(x^k)\right\} = x^k - \alpha_k \nabla\Phi(x^k)$$

$$\text{(2.6)}\quad \textbf{bias} \to \approx x^k - \alpha_k \boxed{\left\{\nabla_1 f(x^k, y^k) - \nabla^2_{12} g(x^k, y^k)z^k\right\}} = x^k - \alpha_k D_x(x^k, y^k, z^k)$$

where (2.4) is the GD step to minimize $g(x^k, \cdot)$, (2.6) is the inexact hyper gradient descent step, and (2.5) is the GD step to minimize a quadratic function with $z^*(x^k)$ being the solution, i.e.,

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

Given that the above update rule, highlighted in blue, does not involve the Hessian matrix inversion, `SOBA` can directly utilize the stochastic oracles of $\nabla_1 f, \nabla_2 f, \nabla_2 g, \nabla_{22}^2 g, \nabla_{12}^2 g$ to obtain unbiased estimators of D_x, D_y, D_z in Eq.(2.4), (2.5), (2.6). This approach circumvents the requirement for a Neumann approximation subroutine or a direct matrix inversion. However, due to the update rule for y, which only utilizes one-step SGD at each iteration k, the value of y^k does not coincide with $y^*(x^k)$. As a result, a certain bias is introduced in the partial gradient of z in Eq.(2.5). Similarly, when estimating the hypergradient $\nabla\Phi(x)$, another bias term arises in Eq.(2.6). Although the bias decreases to zero as $y^k \to y^*(x^k)$ and $z^k \to z^*(x^k)$ under standard smoothness assumptions as indicated by Lemma 3.4 in Dagréou et al. [2022], the current analysis of `SOBA` requires more regularity on f and g to carefully handle the bias; it assume that f has Lipschitz Hessian and g has Lipschitz third-order derivative.

The inability to obtain an unbiased gradient estimator is a common characteristic in stochastic optimization involving nested structures; see, for example, stochastic compositional optimization [Wang et al., 2017, Yang et al., 2019, Ghadimi et al., 2020, Balasubramanian et al., 2022, Chen et al., 2021b] as a specific case of (2.1). One popular approach is to introduce a sequence of dual variables that approximates the true gradient by aggregating all past biased stochastic gradients using a moving averaging technique [Ghadimi et al., 2020, Balasubramanian et al., 2022, Xiao et al., 2022]. Motivated by this approach, we introduce another sequence of variables, denoted as $\{h^k\}$, and update it at k-th iteration given the past iterates $\mathscr{F}_k$ as $h^{k+1} = (1-\theta_k)h^k + \theta_k w^{k+1}$, where $\mathbb{E}[w^{k+1}|\mathscr{F}_k] = D_x(x^k, y^k, z^k), \theta_k \in (0, 1]$. Following the update rule in the constrained setting ($\mathcal{X} \subset \mathbb{R}^{d_x}$) [Ghadimi et al., 2020], the outer variable is updated as $x^{k+1} = x^k + \alpha_k \left(\Pi_{\mathcal{X}}(x^k - \tau h^k) - x^k\right)$, which is reduced to the GD step when $\mathcal{X} \equiv \mathbb{R}^{d_x}$. Denote the stochastic oracles of $\nabla_1 f(x^k, y^k), \nabla_2 f(x^k, y^k), \nabla_2 g(x^k, y^k)$, $\nabla_{22}^2 g(x^k, y^k), \nabla_{12}^2 g(x^k, y^k)$ at k-th iteration as $u_x^{k+1}, u_y^{k+1}, v^{k+1}, H^{k+1}, J^{k+1}$ respectively. We present our method, referred to as **M**oving-**A**verage `SOBA` (`MA-SOBA`), in Algorithm 1.

Algorithm 1: `Moving-Average SOBA`

Input: $x^0, y^0, z^0, h^0 = 0, \{\alpha_k\}, \{\beta_k\}, \{\gamma_k\}, \{\theta_k\}$

for $k = 0, 1, \ldots, K-1$ **do**

$x^{k+1} = x^k + \alpha_k \left(\Pi_{\mathcal{X}}(x^k - \tau h^k) - x^k\right)$ # update x^k via average hypergradient h^k

$y^{k+1} = y^k - \beta_k v^{k+1}$ # update y^k by one-step SGD

$z^{k+1} = z^k - \gamma_k (H^{k+1} z^k - u_y^{k+1})$ # update z^k by one-step SGD

$h^{k+1} = (1-\theta_k) h^k + \theta_k \left(u_x^{k+1} - J^{k+1} z^k\right)$ # update average hypergradient h^k

end

2.3. Theoretical Analysis

In this section, we provide convergence rates of `MA-SOBA` under *standard* smoothness conditions on f, g and *regular* assumptions on stochastic oracles. We also present a proof sketch and have detailed discussions about assumptions made in the literature. The complete proofs are deferred in Section A.1.

2.3.1. Preliminaries and Assumptions. As we consider the general setting in which $\mathcal{X}$ can be either $\mathbb{R}^{d_x}$ or a closed convex set in $\mathbb{R}^{d_x}$, we use the notion of gradient mapping to characterize the first-order stationarity, which is a classical measure widely used in the literature as a convergence criterion when solving nonconvex constrained problems [Nesterov, 2018]. For $\tau > 0$, we define the gradient mapping of at point $\bar{x} \in \mathcal{X}$ as $\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla\Phi(\bar{x}), \tau) := \frac{1}{\tau}(\bar{x} - \Pi_{\mathcal{X}}(\bar{x} - \tau\nabla\Phi(\bar{x})))$. When $\mathcal{X} \equiv \mathbb{R}^d$, the gradient mapping simplifies to $\nabla\Phi(\bar{x})$. Our main goal in this work is to find an ϵ-stationary solution to (2.1), in the sense of $\mathbb{E}[\|\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla\Phi(\bar{x}), \tau)\|^2] \le \epsilon$.

We first state some regularity assumptions on the functions f and g.

Assumption 2. *The functions f and g satisfy:*

(a) **($f \in \mathcal{C}_L^{1,1}$ and $g \in \mathcal{C}_L^{2,2}$)**‡‡ *$\nabla f, \nabla g, \nabla^2 g$ are $L_{\nabla f}, L_{\nabla g}, L_{\nabla^2 g}$ Lipschitz continuous.*

(b) **(SC `LL`)** *g is μ_g-strongly convex.*

(c) *$\|\nabla_2 f(x, y^*(x))\| \le L_f < \infty$ for all $x \in \mathcal{X}$.*

Remark. The above assumption serves as a sufficient condition for the Lipschitz continuity of $\nabla\Phi$, $y^*(x)$, and $z^*(x)$, as well as D_x, D_y, and D_z in Eq. (2.4), (2.5), (2.6). The inclusion of high-order smoothness assumptions ($f \in \mathcal{C}_L^{2,2}$ and $g \in \mathcal{C}_L^{3,3}$) in the current analysis of `SOBA` [Dagréou et al., 2022] is primarily intended to ensure the Lipschitzness of $\nabla z^*(x)$. However, the necessity of such assumptions is subject to doubt, given that $\nabla z^*(x)$ is not involved in designing the algorithm.

Furthermore, the Lipschitzness of f or uniformly boundedness of $\nabla_2 f$ made in several previous works is unnecessary. Instead, the boundedness assumption on $\nabla_2 f$ is only required for all pairs of $(x, y^*(x))$ as demonstrated by Assumption 2(c).

Next, we discuss assumptions made on the stochastic oracles.

ASSUMPTION 3. *For any $k \geq 0$, denote by $\mathscr{F}_k$ the sigma algebra generated by all iterates with superscripts not greater than k : $\mathscr{F}_k = \sigma\left\{h^1, \ldots, h^k, x^1, \ldots, x^k, y^1, \ldots, y^k, z^1, \ldots, z^k\right\}$. The **stochastic oracles** of $\nabla_1 f(x^k, y^k), \nabla_2 f(x^k, y^k), \nabla_2 g(x^k, y^k), \nabla^2_{22} g(x^k, y^k), \nabla^2_{12} g(x^k, y^k)$, denoted as $u_x^{k+1}, u_y^{k+1}, v^{k+1}, H^{k+1}, J^{k+1}$ respectively, used in Algorithm 2 at k-th iteration are **unbiased** with **bounded variance** given $\mathscr{F}_k$, i.e., there exist positive constants $\sigma_{f,1}, \sigma_{f,2}, \sigma_{g,1}, \sigma_{g,2}$ such that*

$$
\begin{aligned}
&\mathbb{E}[u_x^{k+1}|\mathscr{F}_k] = \nabla_1 f(x^k, y^k),\ \mathbb{E}\left[\left\|u_x^{k+1} - \nabla_1 f(x^k, y^k)\right\|^2 |\mathscr{F}_k\right] \leq \sigma_{f,1}^2,\\
&\mathbb{E}\left[u_y^{k+1}|\mathscr{F}_k\right] = \nabla_2 f(x^k, y^k),\ \mathbb{E}\left[\left\|u_y^{k+1} - \nabla_2 f(x^k, y^k)\right\|^2 |\mathscr{F}_k\right] \leq \sigma_{f,2}^2,\\
&\mathbb{E}\left[v^{k+1}|\mathscr{F}_k\right] = \nabla_2 g(x^k, y^k),\ \mathbb{E}\left[\left\|v^{k+1} - \nabla_2 g(x^k, y^k)\right\|^2 |\mathscr{F}_k\right] \leq \sigma_{g,1}^2,\\
&\mathbb{E}\left[H^{k+1}|\mathscr{F}_k\right] = \nabla^2_{22} g(x^k, y^k),\ \mathbb{E}\left[\left\|H^{k+1} - \nabla^2_{22} g(x^k, y^k)\right\|^2 |\mathscr{F}_k\right] \leq \sigma_{g,2}^2,\\
&\mathbb{E}\left[J^{k+1}|\mathscr{F}_k\right] = \nabla^2_{12} g(x^k, y^k),\ \mathbb{E}\left[\left\|J^{k+1} - \nabla^2_{12} g(x^k, y^k)\right\|^2 |\mathscr{F}_k\right] \leq \sigma_{g,2}^2.
\end{aligned}
$$

*In addition, they are conditionally **independent** conditioned on $\mathscr{F}_k$.*

REMARK. The unbiasedness and bounded variance assumptions on stochastic oracles are standard and typically satisfied in several practical stochastic optimization problems [Lan, 2020]. It is important to highlight that we explicitly impose these assumptions on the stochastic oracles, unlike Assumption 3.6 in Dagréou et al. [2022], which assumes $\mathbb{E}[\|v^{k+1}\|^2|\mathscr{F}_k] \leq B_y^2(1 + \|D_y(x^k, y^k, z^k)\|^2)$ and $\mathbb{E}[\|H^{k+1}z^k - u_y^{k+1}\|^2|\mathscr{F}_k] \leq B_z^2(1+\|D_z(x^k, y^k, z^k)\|^2)$. In this case, B_y and B_z represent constants in terms of the Lipschitz constants (L) and variance bounds (σ^2). Moreover, Assumption 3.7 in Dagréou et al. [2022] assumes $\mathbb{E}[\|w^{k+1}\|^2|\mathscr{F}_k] \leq B_x^2$ holds for a constant B_x, which is considerably stronger than our assumptions and may not hold for a broad class of problems.

2.3.2. Convergence Results. We have the following theorem characterizing the convergence results of `MA-SOBA`.

THEOREM 2.3.1. *Define* $x_+^k = \Pi_{\mathcal{X}}(x^k - \tau h^k)$. *Suppose Assumptions 2 and 3 hold. Then there exist positive constants* $c_1, c_2, c_3, \tau > 0$ *such that if* $\alpha_k \equiv \Theta(1/\sqrt{K})$, $\beta_k = c_1\alpha_k, \gamma_k = c_2\alpha_k, \theta_k = c_3\alpha_k$, *in Algorithm 1, then the iterates in Algorithm 1 satisfy*

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau^2}\mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \quad \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|h^k - \nabla\Phi(x^k)\right\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \tag{2.7}$$

which imply

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

That is to say, when uniformly randomly selecting a solution x^R *from* $\{x^1, \ldots, x^K\}$, *the sample complexity of Algorithm 1 for finding an* ϵ*-stationary point is* $\mathcal{O}(\epsilon^{-2})$.

REMARK. In contrast to most existing methods, in `MA-SOBA`, the introduced sequence of dual variables $\{h^k\}$ converges to the exact hypergradient $\nabla\Phi(x)$, even in the presence of estimation bias. This attribute provides reliable terminating criteria in practice. In addition, similar results with an extra factor of $\log(K)$ in the convergence rate can be established under decreasing α_k [Dagréou et al., 2022]. We also note that Algorithm 1 only requires stochastic gradient and Hessian(Jacobian)-vector product oracles, whose computational complexity are typically $\mathcal{O}(\max(d_x, d_y))$ with the help of automatic differentiation techniques [Pearlmutter, 1994, Dagréou et al., 2024]. Moreover, the sample complexity of fully first-order methods for bilevel optimization usually have worse dependency on ϵ [Kwon et al., 2023].

2.3.3. Proof Sketch of Theorem 2.3.1. Define

$$V_k = \frac{1}{\tau^2}\|x_+^k - x^k\|^2 + \|h^k - \nabla\Phi(x^k)\|^2.$$

To obtain (2.7), we consider the merit function:

$$W_k = \Phi(x^k) - \eta_{\mathcal{X}}(x^k, h^k, \tau) + \|y^k - y_*^k\|^2 + \|z^k - z_*^k\|^2,$$

where $\eta_{\mathcal{X}}(x, h, \tau) = \langle h, x_+ - x\rangle + \frac{1}{2\tau}\|x_+ - x\|^2$. By leveraging the moving-average updates of x^k (line 2 of Algorithm 1), we can obtain

$$\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[V_k\right] = \mathcal{O}\left(\sum_{k=0}^{K}\left(\alpha_k\mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Phi(x^k)\right\|^2\right] + \alpha_k^2\right)\right),$$

which reduces the error analysis to controlling the hypergradient estimation bias, i.e., $\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Phi(x^k)\|^2$. This term, by the construction of w^{k+1}, satisfies

$$\sum_{k=0}^K \alpha_k \mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Phi(x^k)\right\|^2\right] = \mathcal{O}\left(\sum_{k=0}^K \alpha_k \mathbb{E}\left[\left\|x_+^k - x^k\right\|^2 + \left\|y^k - y_*^k\right\|^2 + \left\|z^k - z_*^k\right\|^2\right]\right).$$

It is worth noting that Dagréou et al. [2022] requires the existence and Lipschitzness of $\nabla^2 f$ and $\nabla^3 g$ to ensure the Lipschitzness of $\nabla z^*(x)$ (see (2.3)) which is used in proving the sufficient decrease of $\left\|z^k - z_*^k\right\|^2$. In contrast, based on the moving-average updates of x^k and h^k, our refined analysis does not necessitate such assumptions to obtain that

$$\sum_{k=0}^K \alpha_k \mathbb{E}\left[\left\|y^k - y_*^k\right\|^2 + \left\|z^k - z_*^k\right\|^2\right] = \mathcal{O}\Bigg(\sum_{k=0}^K \alpha_k \mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right]\Bigg).$$

The proof of Theorem 2.3.1 can then be completed by choosing appropriate $\alpha_k, c_1, c_2, c_3, \tau > 0$.

2.4. Min-Max Bilevel Optimization

To incorporate robustness in the multi-objective setting where each objective can be expressed as a bilevel optimization problem in (2.1), the following mini-max bilevel problem formulation was proposed in Gu et al. [2023]:

$$\min_{x\in\mathcal{X}} \max_{1\le i\le n} \Phi_i(x) \coloneqq f_i(x, y_i^*(x)) \qquad \text{s.t.} \quad y_i^*(x) = \arg\min_{y_i\in\mathbb{R}^{d_{y_i}}} g_i(x, y_i), 1 \le i \le n. \tag{2.8}$$

Note that (2.8) can be reformulated as a general nonconvex-concave min-max optimization problem (with a bilevel substructure):

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\Delta_n} \Phi(x,\lambda) \coloneqq \sum_{i=1}^n \lambda_i \Phi_i(x). \tag{2.9}$$

Instead of solving (2.9) directly, in this work, we focus on solving the regularized version,

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\Delta_n} \Phi_{\mu_\lambda}(x,\lambda) \coloneqq \Phi(x,\lambda) - \frac{\mu_\lambda}{2}\left\|\lambda - \frac{1}{n}\mathbf{1}_n\right\|^2. \tag{2.10}$$

Note that in (2.10), we include an ℓ^2 regularization term that penalizes the discrepancy between λ and $\frac{\mathbf{1}_n}{n}$. When $\mu_\lambda = 0$, it corresponds to equation (2.8), and as $\mu_\lambda \to +\infty$, it enforces $\lambda = \frac{\mathbf{1}_n}{n}$, leading to direct minimizing of the average loss. It is important to note that minimizing the worst-case loss (i.e., $\max_{1\le i\le n} f_i(x, y_i^*(x))$) does not necessarily imply the minimization of the average loss (i.e.,

$\frac{1}{n}\sum_{i=1}^n f_i(x, y_i^*(x)))$. Therefore, in practice, it may be preferable to select an appropriate $\mu_\lambda > 0$ [Qian et al., 2019, Wang et al., 2021] to strike a balance between these two types of losses. Hu et al. [2022] considered solving a similar problem under stronger assumptions. We defer a detailed discussion to Section A.2.2.

2.4.1. Proposed Framework: the MORMA-SOBA Algorithm. The proposed algorithm, which we refer as to **M**ulti-**O**bjective **R**obust `MA-SOBA` (`MORMA-SOBA`), for solving (2.10) is presented in Algorithm 2. In addition to the basic framework of Algorithm 1, we also maintain a moving-average step in the updates of λ^k for solving the max part of problem 2.4.1. It is worth noting that in its single-level counterpart without the inner variable y, the proposed `MORMA-SOBA` algorithm is fundamentally similar to the single-timescale averaged `SGDA` algorithm proposed in the general nonconvex-strongly-concave setting [Qiu et al., 2020]. Moreover, our algorithmic framework can be leveraged to solve the distributionally robust compositional optimization problem as discussed in Gao et al. [2021].

Remark (Comparison with `MORBiT` [Gu et al., 2023]). In contrast to our approach in (2.10), the work of Gu et al. [2023], for the min-max bilevel problem, attempted to combine `TTSA` [Hong et al., 2023] and `SGDA` [Lin et al., 2020a] to solve the nonconvex-concave problem as (2.9). However, *we identified an issue in* Gu et al. [2023] related to the ambiguity and inconsistency in the expectation and filtration, which may not be easily resolved within their current proof framework. As a consequence, their current proof is unable to demonstrate $\mathbb{E}[\max_{i\in[n]} \|y_i^k - y_i^*(x^{(k-1)})\|^2] \leq \tilde{\mathcal{O}}(\sqrt{n}K^{-2/5})$ as claimed in Theorem 1 (10b) of Gu et al. [2023]. Thus, the subsequent arguments made regarding the convergence analysis of x and λ are incorrect (at least in its current form); see Section A.2 for further discussions. Moreover, the practical implementation of `MORBiT` incorporates momentum and weight decay techniques to optimize the simplex variable λ. This approach can be seen as a means of solving the regularized formulation in (2.10).

2.4.2. Convergence Results. We first present additional assumptions required in the analysis of `MORMA-SOBA`.

Assumption 4. *For any $k \geq 0$, functions $\Phi(x), \nabla\Phi_i(x)$ are bounded, functions f_i are L_f-Lipschitze continuous in the second input, and their stochastic versions are unbiased with bounded*

Algorithm 2: `Multi-Objective Robust Moving-Average SOBA`

Input: $x^0, \lambda^0, \{y_i^0\}, \{z_i^0\}, h_x^0 = 0, h_\lambda^0 = 0, \{\alpha_k\}, \{\beta_k\}, \{\gamma_k\}, \{\theta_k\}$

for $k = 0, 1, \ldots, K-1$ **do**

  $x^{k+1} = x^k + \alpha_k \left(\Pi_{\mathcal{X}}(x^k - \tau_x h_x^k) - x^k\right)$ # update x^k via average hypergradient h_x^k

  $\lambda^{k+1} = \lambda^k + \alpha_k \left(\Pi_{\Delta_n}(\lambda^k + \tau_\lambda h_\lambda^k) - \lambda^k\right)$ # update λ^k via average gradient h_λ^k

  for $i = 1, \ldots, n$ *(in parallel)* **do**

    $y_i^{k+1} = y_i^k - \beta_k v_i^{k+1}$ # update y_i^k by one-step SGD based on (2.4)

    $z_i^{k+1} = z_i^k - \gamma_k\left(H_i^{k+1} z_i^k - u_{y,i}^{k+1}\right)$ # update z_i^k by one-step SGD based on (2.5)

  end

  $h_x^{k+1} = (1-\theta_k) h_x^k + \theta_k \sum_{i=1}^n \lambda_i^k \left(u_{x,i}^{k+1} - J_i^{k+1} z_i^k\right)$

  # update average hypergradient h_x^k

  $h_\lambda^{k+1} = (1-\theta_k) h_\lambda^k + \theta_k \left(s^{k+1} - \mu_\lambda\left(\lambda^k - \frac{\mathbf{1}_n}{n}\right)\right)$ # update average gradient h_λ^k

end

variance, i.e., there exists $L_\Phi, L_f, \sigma_{f,0} \geq 0$ *such that*

$$|\Phi_i(x)| \leq b_\Phi,\ \|\nabla \Phi_i(x)\| \leq L_\Phi,\ |f_i(x,y) - f_i(x,\tilde{y})| \leq L_f \|y - \tilde{y}\|,\ \ \textit{for all } x, y, \tilde{y}, 1 \leq i \leq n,$$

$$s^{k+1} = \left(s_1^{k+1}, ..., s_n^{k+1}\right)^\top,\ \mathbb{E}\left[s_i^{k+1} | \mathscr{F}_k\right] = f_i(x^k, y_i^k),\ \mathbb{E}\left[\left\|s_i^{k+1} - f_i(x^k, y_i^k)\right\|^2 | \mathscr{F}_k\right] \leq \sigma_{f,0}^2.$$

$\bigcup_{i=1}^n \left\{u_{x,i}^{k+1}, u_{y,i}^{k+1}, v_i^{k+1}, H_i^{k+1}, J_i^{k+1}\right\} \cup \left\{s^{k+1}\right\}$ *are conditionally independent under* $\mathscr{F}_k$.

We have the following convergence theorem of `MORMA-SOBA`.

THEOREM 2.4.1. *Suppose Assumptions 2, 3 (for all* f_i, g_i*) and Assumption 4 hold. Then there exist positive constants* $c_1, c_2, c_3, \tau_x, \tau_\lambda > 0$ *such that if* $\alpha_k \equiv \Theta(1/\sqrt{nK}), \beta_k = c_1\alpha_k, \gamma_k = c_2\alpha_k, \theta_k = c_3\alpha_k, \mu_\lambda < 1$ *in Algorithm 2, then the iterates in Algorithm 2 satisfy*

$$\frac{1}{K}\sum_{k=0}^{K} \frac{1}{\tau_x^2} \mathbb{E}\left[\left\|x^k - \Pi_{\mathcal{X}}\left(x^k - \tau_x \nabla \Psi_{\mu_\lambda}(x^k)\right)\right\|^2\right] = \mathcal{O}\left(\frac{n^2}{\mu_\lambda^2 \sqrt{K}}\right),$$

where $\Psi_{\mu_\lambda}(x) := \max_{\lambda \in \Delta_n} \Phi_{\mu_\lambda}(x, \lambda)$. *That is to say, when uniformly randomly selecting a solution* x^R *from* $\{x^1, \ldots, x^K\}$, *the sample complexity (the total number of calls to stochastic oracles) of finding an* ϵ*-stationary point by Algorithm 2 is* $\mathcal{O}(n^5 \mu_\lambda^{-4} \epsilon^{-2})$.

Theorem 2.4.1 indicates that Algorithm 2 is capable of generating an ϵ-first-order stationary point of $\min_x \Psi_{\mu_\lambda}(x)$ with $K \gtrsim n^5 \mu_\lambda^{-4} \epsilon^{-2}$. As $\mu_\lambda \to 0$, the problem (2.10) changes towards the nonconvex-concave problem (2.9) and the sample complexity becomes worse, which to some extent implies the difficulty of directly solving (2.9). We defer the proof details to Section A.1.2. For

Problem (2.9), we adopt the definition of ϵ-stationary point in Definition 3.5 in Lin et al. [2020b], and choose $\mu_\lambda = \mathcal{O}(\sqrt{\epsilon})$ to help shed light on the sample complexity.

COROLLARY 2.4.1. *Under the same setup of Theorem 2.4.1, setting $\mu_\lambda = \mathcal{O}(\sqrt{\epsilon})$, the sample complexity of finding an ϵ-stationary point of Problem (2.9) via Algorithm 2 is $\mathcal{O}(n^5\epsilon^{-4})$.*

REMARK. Note that in Theorem 2.4.1 we explicitly characterize the dependency on n and μ_λ in the convergence rate and the sample complexity. It is worth noting that two variants of stochastic gradient descent ascent (`SGDA`) algorithms for solving the nonconvex-strongly-concave min-max optimization problems (without bilevel substructures), have been studied in Lin et al. [2020a], Qiu et al. [2020]. While such algorithms are not immediately applicable to solve (2.10) due to the presence of the additional bilevel substructure, it is instructive to compare to those methods assuming direct access to $y_i^*(x)$ in (2.8). Specifically, we observe that the sample complexity of `SGDA` with batch size $M = \Theta(n^{1.5}\epsilon^{-1})$ in Lin et al. [2020a] and moving-average `SGDA` with $\mathcal{O}(1)$ batch size in Qiu et al. [2020] for solving (2.10) assuming direct access to $y_i^*(x)$ will be $\mathcal{O}\left(n^4\mu_\lambda^{-2}\epsilon^{-2}\right)$ and $\mathcal{O}\left(n^5\mu_\lambda^{-4}\epsilon^{-2}\right)$* respectively. Our results in Theorem 2.4.1 indicate that the sample complexity of the proposed algorithm `MORMA-SOBA` for solving min-max bilevel problems has the same dependency on n and μ_λ as the sample complexity of the moving-average `SGDA` introduced in Qiu et al. [2020] for solving min-max single-level problems, while also computing $y_i^*(x)$ instead of assuming direct access.

2.5. Experiments

While our contributions primarily focus on theoretical aspects, we also conducted experiments to validate our results. We first compare the performance of `MA-SOBA` with other benchmark methods on two common tasks proposed in previous works [Ji et al., 2021, Hong et al., 2023, Dagréou et al., 2022], *hyperparameter optimization* for ℓ^2 penalized logistic regression and *data hyper-cleaning* on the corrupted MNIST data set. To demonstrate the practical performance of `MORMA-SOBA`, we then conduct experiments in *robust multi-task representation learning* introduced in Gu et al. [2023] on the FashionMNIST data set [Xiao et al., 2017].

2.5.1. Experimental Details for MA-SOBA. Our experiments for `MA-SOBA` are performed with the aid of the recently developed package `Benchopt` [Moreau et al., 2022] and the open-sourced

*Note that $\Phi_{\mu_\lambda}(x,\lambda)$ in (2.9) is quadratic in λ, and these two sample complexities are obtained under this special case, i.e., $\nabla_2^2 f(x,y) = -\mu\mathbf{I}$ applied to Lin et al. [2020a], Qiu et al. [2020].

bilevel optimization benchmark[†]. For a fair comparison, we exclusively consider benchmark methods that do not utilize variance reduction techniques in Table 2.1: (i) `BSA` [Ghadimi and Wang, 2018]; (ii) `stocBiO` [Ji et al., 2021]; (iii) `TTSA` [Hong et al., 2023]/`ALSET` [Chen et al., 2021a]; (iv) `SOBA` [Dagréou et al., 2022]. Noting that `ALSET` only differs from `TTSA` regarding time scales, we use `TTSA` to represent this class of approach. Also, we omit the comparison with `AmIGO` [Arbel and Mairal, 2022] below, given that it is essentially a double-loop `SOBA` with increasing batch sizes. The tunable parameters in benchmark methods are selected in the same manner as those in `benchmark_bilevel`[†].

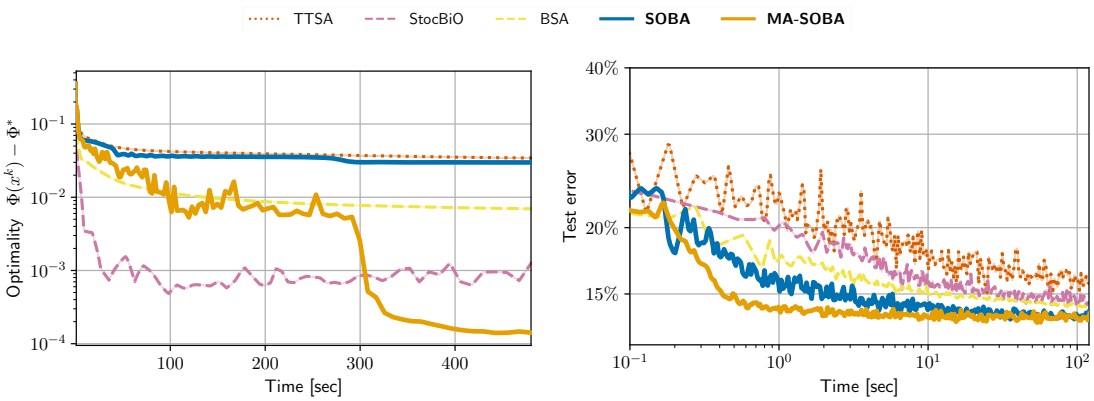
Setup. We strictly adhere to the settings provided in `benchmark_bilevel`, as detailed in Appendix B.1 of Dagréou et al. [2022]. The previous results and setups of Dagréou et al. [2022] have also been available in `https://benchopt.github.io/results/benchmark_bilevel.html`. For completeness, we provide a summary of the setup below.

- To avoid redundant computations, we utilize oracles for functions F_ξ, G_ϕ, which provide access to quantities such as $\nabla_1 F_\xi(x,y)$, $\nabla_2 F_\xi(x,y)$, $\nabla_2 G_\phi(x,y)$, $\nabla^2_{22} G_\phi(x,y)v$, and $\nabla^2_{12} G_\phi(x,y)v$, although this may violate the independence assumption in Assumption 3.
- In all our experiments, we employ a batch size of 64 for all methods, even for `BSA` and `AmIGO` that theoretically require increasing batch sizes.
- For methods involving an inner loop (`stocBiO`, `BSA`, `AmIGO`), we perform 10 inner steps per each outer iteration as proposed in those papers.
- For methods that involve Neumann approximation for Hessian-vector product (such as `BSA`, `TTSA`, `SUSTAIN`, and `MRBO`), we perform 10 steps of the subroutine per outer iteration. For `AmIGO`, we perform 10 steps of SGD to approximate the inversion of the linear system.
- The step sizes and momentum parameters used in all benchmark algorithms are directly adopted from the fine-tuned parameters provided by Dagréou et al. [2022]. From a grid search, we select the best constant step sizes for `MO-SOBA`.

We have excluded `SRBA` [Dagréou et al., 2023] from the benchmark due to its limited reported improvement over `SABA`.

2.5.1.1. *Hyperparameter Optimization on IJCNN1.* In the first task, we fit a multi-regularized logistic regression model (for binary classification), and select the regularization parameters (one

[†] `https://github.com/benchopt/benchmark_bilevel`



(a) ℓ^2 Penalized Logistic Regression on IJCNN1 (b) Data Hyper-Cleaning on MNIST

FIGURE 2.1. Comparison of `MA-SOBA` with other stochastic bilevel optimization methods without using variance reduction techniques. For each algorithm, we plot the median performance over 10 runs. **Left:** Hyperparameter optimization for ℓ^2 penalized logistic regression on IJCNN1 data set. **Right.** Data hyper-cleaning on MNIST with $p = 0.5$ (corruption rate).

hyperparameter per feature) on the IJCNN1 data set[‡]. The functions f and g of the problem (2.1) are the average logistic loss on the validation set and training set respectively, with ℓ^2 regularization for g. Specifically, the problem can be formulated as:

$$\min_{\nu \in \mathbb{R}^d} \quad \Phi(\nu) := \underbrace{\mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}_{\text{val}}} \left[\ell\left(\langle \omega^*(\nu), X\rangle, Y\right)\right]}_{f(\nu, \omega^*(\nu))}$$

$$\text{s.t.} \quad \omega^*(\nu) = \arg\min_{\omega \in \mathbb{R}^d} \underbrace{\mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}_{\text{train}}} \left[\ell\left(\langle \omega, X\rangle, Y\right)\right] + \frac{1}{2}\omega^\top \operatorname{diag}\left(e^{\nu_1}, \ldots, e^{\nu_d}\right)\omega}_{g(\nu,\omega)}.$$

In this case, $|\mathcal{D}_{\text{train}}| = 49,990$, $|\mathcal{D}_{\text{val}}| = 91,701$, and $d = 22$. For each sample, the covariate and label are denoted as (X, Y), where $X \in \mathbb{R}^{22}$ and $Y \in \{0, 1\}$. The inner variable ($\omega \in \mathbb{R}^{22}$) is the regression coefficient. The outer variable ($\nu \in \mathbb{R}^{22}$) is a vector of regularization parameters. The loss function $\ell(y', y) = -y \log(y') - (1 - y) \log(1 - y')$ is the log loss.

In Figure 2.1(a), we plot the suboptimality gap against the runtime for each method. Surprisingly, we observed that `MA-SOBA` achieves lower objective values after several iterations compared to

[‡] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

all benchmark methods. This improvement can be attributed to the convergence of average hypergradients $\{h^k\}$. These findings demonstrate the practical superiority of our algorithm framework, even with the same sample complexity results.

To supplement the comparison, we conducted additional experiments that involved comparing all benchmark methods, including the variance reduction based method. In Figure 2.2, we plot the suboptimality gap $(\Phi(x) - \Phi^*)$ against runtime and the number of calls to oracles. Unfortunately, the previous results obtained for `MRBO` and `AmIGO` on the IJCNN1 data set are not reproducible at the moment due to some conflicts in the current developer version of `Benchopt`. As reported in Dagréou et al. [2022], `MRBO` exhibits similar performance to `SUSTAIN`, while the curve of `AmIGO` initially follows a similar trend as `SUSTAIN` and eventually reaches a similar level as `SABA` towards the end. Following a grid search, we have selected the parameters in `MA-SOBA` as $\alpha_k\tau = 0.02$, $\beta_k = \gamma_k = 0.01$, and $\theta_k = 0.1$. As shown in Figure 2.2, our proposed method `MA-SOBA` outperforms `SOBA` significantly, achieving a slightly lower suboptimality gap compared to the state-of-the-art variance reduction-based method `SABA`.

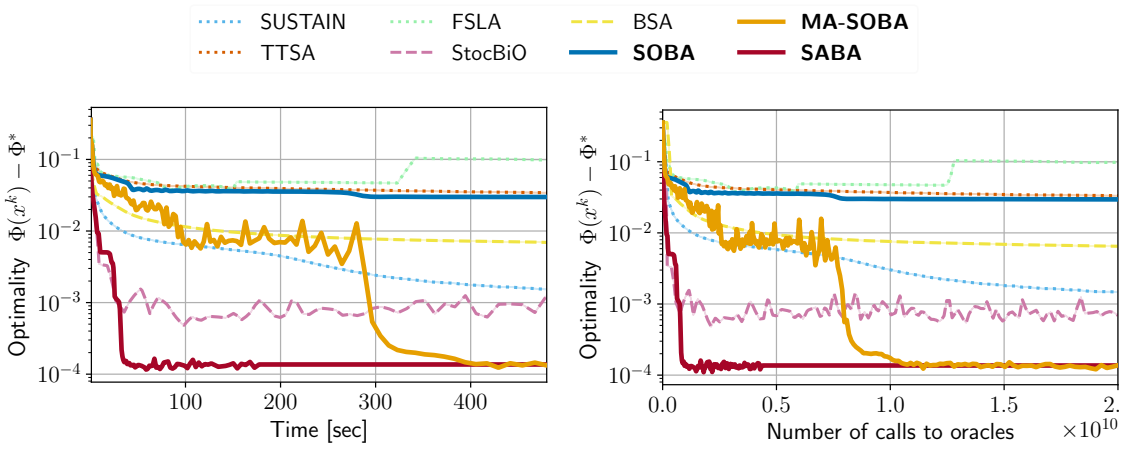


FIGURE 2.2. Comparison of `MA-SOBA` with other stochastic bilevel optimization methods in the problem of hyperparameter optimization for ℓ^2 regularized logistic regression on the IJCNN1 data set. We plot the median performance over 10 runs for each method. **Left:** Performance in runtime; **Right:** Performance in the number of gradient/Hessian(Jacobian)-vector products sampled.

2.5.1.2. *Data Hyper-Cleaning on MNIST.* In the second task, we conduct data hyper-cleaning on the MNIST data set introduced in Franceschi et al. [2017]. Data cleaning aims to train a multinomial logistic regression model on the corrupted training set and determine a weight for

each training sample. These weights should approach zero for samples with corrupted labels. The data set is partitioned into a training set $\mathcal{D}_{\text{train}}$, a validation set $\mathcal{D}_{\text{val}}$, and a test set $\mathcal{D}_{\text{test}}$, where $|\mathcal{D}_{\text{train}}| = 20,000$, $|\mathcal{D}_{\text{val}}| = 5,000$, and $|\mathcal{D}_{\text{test}}| = 10,000$. Each sample is represented as a vector X of dimension 784, where the input image is flattened. The corresponding label takes values from the set $\{0, 1, \ldots, 9\}$. We use $Y \in \mathbb{R}^{10}$ to denote its one-hot encoding. Each sample in the training set is corrupted with probability p by replacing its label with a random label $\{0, 1, \ldots, 9\}$.

The task can be formulated into the bilevel optimization problem (2.1) with the inner variable y being the regression coefficients and the outer variable x being the sample weight. The `LL` function g is the sample-weighted cross-entropy loss on the corrupted training set with ℓ^2 regularization. The `UL` function f is the cross-entropy loss on the validation set. Precisely, the task can be formulated into the bilevel optimization problem as below:

$$\min_{\nu \in \mathbb{R}^{|\mathcal{D}_{\text{train}}|}} \quad \Phi(\nu) \coloneqq \underbrace{\mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}_{\text{val}}} [\ell(W^*(\nu)X, Y)]}_{f(\nu, W^*(\nu))}$$

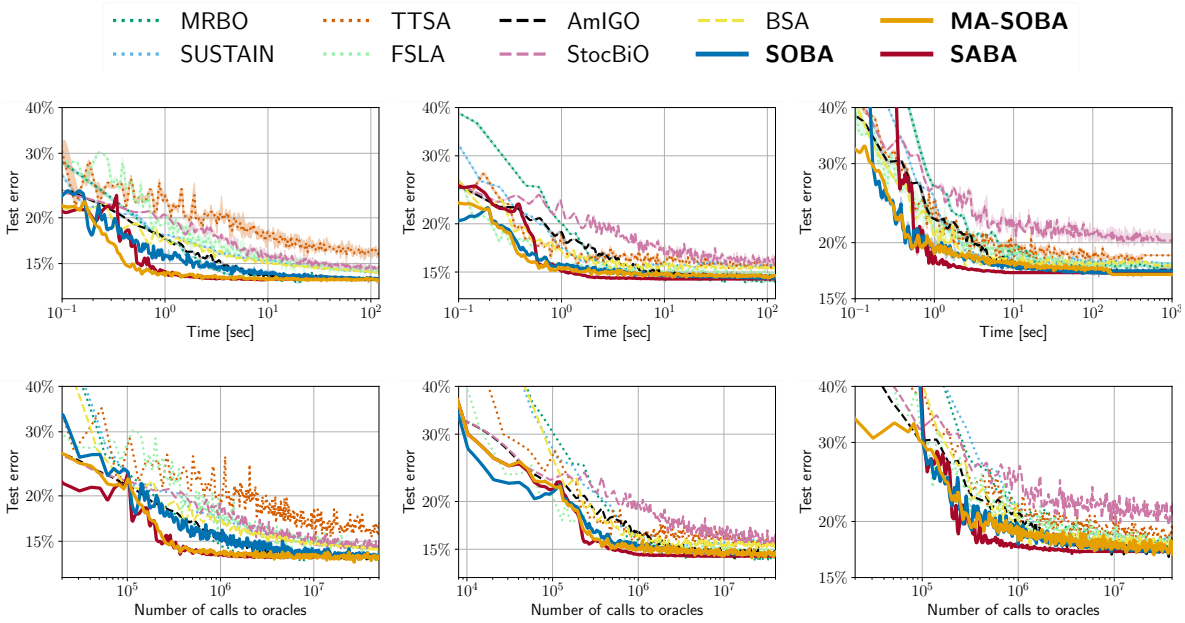
$$\text{s.t.} \quad W^*(\nu) = \arg\min_{\omega \in \mathbb{R}^d} \underbrace{\frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(X_i,Y_i)\sim\mathcal{D}_{\text{train}}} \sigma(\nu_i)\ell(WX_i, \overbrace{\widetilde{Y}_i}^{\text{corrupted}}) + C_r \|W\|^2}_{g(\nu, W)},$$

where the outer variable ($\nu \in \mathbb{R}^{20,000}$) is a vector of sample weights for the training set, the inner variable $W \in \mathbb{R}^{10\times 784}$, and ℓ is the cross entropy loss and σ is the sigmoid function. The regularization parameter $C_r = 0.2$ following Dagréou et al. [2022]. The objective of data hyper-cleaning is to train a multinomial logistic regression model on the training set and determine a weight for each training sample using the validation set. The weights are designed to approach zero for corrupted samples, thereby aiding in the removal of these samples during the training process.

We report the test error in Figure 2.1(b). We observe that `MA-SOBA` outperforms other benchmark methods by achieving lower test errors faster.

To supplement the comparison, we conducted additional experiments that involved comparing all benchmark methods, including the variance reduction-based method. Following a grid search, we have selected the parameters in `MA-SOBA` as $\alpha_k\tau = 10^3$, $\beta_k = \gamma_k = 10^{-2}$, and $\theta_k = 10^{-1}$. In Figure 2.3, we plot the test error against runtime and the number of calls to oracles with different corruption probability $p \in \{0.5, 0.7, 0.9\}$. We observe that `MA-SOBA` has comparable performance to

the state-of-the-art method SABA. Remarkably, MA-SOBA is the fastest algorithm to reach the best test accuracy when $p = 0.5$.



(a) $p = 0.5$ (b) $p = 0.7$ (c) $p = 0.9$

FIGURE 2.3. Comparison of MA-SOBA with other stochastic bilevel optimization methods in the problem of data hyper-cleaning on the MNIST data set when the corruption probability $p \in \{0.5, 0.7, 0.9\}$. We plot the median performance over 10 runs for each method. **Top:** Performance in runtime; **Bottom:** Performance in the number of gradient/Hessian(Jacobian)-vector products sampled.

2.5.2. Experimental Details for MORMA-SOBA. To demonstrate the practical performance of MORMA-SOBA as compared to MORBiT [Gu et al., 2023], we conduct experiments in *robust multi-task representation learning* introduced in Gu et al. [2023] on the FashionMNIST data set [Xiao et al., 2017]. We adopt the same setup as described in Gu et al. [2023], which can be summarized as follows.

Setup. We consider binary classification tasks generated from FashionMNIST where we select 8 "easy" tasks (lowest loss ~ 0.3 from independent training) and 2 "hard" tasks (lowest loss ~ 0.45 from independent training) for multi-objective robust representation learning:

- "easy" tasks: (0, 9), (1, 7), (2, 7), (2, 9), (4, 7), (4, 9), (3, 7), (3, 9)
- "hard" tasks: (0, 6), (2, 4)

For each task $i \in [10]$ above, we partition its data set into the training set $\mathcal{D}_i^{\text{train}}$, validation set $\mathcal{D}_i^{\text{val}}$, and test set $\mathcal{D}_i^{\text{test}}$. We also generate 7 (unseen) binary classification tasks for testing:

- "easy" tasks: (1, 9), (2, 5), (4, 5), (5, 6)
- "hard" tasks: (2, 6), (3, 6), (4, 6)

We train a shared representation network that maps the 784-dimensional (vectorized 28x28 images) input to a 100-dimensional space. To learn a shared representation and per-task models that generalize well on each task, we aim to solve the following problem:

$$\min_{E \in \mathbb{R}^{100\times 784}} \max_{1\le i\le n} \Phi_i(E) := \underbrace{\mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}_i^{\text{val}}} \left[\ell \left(W_i^*(E) \circ \overbrace{\text{ReLU}(EX)}^{\text{representation}} + b_i^*(E), Y \right) \right]}_{f_i(E,(W_i^*,b_i^*))}$$

$$\text{s.t.} \quad \begin{pmatrix} W_i^*(E) \\ b_i^*(E) \end{pmatrix} =$$

$$\underset{W_i\in\mathbb{R}^{10\times 100}, b_i\in\mathbb{R}^{10}}{\arg\min} \underbrace{\mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}_i^{\text{train}}} \left[\ell \left(\overbrace{W_i}^{\text{weight}} \circ \text{ReLU}(EX) + \overbrace{b_i}^{\text{bias}}, Y \right) \right] + \rho \left\| W_i \right\|_F^2}_{g_i(E,(W_i,b_i))}, 1 \le i \le n.$$

Each bilevel objective Φ_i in this setup represents a distinct binary classification "task" $i \in [n]$ with its own training and validation sets. The optimization variable is engaged in a shared representation network, parameterized by the outer variable $E \in \mathbb{R}^{100\times 784}$, along with per-task linear models parameterized by each inner variable (W_i, b_i). The `UL` function f_i is the average cross-entropy loss over the $\mathcal{D}_i^{\text{val}}$, and the `LL` function g_i is the ℓ^2 regularized cross-entropy loss over $\mathcal{D}_i^{\text{train}}$. Each sample is represented as a vector X of dimension 784, where the input image is flattened. The corresponding label takes values from the set $\{0, 1, \ldots, 9\}$. We use $Y \in \mathbb{R}^{10}$ to denote its one-hot encoding.

In the experiment, the regularization parameter in the `LL` function $\rho = 5 \times 10^{-4}$. The implementation of `MORBiT` follows the same manner described in Gu et al. [2023]. Specifically, the code of `MORBiT` [Gu et al., 2023] uses vanilla SGD with a learning rate scheduler and incorporates momentum and weight decay techniques to optimize each variable:

- Outer variable: learning rate = 0.01, momentum = 0.9, weight_decay = 10^{-4}
- Inner variable: learning rate = 0.01, momentum = 0.9, weight_decay = 10^{-4}
- Simplex variable: learning rate = 0.3, momentum = 0.9, weight_decay = 10^{-4}

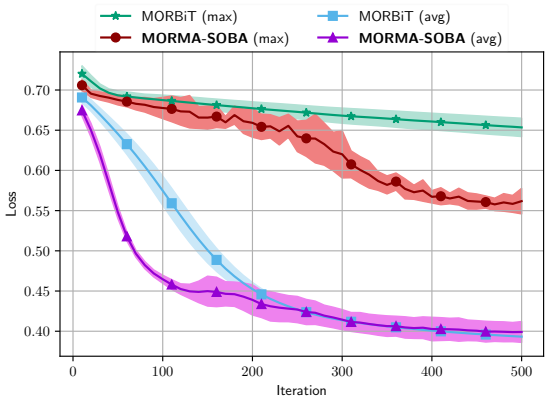


FIGURE 2.4. `MORMA-SOBA` ($\mu_\lambda = 0.01$) vs. `MORBiT` on robust multi-task representation learning.

In addition, `MORBiT` adopts a straightforward iterative auto-differentiation to calculate the hypergradient without using Neumann approximation of the Hession inversion.

For the implementation of `MORMA-SOBA`, the regularization parameter μ_λ in 2.10 is set to be 0.01. All remaining parameters are chosen as constant values, as listed below:

- Outer variable: $\tau_x = 1, \alpha_k = 0.02$,
- Inner variable: $\beta_k = 0.02$
- Auxiliary variable: $\gamma_k = 0.02$
- Simplex variable: $\tau_\lambda = 1, \alpha_k = 0.02$
- Average gradient: $\theta_k = 0.6$

Both evaluated methods use batch sizes of 8 and 128 to compute g_i for each inner step and f_i for each outer iteration, respectively.

In Figure 2.4, we compare our algorithm with the existing min-max bilevel algorithm `MORBiT` [Gu et al., 2023] in terms of the average loss ($(1/n)\sum_i \Phi_i$) and maximum loss ($\max_i \Phi_i$). The results demonstrate the superiority of `MORMA-SOBA` over `MORBiT` in terms of lowering both the max loss and average loss at a faster rate. In addition to Figure 2.4, which showcases the performance on 10 seen tasks used for representation learning, we present Figure 2.5. This figure displays the maximum/average loss values against the number of iterations on test sets consisting of 10 seen tasks and 7 unseen tasks. Our approach, `MORMA-SOBA`, demonstrates superior performance in terms of faster reduction of both maximum and average loss.

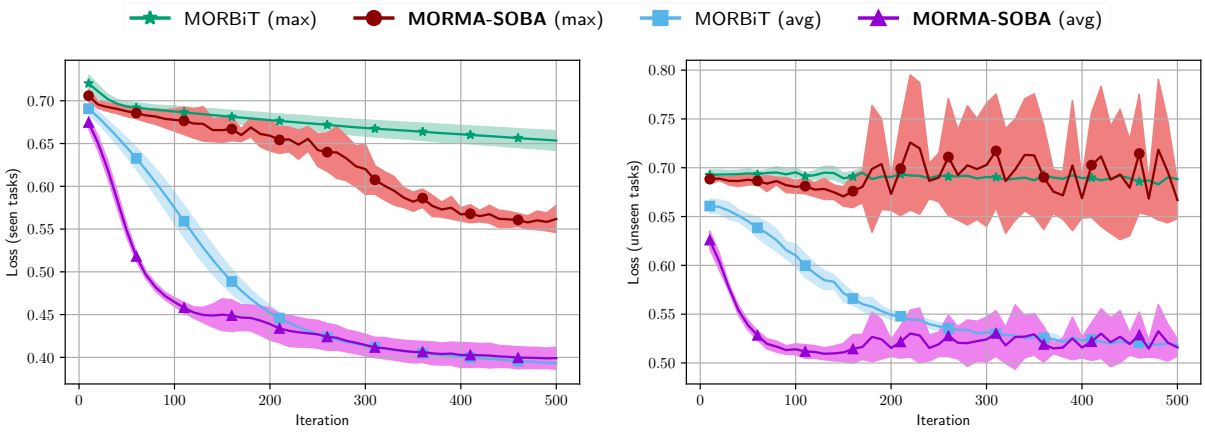


FIGURE 2.5. Comparison of `MORMA-SOBA` with `MORBiT` in the problem of multi-objective robust representation learning for binary classification tasks on the FashionMNIST data set. We aggregate the results over 10 runs for each method. **Left:** Performance on test sets of seen tasks; **Right:** Performance on unseen tasks.

2.5.3. Moving Average vs. Variance Reduction. Through empirical studies, we have demonstrated that our proposed method, `MA-SOBA`, achieves comparable performance to the state-of-the-art variance reduction-based approach `SABA` using `SAGA` updates [Defazio et al., 2014]. In this context, we would like to highlight the key difference and relationship between these two methods.

We start with presenting the update rules of the sequence of estimated gradients $\{g^k\}$ for the variance reduction techniques `SAGA` [Defazio et al., 2014] and our moving-average method (`MA`) for the single-level problem:

SAGA (finite-sum:) $\min \ \frac{1}{n}\sum_{i=1}^{n} f_i(x)$

$$g^k = \nabla f_{i_k}(x^k) - \nabla f_{i_k}(\bar{x}_{i_k}) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(\bar{x}_j)$$

The `SAGA` update is designed for finite-sum problems with offline batch data. At each iteration k, the algorithm randomly selects an index $i_k \in [n]$ and updates the gradient variable g^k using a reference point $\bar{x}_{i_k}$, which corresponds to the last evaluated point for ∇f_{i_k}. However, it should be noted that `SAGA` requires storing the previously evaluated gradients $\nabla f_j(\bar{x}_j)$ in a table, which can be memory-intensive when sample size n or dimension d is large. In the finite-sum setting, there exist several other variance reduction methods, such as `SARAH` [Dagréou et al., 2023], that can be employed to further enhance the dependence on the number of samples, n, for bilevel optimization

problems. However, the SARAH-type method requires double gradient evaluations on each iteration of x^k and x^{k-1}:

MA (expectation): min $\mathbb{E}_\xi[f(x;\xi)]$

$$g^k = (1-\alpha_k)g^{k-1} + \alpha_k \nabla f(x^k;\xi^{k+1})$$

Unlike variance reduction techniques, the moving-average methods can solve the general expectation-form problem with online and streaming data using a simple update per iteration. In addition, the moving-average techniques offer two more advantages:

Theoretical Assumption. All variance reduction methods, including SVRG [Reddi et al., 2016], SAGA [Defazio et al., 2014], SARAH [Nguyen et al., 2017], STORM [Cutkosky and Orabona, 2019], and others, typically rely on assuming mean-squared smoothness assumptions. In particular, for stochastic optimization problems in the form of $\min_x\{f(x) = \mathbb{E}[F(x,\xi)]\}$, the definition of mean-squared smoothness (MSS) is: (MSS) $\mathbb{E}_\xi[\|\nabla F(x,\xi) - \nabla F(x',\xi)\|^2] \leq L^2\|x-x'\|^2$. However, MSS is a stronger assumption than the general smoothness assumption on f: $\|\nabla f(x) - \nabla f(x')\| \leq L\|x-x'\|$. By Jensen's inequality, we have that MSS is stronger than the general smoothness assumption on f: $\|\nabla f(x) - \nabla f(x')\|^2 \leq \mathbb{E}_\xi[\|\nabla F(x,\xi) - \nabla F(x',\xi)\|^2]$. In this work, the theoretical results of the proposed methods are only built on the smoothness assumption on the UL and LL functions f, g without further assuming MSS on F_ξ and G_ϕ. It is worth noting that a clear distinction in the lower bounds of sample complexity for solving the single-level stochastic optimization has been proven in Arjevani et al. [2023]. Specifically, they establish a separation under the MSS assumption on F_ξ and smoothness assumptions on f ($\mathcal{O}(\epsilon^{-1.5})$ vs. $\mathcal{O}(\epsilon^{-2})$). Thus, it is important to emphasize that MA-SOBA achieves the optimal sample complexity $\mathcal{O}(\epsilon^{-2})$ under our weaker assumptions.

Practical Implementation. Variance reduction methods often entail additional space complexity, require double-loop implementation or double oracle computations per iteration. These requirements can be unfavorable for large-scale problems with limited computing resources. For instance, in the second task, the runtime improvement achieved by using SABA is limited. This limitation can be attributed to the dimensionality of the variables ν (with a dimension of $20,000$) and W (with a dimension of 10×784). The benefit of using variance reduction methods is expected to be less significant for more complex problems involving computationally expensive oracle evaluations.

2.6. Conclusion

In this work, we propose a novel class of algorithms (`MA-SOBA`) for solving stochastic bilevel optimization problems in (2.1) by introducing the moving-average step to estimate the hypergradient. We present a refined convergence analysis of our algorithm, achieving the optimal sample complexity without relying on the high-order smoothness assumptions employed in the literature. Furthermore, we extend our algorithm framework to tackle a generic min-max bilevel optimization problem within the multi-objective setting, identifying and addressing the theoretical gap present in the literature.

CHAPTER 3

Decentralized Stochastic Bilevel Optimization

3.1. Introduction

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

$$\begin{aligned} &\min_{x\in\mathbb{R}^p} \quad \Phi(x) = f(x, y^*(x)) \\ &\text{s.t.} \quad y^*(x) = \arg\min_{y\in\mathbb{R}^q} g(x, y), \end{aligned} \tag{3.1}$$

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., 2021a] to modern machine learning problems such as reinforcement learning [Hong et al., 2020], meta learning [Snell et al., 2017, Bertinetto et al., 2019, Rajeswaran et al., 2019, Ji et al., 2020], 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., 2021a], to name a few.

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, Altae-Tran et al. [2017], Zhang et al. [2019], 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:

$$\begin{aligned} \min_{x\in\mathbb{R}^p} \quad & \Phi(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x, y^*(x)) \\ \text{s.t.} \quad & y^*(x) = \arg\min_{y\in\mathbb{R}^q} g(x,y) := \frac{1}{n}\sum_{i=1}^{n} g_i(x,y), \end{aligned} \tag{3.2}$$

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].

3.1.1. Our contributions. Our contributions in this work are as follows.

- 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 (both computational and communication) 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 5), and in Theorem 3.3.2 and Corollary 3.3.1 we show the sample complexity is of order $\mathcal{O}(\epsilon^{-2}\log\frac{1}{\epsilon})$, which matches the currently well-known results of the single-agent bilevel optimization [Chen et al., 2021a]. 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].

3.1.2. 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., 2019, 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., 2021a, 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 $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity [Chen et al., 2021a, Arbel and Mairal, 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., 2021a, 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 stochastic bilevel optimization. FEDNEST [Tarzanagh et al., 2022] and FedBiO [Li et al., 2022b] impose federated learning, which is essentially a centralized setting, on stochastic bilevel optimization. Existing work on DSBO can be classified to two categories: global DSBO and personalized DSBO. Problem (3.2) that we consider in this paper is a global DSBO, where both lower-level and upper-level functions are not directly accessible to any local agent. Other works on global DSBO include Chen et al. [2022b], Yang et al. [2022], Gao et al. [2022]*. The personalized DSBO [Lu 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 (3.2), which leads to

$$\begin{aligned} &\min_{x\in\mathbb{R}^p} \quad \Phi(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x, y_i^*(x)) \\ &\text{s.t.} \quad y_i^*(x) = \arg\min_{y\in\mathbb{R}^q} g_i(x,y), i = 1,\ldots,n. \end{aligned} \tag{3.3}$$

To solve global DSBO (3.2), 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]. 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., 2021a] to avoid computing the Hessian inverse. In personalized DSBO (3.3), 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

*Here we point out that although Gao et al. [2022] claim that they solve the global DSBO, based on equations (2) and (3) in their paper (https://arxiv.org/abs/2206.15025v1), it is clear that they are only solving a special case of global DSBO problem. See appendix A.5.2 for detailed discussion.

TABLE 3.1. We compare our Algorithm 5 (MA-DSBO) with existing distributed bilevel optimization algorithms: SPDB [Lu et al., 2022], DSBO-JHIP [Chen et al., 2022b], and GBDSBO [Yang et al., 2022]. The problem types include Personalized-Decentralized Stochastic Bilevel Optimization (P-DSBO), and Global-Decentralized Stochastic Bilevel Optimization (G-DSBO). In the table we define $d = \max(p, q)$. 'Computation' (See Section A.5.3 for details) and 'Samples' represent the computational and sample complexity of finding an ϵ-stationary point, respectively. $\tilde{\mathcal{O}}$ hides the $\log(\frac{1}{\epsilon})$ factor. '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}.

ALGORITHM	PROBLEM	COMPUTATION	SAMPLES	JACOBIAN	MINI-BATCH
SPDB	P-DSBO	$\tilde{\mathcal{O}}(dn^{-1}\epsilon^{-2})$	$\tilde{\mathcal{O}}(n^{-1}\epsilon^{-2})$	NO	YES
DSBO-JHIP	G-DSBO	$\tilde{\mathcal{O}}(pq\epsilon^{-3})$	$\tilde{\mathcal{O}}(\epsilon^{-3})$	YES	NO
GBDSBO	G-DSBO	$\mathcal{O}((q^2\log(\frac{1}{\epsilon})+pq)n^{-1}\epsilon^{-2})$	$\tilde{\mathcal{O}}(n^{-1}\epsilon^{-2})$	YES	NO
MA-DSBO	G-DSBO	$\tilde{\mathcal{O}}(d\epsilon^{-2})$	$\tilde{\mathcal{O}}(\epsilon^{-2})$	NO	NO

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 3.1.

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, $\bar{x}_k = \frac{1}{n}\sum_{i=1}^n x_{i,k}$, $\bar{y}_k^{(t)} = \frac{1}{n}\sum_{i=1}^n y_{i,k}^{(t)}$.

3.2. Preliminaries

The following assumptions are used throughout this paper. They are standard assumptions that are made in the literature on bilevel optimization [Ghadimi and Wang, 2018, Hong et al., 2020, Chen et al., 2021a, Ji et al., 2021, Huang et al., 2023] and decentralized optimization [Qu and Li, 2017, Nedic et al., 2017, Lian et al., 2017, Tang et al., 2018].

ASSUMPTION 5 (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 6 (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 6 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' [Lian et al., 2017], 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, which eventually guarantees the consensus among agents.

ASSUMPTION 7 (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

$$\begin{aligned}\|\nabla_y g(x, y)\|^2 \leq& \mathbb{E}\left[\|\nabla_y g(x, y) - \nabla_y g(x, y; \xi)\|^2\right] + \|\nabla_y g(x, y)\|^2 \\ =& \mathbb{E}\left[\|\nabla_y g(x, y; \xi)\|^2\right] \text{ – uniformly bounded,}\end{aligned}$$

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 7. However Assumption 7 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 7 but does not have bounded gradient.)

ASSUMPTION 8 (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.

3.3. 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., 2021a] that gives a closed form expression of the hypergradient $\nabla\Phi(x)$, making gradient-based bilevel optimization tractable.

LEMMA 3.3.0.1. *Suppose Assumption 5 holds. The hypergradient $\nabla\Phi(x)$ of (3.2) takes the form:*

(3.4)
$$\nabla\Phi(x) = \frac{1}{n}\left(\sum_{i=1}^{n}\nabla_x f_i(x, y^*(x))\right) - \nabla^2_{xy} g(x, y^*(x))\left(\nabla^2_y g(x, y^*(x))\right)^{-1}\left[\frac{1}{n}\left(\sum_{i=1}^{n}\nabla_y f_i(x, y^*(x))\right)\right].$$

We also include smoothness properties of $\nabla\Phi(x)$ and $y^*(x)$ in Section A.4 in the appendix.

3.3.1. 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

$$\nabla^2_{xy} g(x, y^*(x))\left(\nabla^2_y g(x, y^*(x))\right)^{-1} \neq \frac{1}{n}\sum_{i=1}^{n}\nabla^2_{xy} g_i(x, y_i^*(x))\left(\nabla^2_y g_i(x, y_i^*(x))\right)^{-1}, \tag{3.5}$$

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

$$z = \left(\sum_{i=1}^{n}\nabla^2_y g_i(x, y^*(x))\right)^{-1}\left(\sum_{i=1}^{n}\nabla_y f_i(x, y^*(x))\right). \tag{3.6}$$

According to (3.4), the global hypergradient is given by

$$\nabla\Phi(x) = \frac{1}{n}\sum_{i=1}^{n}(\nabla_x f_i(x, y^*(x)) - \nabla^2_{xy} g_i(x, y^*(x))z). \tag{3.7}$$

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.

3.3.2. Hessian-Inverse-Gradient-Product oracle. Solving (3.6) 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

$$\sum_{i=1}^{n} H_i z = \sum_{i=1}^{n} b_i, \text{ or } z = \left(\sum_{i=1}^{n} H_i\right)^{-1}\left(\sum_{i=1}^{n} b_i\right). \tag{3.8}$$

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

$$\min_{z\in\mathbb{R}^q} \frac{1}{n}\sum_{i=1}^{n} h_i(z), \text{ where } h_i(z) = \frac{1}{2}z^\top H_i z - b_i^\top z. \tag{3.9}$$

Hence we can design a decentralized algorithm to solve for z without the presence of a central server. Based on this observation and (3.7), we present our Hessian-Inverse-Gradient Product oracle in Algorithm 3.

Algorithm 3: 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, number of total iterations N, $d_{i,0}^{(k)} = -b_{i,0}^{(k)}$, $s_{i,0}^{(k)} = -b_{i,0}^{(k)}$, and $z_{i,0}^{(k)} = 0$.
for $t = 0, 1, ..., N-1$ **do**
  for $i = 1, ..., n$ **do**
    $z_{i,t+1}^{(k)} = \sum_{j=1}^{n} w_{ij} z_{j,t}^{(k)} - \gamma d_{i,t}^{(k)}$,
    $s_{i,t+1}^{(k)} = H_{i,t+1}^{(k)} z_{i,t+1}^{(k)} - b_{i,t+1}^{(k)}$,
    $d_{i,t+1}^{(k)} = \sum_{j=1}^{n} w_{ij} d_{j,t}^{(k)} + s_{i,t+1}^{(k)} - s_{i,t}^{(k)}$,
  end for
end for
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, Shi et al., 2015, Di Lorenzo and Scutari, 2016, 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 3. Instead of using the local stochastic gradient in line 4 of Algorithm 3, we maintain another set of variables $d_{i,t+1}^{(k)}$ in line 6 as the gradient tracking step. We will utilize the linear convergence property of gradient tracking in our convergence analysis.

Algorithm 4: Hypergradient Estimation

1: **Input:** Samples $\phi = (\phi_{i,0}, ..., \phi_{i,N})$, $\xi = (\xi_{i,0}, ..., \xi_{i,N})$ on node i.
2: Run Algorithm 3 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 $z_{i,N}^{(k)}$.
3: Set $u_{i,k} = \nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,0}) - \nabla_{xy}^2 g_i(x_{i,k}, y_{i,k}^{(T)}; \xi_{i,0}) z_{i,N}^{(k)}$.
4: **Output:** $u_{i,k}$ on node i.

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 3 of Algorithm 4, however, the real implementation only requires Hessian-vector products, as shown in Algorithm 3, and we do not need to compute the full Hessian.

3.3.3. 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 5 we adopt the basic structure of double-loop bilevel optimization algorithm [Ghadimi and Wang, 2018, Ji et al., 2021, Chen et al., 2021a] – we first run T-step inner loop (line 4-8) to obtain a good approximation of y^*. Next, we run Algorithm 4 to estimate the hypergradient. To reduce the order of the bias in hypergradient estimation error (see Section 3.3.5.1 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]. Note that all communication steps of our Algorithms (lines 4 and 6 of Algorithm 3, lines 6 and 11 of Algorithm 5) only include sending (resp. receiving) vectors to (resp. from) neighbors, which greatly

Algorithm 5: MA-DSBO Algorithm

Input: Stepsizes α_k, β_k, iteration numbers K, T, N, $y_{i,k}^{(0)} = 0$, and $x_{i,0} = r_{i,0} = 0$.
for $k = 0, 1, ..., K-1$ **do**
  $y_{i,k}^{(0)} = y_{i,k-1}^{(T)}$.
  for $t = 0, 1, ..., T-1$ **do**
    for $i = 1, ..., n$ **do**
      $y_{i,k}^{(t+1)} = \sum_{j=1}^{n} w_{ij} y_{j,k}^{(t)} - \beta_k v_{i,k}^{(t)}$ with $v_{i,k}^{(t)} = \nabla_y g_i(x_{i,k}, y_{i,k}^{(t)}; \tilde{\xi}_{i,k}^{(t)})$
    end for
  end for
  Run Algorithm 4 and set the output as $u_{i,k}$.
  for $i = 1, ..., n$ **do**
    $x_{i,k+1} = \sum_{j=1}^{n} w_{ij} x_{j,k} - \alpha_k r_{i,k}$.
    $r_{i,k+1} = (1 - \alpha_k) r_{i,k} + \alpha_k u_{i,k}$.
  end for
end for
Output: $\bar{x}_K = \frac{1}{n} \sum_{i=1}^{n} x_{i,K}$.

reduce the per-iteration communication complexity from $\max\{pq, q^2\}$ of GBDSBO (see line 8 and 11 of Algorithm 1 in Yang et al. [2022]).) to $\max\{p, q\}$.

We now introduce our notion of convergence. Specifically, the ϵ-stationary point of (3.3) is defined as follows.

DEFINITION 3.3.1. *For a sequence $\{\bar{x}_k\}_{k=0}^{K}$ generated by Algorithm 5, 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 (3.3).

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 5 is given in Theorem 3.3.2.

THEOREM 3.3.2. *Suppose Assumptions 5, 6, 7, and 8 hold. There exist constants* [†] $0 < c_1 < c_2$ *such that in Algorithm 5 if we set $\gamma \in (c_1, c_2)$, $T \ge 1$, and*

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

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

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 5 for obtaining an ϵ-stationary point of (3.2) is $\mathcal{O}(\epsilon^{-2})$. Moreover, we have the following corollary that gives the sample complexity of our algorithm.

COROLLARY 3.3.1. *Suppose the conditions of Theorem 3.3.2 hold. For any $\epsilon>0$, if we set $K=\mathcal{O}\left(\epsilon^{-2}\right)$, $N=\Theta(\log\frac{1}{\epsilon})$, and $T=1$, then in Algorithm 5 the sample complexity to find an ϵ-stationary point is $\mathcal{O}(\epsilon^{-2}\log(\frac{1}{\epsilon}))$.*

It is worth noting that $T\geq 1$ in Theorem 3.3.2 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]. It is interesting to study how setting T dependent on other problem parameters will improve the dependency on problem parameters in the final convergence rate. The hypergradient estimation algorithms (i.e., HIGP oracle and Algorithm 4) provide an additional $\mathcal{O}(\log\frac{1}{\epsilon})$ factor in the sample complexity, which matches Chen et al. [2021a]. To further remove the log factor, Arbel and Mairal [2021] applies warm start to hypergradient estimation and uses mini-batch method (whose batch sizes are dependent on ϵ^{-1}) 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 3.3.2 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 A.5 for a detailed discussion). Besides, according to Table 3.1, our Algorithm is more efficient and preferable when $\min\{p,q\}>n$ since we improve the per-iteration computational and communication complexity from $\max\{pq,q^2\}$ in Yang et al. [2022] to $\max\{p,q\}$. It would be interesting to study how to incorporate Jacobian-computing-free algorithm in DSBO under the mild assumptions without affecting linear speedup.

3.3.4. Consequences for Decentralized Stochastic Compositional Optimization. Note that our algorithm can be used to solve Decentralized Stochastic Compositional Optimization (DSCO) problem:

$$\min_{x\in\mathbb{R}^p} \quad \Phi(x) = \frac{1}{n}\sum_{i=1}^{n} f_i\left(\frac{1}{n}\sum_{j=1}^{n} g_j(x)\right), \tag{3.10}$$

which can be written in a bilevel formulation:

$$\begin{aligned} \min_{x\in\mathbb{R}^p} \quad & \Phi(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(y^*(x)) \\ \text{s.t.} \quad & y^*(x) = \arg\min_{y\in\mathbb{R}^q} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}y^\top y - g_i(x)^\top y\right), \end{aligned} \tag{3.11}$$

To solve DSCO, Zhao and Liu [2022] proposes D-ASCGD and its compressed version. Both of them have $\mathcal{O}(\epsilon^{-2})$ sample complexity. However, their algorithm requires stronger assumptions (see Assumption 1 (a) in Zhao and Liu [2022]) and needs to compute full Jacobians (i.e., $\nabla g_i(x;\xi)$), which lead to $\mathcal{O}(pq\epsilon^{-2})$ computational complexity. By using our Algorithm 5, we can obtain $\tilde{\mathcal{O}}(\max(p,q)\epsilon^{-2})$ computational complexity, which is preferable in high dimensional problems. We state the result formally in the corollary below; the proof is immediate.

COROLLARY 3.3.2. *Suppose the conditions of Theorem 3.3.2 hold. For any $\epsilon > 0$, if we set $K = \mathcal{O}\left(\epsilon^{-2}\right)$, $N = \Theta(\log\frac{1}{\epsilon})$, and $T = 1$, then the sample complexity of using Algorithm 5 to find an ϵ-stationary point of Problem (3.11) is $\mathcal{O}(\epsilon^{-2}\log(\frac{1}{\epsilon}))$, and the computational complexity is $\tilde{\mathcal{O}}(\max(p,q)\epsilon^{-2})$.*

3.3.5. Proof sketch. In this section we briefly introduce a sketch of our proof for Theorem 3.3.2 as well as the ideas of the algorithm design. Throughout our analysis, we define the filtration as

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

3.3.5.1. *Moving average method.* The moving average method used in line 12 of Algorithm 5 serves as a key step in setting up the convergence analysis framework. We focus on estimating

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

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

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

It can then be shown that by appropriately choosing parameters (see Lemma A.4.0.11 and A.4.0.12 for details), we 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.

3.3.5.2. *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.4.0.13) shows that

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

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.4.0.7). The third term represents the inner loop estimation error, which can be bounded by considering its decrease as k increases (Lemma A.4.0.8). Our novelty lies in bounding the last two terms – the consensus and convergence analysis of the HIGP oracle. Observe that by setting

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

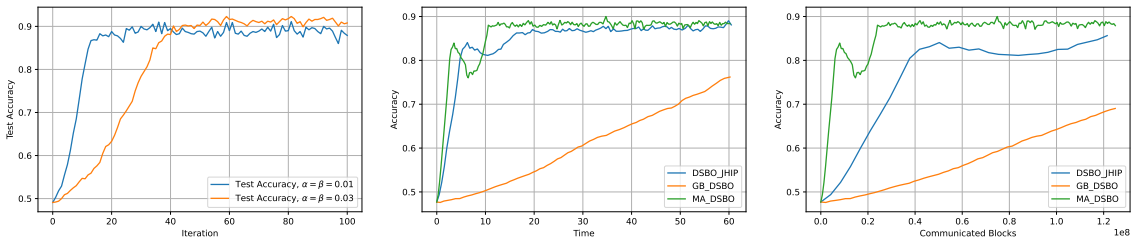
we know from Algorithm 3

$$\dot{z}_{i,t+1}^{(k)} = \sum_{j=1}^{n} w_{ij}\dot{z}_{j,t}^{(k)} - \gamma\dot{d}_{i,t}^{(k)},\ Z_0^{(k)} = 0,$$

$$\dot{d}_{i,t+1}^{(k)} = \sum_{i=1}^{n} w_{ij} \dot{d}_{j,t}^{(k)} + \dot{s}_{i,t+1}^{(k)} - \dot{s}_{i,t}^{(k)},$$

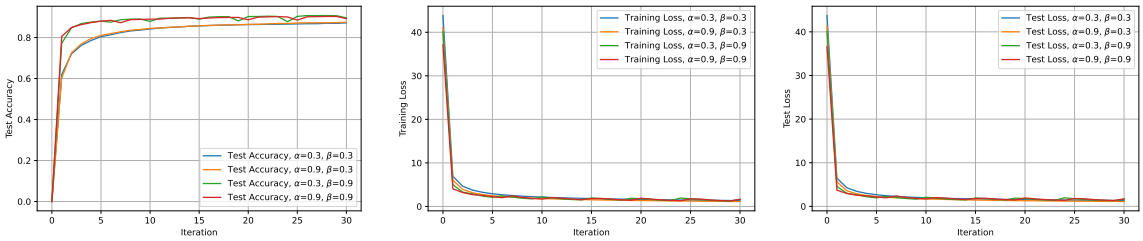
$$\dot{s}_{i,t}^{(k)} = \nabla_y^2 g_i(x_{i,k}, y_{i,k}^{(T)}) \dot{z}_{i,t}^{(k)} - \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}),$$

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.4.0.9 and A.4.0.13 we explicitly characterize the error and eventually obtain the final convergence result in Theorem 3.3.2.



(a) (b) (c)

FIGURE 3.1. ℓ^2-regularized logistic regression on synthetic data.



(a) (b) (c)

FIGURE 3.2. ℓ^2-regularized logistic regression on MNIST.

3.4. Numerical experiments

In this section we study the applications of Algorithm 5 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 3) 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.

3.4.1. 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. The data distribution of x_e on node i is $\mathcal{N}(0, i^2)$. 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 $\lambda \in \mathbb{R}^p$. We also compare our Algorithm 5 with GBDSBO [Yang et al., 2022] and DSBO-JHIP [Chen et al., 2022b] under this setting with dimension parameter $p = 100$. Figures 3.1(a), 3.1(b) and 3.1(c)[‡] demonstrate the efficiency of our algorithm in both time and space complexity. Due to space limit, we include our additional experiments in Section A.3.

[‡]The word "block" is a term used in tracemalloc module in Python (see https://docs.python.org/3/library/tracemalloc.html) to measure the memory usage, and we keep track of the number of the communicated blocks between different agents as a direct measure for communication cost.

3.4.2. 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 3.2. Note that in previous algorithms [Chen et al., 2022b, Yang et al., 2022] one Hessian matrix of the lower level function requires $\mathcal{O}(c^2p^2)$ 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 MA-DSBO Algorithm 5 has a considerably good performance on real world dataset. Note that this problem has larger dimension, and the other algorithms took more time so we do not do the comparison.

3.5. 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 7 (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]. It is also interesting to incorporate Hessian-free methods [Sow et al., 2022c] in DSBO, and we leave it as future work.

CHAPTER 4

Training Dynamics of Gradient Descent for Quadratic Regression

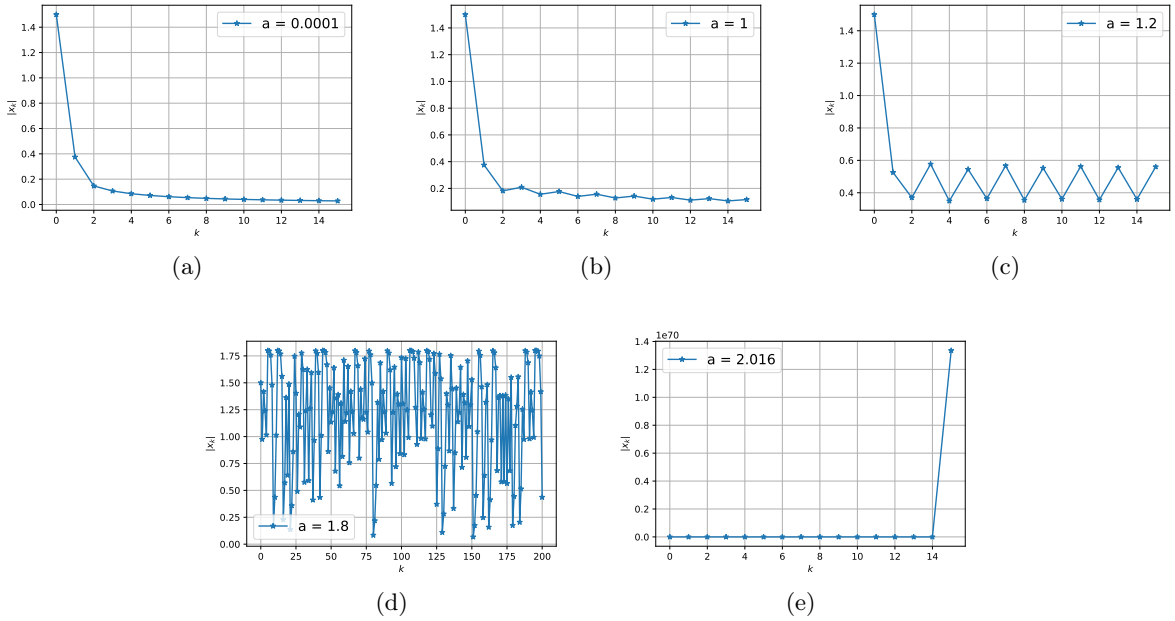
4.1. Introduction

Iterative algorithms like the gradient descent and its stochastic variants are widely used to train deep neural networks. For a given step-size (or learning rate) parameter $\eta > 0$, the gradient descent algorithm is of the form $w^{(t+1)} = w^{(t)} - \eta \nabla \ell(w^{(t)})$ where ℓ is the training objective function being minimized, which depends on the loss function and the neural network architecture and the dataset. Classical optimization theory operates under small-order step-sizes. In this regime, one can think of the gradient descent algorithm as a discretization of so-called gradient flow equation given by $\dot{w}^{(t)} = -\nabla \ell(w^{(t)})$, which could be obtained from the gradient descent algorithm by letting $\eta \to 0$. Additionally, assuming that the objective function ℓ has gradients that are L-Lipschitz, selecting a step-size $\eta < 1/L$ guarantees convergence to stationarity.

In stark contrast to traditional optimization, recent empirical studies in deep learning have revealed that training deep neural networks with large-order step-sizes yields superior generalization performance. Unlike the scenario with small step-sizes, where gradient descent dynamics follow a monotonic pattern, larger step-sizes introduce a more intricate behavior. Various patterns like catapult, (also related to *edge of stability*), periodicity and chaotic dynamics in neural network training with large step-sizes have been observed empirically, for example, by Lewkowycz et al. [2020], Jastrzebski et al. [2020], Cohen et al. [2021], Lobacheva et al. [2021], Gilmer et al. [2022], Zhang et al. [2022a], Kodryan et al. [2022], Herrmann et al. [2022]. A recent work by Sohl-Dickstein [2024] also empirically observe that the boundary between stable and divergent training behaviour, in terms of hyperparameters (including the step-size parameter), exhibits a fractal structure. Furthermore, the necessity for step-size schedules to include large-order step-sizes to expedite convergence and the ensuing chaotic behavior has also been observed empirically outside the deep learning community by Van Den Doel and Ascher [2012], much earlier.

Faster convergence of gradient descent with iteration-dependent step-size schedules that have specific patterns (including cyclic and fractal patterns) has been examined empirically by Lebedev and Finogenov [1971], Smith [2017], Oymak [2021], Agarwal et al. [2021], Goujaud et al. [2022], and Grimmer [2023], with Altschuler and Parrilo [2023] and Grimmer et al. [2023] proving the state-of-the art remarkable results; see also Altschuler and Parrilo [2023, Section 1.2] for a historical overview. Notably, the stated faster convergence behavior of gradient descent requires large order step-sizes, very much violating the classical case. More importantly, the corresponding optimization trajectory, while being non-monotonic, exhibits intriguing patterns [Van Den Doel and Ascher, 2012].

Considering the aforementioned factors, gaining insight into the dynamics of gradient descent with large-order step-sizes emerges as a pivotal endeavor. A precise theoretical characterizing of the gradient descent dynamics in the large step-size regime for deep neural network, and other such non-convex models, is a formidably challenging problem. Existing findings (as detailed in Section 4.1.1) often rely on strong assumptions, even when attempting to delineate a subset of the aforementioned patterns, and do not provide a comprehensive account of the entire narrative underlying the training dynamics. Recent research, such as Agarwala et al. [2023], Zhu et al. [2024], and Zhu et al. [2023b], has pivoted towards comprehending the dynamics of quadratic regression models based on a *local* analysis. These models offer a valuable testing ground due to their ability to provide tractable approximations for various machine learning models, including phase retrieval, matrix factorization, and two-layer neural networks, all of which exhibit unstable training dynamics. Despite their seeming simplicity, a fine-grained understanding of their training dynamics is far from trivial. Building in this direction, the primary aim of our work is to attain a precise characterization of the training dynamics of gradient descent in quadratic models, thereby fostering a deeper comprehension of the diverse phases involved in the training process.



(a) (b) (c)

(d) (e)

FIGURE 4.1. Phases of cubic-map based dynamical system in (4.2) parameterized by a. Sub-figure 4.1(a) corresponds to the monotonic phases, where the dynamics monotonically decays to zero. Sub-figure 4.1(b) corresponds to the generalized catapult phase where the dynamics decays to zero but is non-monotonic in a specific manner. Sub-figure 4.1(c) corresponds to the periodic phase, where the dynamics decays and settles in a period-2 orbit (i.e., shuttles between two points) but never decays to zero. Sub-figures 4.1(d) and 4.1(e) correspond to the chaotic phase (see Definition 4.2.1) and divergent phases, respectively. Note that the order of x-axis and y-axis in Sub-figures 4.1(d) and 4.1(e) are different from the rest.

Contribution 1. We perform a *fine-grained, global theoretical analysis* of a cubic-map-based dynamical system (see Equation 4.2), and identify the precise boundaries of the following five phases: (i) monotonic, (ii) generalized catapult, (iii) periodic, (iv) Li-Yorke chaotic, and (v) divergent. See Figure 4.1 for an illustration, and Definition 4.2.2 and Theorem 4.2.3 for formal results. We show in Theorem 4.3.2 and 4.3.3, that the dynamics of gradient descent for two non-convex statistical problems, namely phase retrieval and two-layer neural networks with constant outer layers and quadratic activation functions, with orthogonal training data is captured by the cubic-map-based dynamical system. We provide empirical evidence of the presence of similar phases in training with non-orthogonal data.

We also empirically examine the effect of training models in the above-mentioned phases, in particular the non-monotonic ones, on the generalization error. Indeed, provable model-specific statistical benefits for training in catapult phase are studied in Lyu et al. [2022] and Ahn et al. [2022]. Lim et al. [2022] proposed to induce controlled chaos in the training trajectory to obtain better generalization. Approaches to explain generalization with chaotic behavior are examined in Chandramoorthy et al. [2022] based on a relaxed notion of statistical algorithmic stability. Although our focus is on gradient descent, related notions of generalization of stochastic gradient algorithms, based on characterizing the fractal-like properties of the invariant measure they converge to (with larger-order constant step-size choices) have been explored, for example, in Birdal et al. [2021], Camuto et al. [2021], Dupuis et al. [2023], and Hodgkinson et al. [2022]. Hence, we also conduct empirical investigations into the performance of generalization when training within the different non-monotonic (and non-divergent) phases and make the following contribution.

Contribution 2. We propose a natural ergodic trajectory averaging based prediction mechanism (see Section 4.4.2) to stabilize the predictions when operating in any non-monotonic (and non-divergent) phase.

4.1.1. Related works. General results. Lewkowycz et al. [2020] empirically examine the catapult phase particularly in neural networks with one hidden layer and linear activations, the phase in which the linear approximation of the model becomes less informative. In this case, they observe that the loss does not have monotonic decrease but eventually converges when the curvature (maximum of the eigenvalue of the Neural Tangent Kernel [Jacot et al., 2018]) stabilizes at a value less than 2/(step-size). Similar oscillations with convergence behavior have been also observed in Cohen et al. [2021], which empirically demonstrate that the sharpness (largest eigenvalue of the Hessian matrix of the loss) in gradient descent on neural networks training hovers just above the value 2/(step-size), indicating that gradient descent usually operates in the regime they call Edge of Stability (EoS). This is also formally studied in Ahn et al. [2022]. Damian et al. [2023] propose self-stabilization as a phenomenological reason for the occurrence of catapults and EoS in gradient descent dynamics. Kreisler et al. [2023] investigate how gradient descent monotonically decreases the sharpness of Gradient Flow solutions, specifically in one-dimensional deep neural networks. Although they do not formally prove the existence of chaos in the dynamics, they conjecture its

possibility. Arora et al. [2022] and Lyu et al. [2022] explore sharpness reduction flows, related to the above findings. Andriushchenko et al. [2023] prove that large step-sizes in gradient descent can lead to the learning of sparse features. Wu et al. [2023] investigate the EoS phenomenon for logistic regression. Kong and Tao [2020] theoretically explore the chaotic dynamics (and related stochasticity) in gradient descent for minimizing multi-scale functions under additional assumptions. While being extremely insightful, their results are fairly qualitative and are not directly applicable to the cubic maps analyzed in our work. As we focus on specific models, our results are more precise and quantitative.

Specific Models. Zhu et al. [2023b] and Chen and Bruna [2023] studied gradient descent dynamics for minimizing the functions $\ell(u, v) = (u^2v^2 - 1)^2$ and $\ell(u) = (u^2 - 1)^2$, respectively. Both works primarily focused on characterizing period-2 orbits and hint at the possibility of chaos without rigorous theoretical justifications. Furthermore, their proofs are relatively ad-hoc and significantly different from ours. Song and Yun [2024] provided empirical evidence of periodicity and chaos for training a fully-connected neural network using gradient descent. However, their theoretical results are not applicable to quadratic regression models. Ahn et al. [2024] examined the Edge of Stability (EoS) between the monotonic and catapult phase for minimizing $\ell(u, v) = l(uv)$, where l is convex, even, and Lipschitz. Their analysis is not directly extendable to the quadratic regression models we consider in this work. See also the discussion below Theorem 4.2.3 for important technical comparisons. Wang et al. [2022] analyzed additional benefits (e.g., taming homogeneity) of gradient descent with large step-sizes for matrix factorization. Ziyin et al. [2022] also studied *stochastic* gradient descent with large step-sizes for the case when the loss function $\ell(u) = au^2$ for $a \in \mathbb{R}$. Note in this case that the point 0 is the minimum when $a > 0$. However, when $a < 0$, the point 0 is a maximum. In this setup, Ziyin et al. [2022] precisely characterize the behaviour of SGD for converging to a minimum or a maximum, in terms of the step-size parameter, initialization and the noise distribution of the stochastic gradient.

Agarwala et al. [2023] explored gradient descent dynamics for a class of quadratic regression models and identified the EoS. Zhu et al. [2023a,b] also studied the catapult phase and EoS for a class of quadratic regression models. Agarwala and Dauphin [2023] examined the EoS in the context of Sharpness Aware Minimization for quadratic regression models. The above works are related to our work in terms of the model that they study. However, none of the above works characterize the

five distinct phases (with precise boundaries) like we do, along with precise boundaries. Furthermore, our analysis is distinct (and is also global*) from the above works and is firmly grounded in the rich literature on dynamical systems.

Dynamical systems. Our results draw upon the rich literature available in the field of dynamical systems. We refer the interested reader to Alligood et al. [1997], Lasota and Mackey [1998], Devaney [1989], Ott [2002], and De Melo and Van Strien [2012] for a book-level introduction. Birfurcation analysis of some classes of cubic maps has been studied, for example, by Skjolding et al. [1983], Rogers and Whitley [1983], Branner and Hubbard [1988] and Milnor [1992]. Some of the above works are rather empirical, and the exact maps considered in the above works differ significantly from our case.

4.2. Analyzing a discrete dynamical system with cubic map

Notations and definitions. We say a sequence $\{x_k\}_{k=0}^{\infty}$ is increasing (decreasing), if $x_{t+1} \geq x_t$ ($x_{t+1} \leq x_t$) for any t. Moreover, it is strictly increasing (decreasing) if the equalities never hold. For a real-valued function f and a set S, define $f(S) = \{f(x) : x \in S\}$, and $f^{(k)}(x) := f(f^{(k-1)}(x))$ for any $k \in \mathbb{N}_+$ with $f^{(0)}(x) = x$. The preimage of x under f on S is the set $f^{-1}(x) := \{y \in S : f(y) = x\}$. We say a property P holds for almost every $x \in S$ or almost surely in S, if the subset $\{x \in S : \text{property } P \text{ does not hold for } x\}$ is Lebesgue measure zero. A critical point of f is a point x satisfying $f'(x) = 0$. We call x_0 a period-k point of f, when $f^{(k)}(x_0) = x_0$ and $f^{(i)}(x_0) \neq x_0$ for any $0 \leq i \leq k-1$. The orbit of a point x_0 denotes the sequence $\{f^{(t)}(x_0)\}_{t=0}^{\infty}$. A point x_0 is called asymptotically periodic if there exists a periodic point y_0 such that $\lim_{t\to\infty} |f^{(t)}(x_0) - f^{(t)}(y_0)| = 0$. The stable set of a period-k point x_0 is defined as $W^s(x_0) := \left\{x : \lim_{n\to\infty} f^{(kn)}(x) = x_0.\right\}$.

The stable set of the orbit of a periodic point x_0 is the union of the stable sets of all points in the orbit of x_0. A point x_0 is an aperiodic point if it is not an asymptotically periodic point and the orbit of x_0 is bounded. We say a fixed point x_0 of f is stable if, for any $\epsilon > 0$, there is a $\delta > 0$ such that for any x satisfying $|x - x_0| < \delta$, we have $|f^{(n)}(x) - x_0| < \epsilon$ for all $n \geq 0$. The fixed point x_0 is said to be unstable if it is not stable. The fixed point x_0 is asymptotically stable if it is stable and there is an $\delta > 0$ such that $\lim_{n\to\infty} f^{(n)}(x) = x_0$ for all x satisfying $|x - x_0| < \delta$. A period-p point x_0 and its associated periodic orbit are asymptotically stable if x_0 is an asymptotically stable fixed

*Analysis in Wang et al. [2022] and Chen and Bruna [2023] is also global, but not applicable to our model.

point of $f^{(p)}$. A point $x_0 \in \mathbb{R} \bigcup \{+\infty, -\infty\} \backslash S$ is called an absorbing boundary point of S for f with period p, for some $p \in \{1, 2\}$, if there exists an open set $U \subseteq S$ such that $\lim_{k\to\infty} f^{(pk)}(y) \to x$ for all $y \in U$.

We now introduce two quantities that are common in dynamical systems theory to study the stability properties. The Schwarzian derivative of a three-times continuously differentiable function f is defined (at non-critical points) as

$$\mathsf{S}f(x) \coloneqq (f'''(x)/f'(x)) - 1.5\left(f''(x)/f'(x)\right)^2, \text{ where } f'(x) \neq 0. \tag{4.1}$$

It is widely used for its sign-preservation property under compositions; see, for example, De Melo and Van Strien [2012]. Specifically, the stability of a fixed point is related to the sign of the Schwarzian derivative at that point. Positive values may indicate instability, while negative values suggest stability. The Lyapunov exponent of a given orbit with initialization x_0 is defined as

$$\mathsf{L}f(x_0) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n-1} \log |f'(x_i)|.$$

It is another related quantity associated with the stability properties of dynamical systems and is used to measure the sensitive dependence on initial conditions [Strogatz, 2018]. Chaotic systems typically exhibit positive Lyapunov exponents, reflecting their sensitive dependence on initial conditions. Similarly, a negative Lyapunov exponent is a characteristic of stable systems. Finally, we also define the sharpness of a loss function is defined as the maximum eigenvalue of the Hessian matrix of the loss.

Bifurcation analysis. Our main goal in this section is to undertake a bifurcation analysis of the following discrete dynamics system defined by a cubic map. For $a > 0$, first define the functions g and f, parameterized by a, as

$$g_a(z) = z^2 + (a-2)z + 1 - 2a = (z+a)(z-2) + 1 \quad \text{and} \quad f_a(z) = z g_a(z). \tag{4.2}$$

Next, consider the discrete dynamical system given by

$$z_{t+1} = f_a(z_t) = z_t g_a(z_t). \tag{4.3}$$

Note that for any $a, \epsilon > 0$ and $z_0 \geq 2 + \epsilon$ or $z_0 \leq -a - \epsilon$, we will have $\lim_{t\to\infty} |z_t| = +\infty$. Hence, we only study the case when $z_0 \in [-a, 2]$. We will show in Section 4.3 that the dynamics of the training loss for several quadratic regression models could be captured by (4.3). The parameter a in (4.2) for the models will naturally correspond to the step-size of the gradient descent algorithm.

We next introduce the precise definitions of the five phase that arise in the bifurcation analysis of (4.2). To do so, we need the following definition of chaos in the Li-Yorke sense [Li and Yorke, 1975]. Li-Yorke chaos is widely used in the study of dynamical systems and is also directly related to important measures of the complexity of dynamical systems, like the topological entropy [Adler et al., 1965, Franzová and Smítal, 1991]. We also refer to Aulbach and Kieninger [2001] and Kolyada [2004] for its relationship to other notions of chaos and related history.

DEFINITION 4.2.1 (Li-Yorke Chaos [Li and Yorke, 1975]). *Suppose we are given a function $f(x)$. If there exists a compact interval I such that $f : I \to I$, then it is called Li-Yorke chaotic [Li and Yorke, 1975, Aulbach and Kieninger, 2001] when it satisfies*

- *For every $k = 1, 2, ...$ there is a periodic point in I having period-k.*
- *There is an uncountable set $S \subseteq I$ (containing no periodic points), which satisfies for any $p, q \in S$ with $p \neq q$, $\limsup_{t\to\infty} |f^{(t)}(p) - f^{(t)}(q)| > 0$, $\liminf_{t\to\infty} |f^{(t)}(p) - f^{(t)}(q)| = 0$, and for any $p \in S$ and periodic point $q \in I$, $\limsup_{n\to\infty} |f^{(t)}(p) - f^{(t)}(q)| > 0$.*

To define the 5 phases in particular, we consider the orbit $\{f^{(k)}(x)\}_{k=0}^{+\infty}$ generated by a given function f defined over a set I, in which the initial point x belongs to.

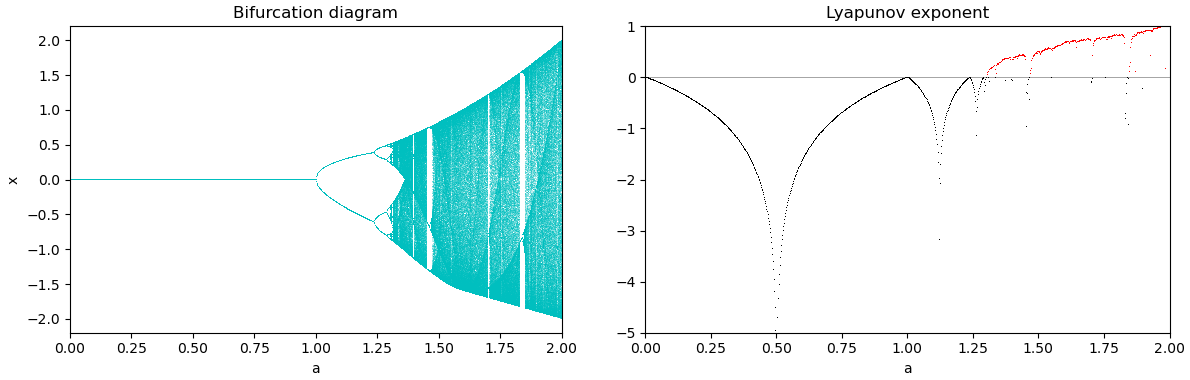
DEFINITION 4.2.2. *Given a function $f(x)$ defined on a set I, we say the discrete dynamics is in the*

- ***Monotonic phase**, when $\{|f^{(k)}(x)|\}_{k=0}^{\infty}$ is decreasing and $\lim_{n\to\infty} |f^{(n)}(x)| = 0$ for almost every $x \in I$.*
- ***Generalized[†] catapult phase**, when $\{|f^{(k)}(x)|\}_{k=m}^{\infty}$ is not decreasing for any m, and for almost every $x \in I$ $\lim_{n\to\infty} |f^{(n)}(x)| = 0$. We say such sequences have catapults.*
- ***Periodic phase**, when f is not Li-Yorke chaotic, $\{|f^{(k)}(x)|\}_{k=0}^{\infty}$ is bounded and does not have a limit for almost every $x \in I$, and there exists period-2 points in I.*

[†]Here, we use the term *generalized* to distinguish from Lewkowycz et al. [2020] who consider the case of a single spike in the training loss.

- ***Chaotic phase***, *when the function f is Li-Yorke chaotic and $\{|f^{(k)}(x)|\}_{k=0}^{\infty}$ is bounded for almost every $x \in I$.*
- ***Divergent phase***[‡], *when $\lim_{n\to\infty} |f^{(n)}(x)| = +\infty$ for almost every $x \in I$.*

We emphasize here that our use of the word "phase" refers to the whole sequence $\{|f^{(k)}(x)|\}_{k=0}^{\infty}$, and the categorization is with respect to the different step-sizes. As an illustration, in Figure 4.1, we plot the five phases for the parameterized function and its discrete dynamical system defined in (4.2) with initialization 1.9, i.e., $x_k = f_a^{(k)}(x_0),\ x_0 = 1.9$. We have the following main result for different phases of dynamics.



(a) (b)

FIGURE 4.2. Bifurcation diagram and Lyapunov exponent. Initialization $z_0 = 0.1$.

THEOREM 4.2.3. *Suppose $f_a(z)$ is defined in (4.2). Define $z_{t+1} = f_a(z_t)$ with z_0 sampled uniformly at random in $(-a, 2)$. Then there exists $a_* \in (1, 2)$ such that the following holds.*

- *If $a \in (0, 2\sqrt{2} - 2]$, then almost surely $\lim_{t\to\infty} |z_t| = 0$ and $|z_t|$ is decreasing, and hence the dynamics is in the monotonic phase.*
- *If $a \in (2\sqrt{2} - 2, 1]$, then almost surely $\lim_{t\to\infty} |z_t| = 0$ and $|z_t|$ have catapults, and hence the dynamics is in the generalized catapult phase.*
- *If $a \in (1, a_*)$, then there exists a period-2 point in $(0, 1)$. $z_t \in (-a, 2)$ for all t. If there exists an asymptotically stable periodic orbit, then the orbit of z_0 is asymptotically periodic almost surely, and hence the dynamics is in the periodic phase.*

[‡]We do not further sub-characterize the divergent phase as it is uninteresting.

- *If $a \in (a_*, 2]$, f_a is Li-Yorke chaotic. $z_t \in (-a, 2)$ for all t. If there exists an asymptotically stable periodic orbit, then the orbit of z_0 is asymptotically periodic almost surely, and hence the dynamics is in the chaotic phase.*
- *If $a \in (2, +\infty)$, then $\lim_{t\to\infty} |z_t| = +\infty$ almost surely, and hence the dynamics is in the divergent phase.*

From a pure optimization perspective, Phase 1 and 2 are the most relevant, as training loss actually minimized. However, from a generalization perspective, similar to other works [Lyu et al., 2022, Lim et al., 2022, Chandramoorthy et al., 2022] we empirically observe that often times phases 2, 3 and 4 lead to comparatively improved generalization for various models.

Connections with sharpness and EoS. As we will see in Section 4.3, the training loss and sharpness of a special class of quadratic regression models can be written as functions of z_t, and hence their dynamics can be explicitly given by Theorem 4.2.3. As a byproduct of our theory, we reveal that the EoS phenomenon happens in the catapult phase and quantify the limit that the sharpness eventually converges to, which matches the empirical observations in Cohen et al. [2021] and Ahn et al. [2022].

As a direct application of Theorem 4.2.3, we have the following result characterizing the dynamics generated by n different functions.

COROLLARY 4.2.1. *Suppose $f_a(z)$ is defined in (4.2), and we are given $2n$ positive scalars a_i, ρ_i for $1 \le i \le n$. Define $z_i^{(t+1)} = f_{a_i}(z_i^{(t)})$, $L(z^{(t)}, \rho) = \sum_{i=1}^n \rho_i (z_i^{(t)})^2$. Then for almost all $z^{(0)} \in \{z : -a_i \le z_i \le 2\}$ we have*

- *If $0 < \max_{1\le i\le n} a_i \le 1$, then $\lim_{t\to\infty} L(z^{(t)}, \rho) = 0$. Moreover, if $0 < \max_{1\le i\le n} a_i \le 2\sqrt{2} - 2$, the sequence $\{L(z^{(t)}, \rho)\}_{t=0}^{\infty}$ is decreasing.*
- *If $1 < \max_{1\le i\le n} a_i \le 2$, then $\{L(z^{(t)}, \rho)\}_{t=0}^{\infty}$ is bounded and does not converge to 0.*
- *If $\max_{1\le i\le n} a_i > 2$, then $\lim_{t\to\infty} L(z^{(t)}, \rho) = +\infty$.*

We highlight here that even if we know from Theorem 4.2.3 the dynamics of each individual $z_i^{(t)}$, explicitly characterizing the phase of $L(z^{(t)}, \rho)$ is not trivial. To see this, we provide one simple example as follows.

$$S_1 := \{S_1^{(n)}\} = \{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, ...\},\ S_2 := \{S_2^{(n)}\} = \{1, \frac{1}{2^2}, \frac{1}{3^2}, \frac{1}{4^2}, ...\},\ S_3 := \{S_3^{(n)}\} = \{\frac{1}{2}, 1, \frac{1}{4}, \frac{1}{3}, ...\},$$

where S_3 is obtained by switching the $(2i-1)$-th and $2i$-th terms in S_1. Sequences S_1 and S_2 are decreasing to 0, and S_3 is in the catapult phase. We can verify that both $\{S_1^{(n)} + S_3^{(n)}\}$ and $\{S_2^{(n)} + S_3^{(n)}\}$ are converging to 0 but the former is decreasing while the latter is in the generalized catapult phase. This implies that the summation of a decreasing sequence and a catapult sequence can be either decreasing or catapult, which makes analyzing the dynamics of the weighted summation $L(z^{(t)}, \rho)$ non-obvious. As we will see in Section 4.3.1, the above result gives the training dynamics of generalized phase retrieval and a two-layer neural network with quadratic activation functions on n orthogonal data points.

In Figures 4.2(a) and 4.2(b) we numerically plot a bifurcation diagram for $a \in (0, 2)$ and Lyapunov exponent scatter plot with initialization $z_0 = 0.1$. The main ingredients in proving Theorem 4.2.3 are the following Lemmas 4.2.3.1, 4.2.3.2, and 4.2.3.3. Note that by straightforward computations, we have

$$f_a'(0) = 1 - 2a \in (-1, 1) \Leftrightarrow a \in (0, 1). \tag{4.4}$$

This implies 0 is a asymptotically stable fixed point when $a \in (0, 1)$. This type of local stability analysis is standard in dynamical systems literature [Hale and Koçak, 2012, Strogatz, 2018], and has been used in analyzing the training dynamics of gradient descent recently [Zhu et al., 2024, Song and Yun, 2024]. However, such results are limited to only local regions. In contrast, the following results provide a global convergence analysis.

LEMMA 4.2.3.1. *Suppose $0 < a \leq 1$ and $-a \leq z_0 \leq 2$. Then we have*

- *(i) $-a \leq z_t \leq 2$ for any t, and f_a does not have a period-2 point on $[-a, 2]$.*
- *(ii) If z_0 is chosen from $[-a, 2]$ uniformly at random, then $\lim_{t\to\infty} z_t = 0$ almost surely. Moreover, if $0 < a \leq 2\sqrt{2} - 2$, then almost surely $|z_{t+1}| \leq |z_t|$ for all t. If $2\sqrt{2} - 2 < a \leq 2$, then almost surely $\{|z_t|\}_{t=0}^{\infty}$ has catapults.*

LEMMA 4.2.3.2. *Suppose $1 < a \leq 2$ and $-a \leq z_0 \leq 2$. Then we have*

- *(i) $-a \leq z_t \leq 2$ for any t, and $f_a(z)$ has a period-2 point on $[0, 1]$.*
- *(ii) There exists $a_* \in (1, 2)$ such that for any $a \in (a_*, 2)$, f_a is Li-Yorke chaotic, and for any $a \in (1, a_*)$, f_a is not Li-Yorke chaotic.*

- *(iii) If there exists an asymptotically stable orbit and z_0 is chosen from $[-a, 2]$ uniformly at random, then the orbit of z_0 is asymptotically periodic almost surely.*

LEMMA 4.2.3.3. *Suppose $a > 2$. z_0 is chosen from $[-a, 2]$ uniformly at random. Then $\lim_{t\to\infty} |z_t| = +\infty$ almost surely.*

In Lemma 4.2.3.2, part (iii), we assume the existence of an asymptotically stable periodic point. Note that such a point must have negative Lyapunov exponent [Strogatz, 2018]. It is possible to obtain particular values for a under which $f_a(z)$ has an asymptotically stable orbit. For example, a can be chosen such that $|f_a'(p)f_a'(q)| < 1$, where $p \in (0, 1)$ is a period-2 point with $f_a(p) = q$. In Figure 4.2(b) we plot the Lyapunov exponent of f_a at the orbit starting from $z_0 = 0.1$. It is interesting to explicitly characterize the set of a values in $(1, 2)$ such that $f_a(z)$ has an asymptotically stable periodic orbit. Furthermore, we conjecture that a_* defined in Lemma 4.2.3.2 is the smallest number $a \in (1, 2)$ such that $(1 - 2a)/3$ is a period-3 point. The above two problems are challenging and left as future work.

4.3. Applications to quadratic regression models

We now provide illustrative examples based on quadratic or second-order regression models, motivated by the works of Zhu et al. [2024] and Agarwala et al. [2023]. Specifically, we consider a generalized phase retrieval model and training hidden-layers of 2-layer neural networks with quadratic activation function as examples.

4.3.1. Example 1: Generalized phase retrieval. Single Data Point. Following Zhu et al. [2024], it is instructive to study the dynamics with a single training sample. Consider the following optimization problem on a single data point (X, y):

$$\min_w \left\{\ell(w) = \frac{1}{2}(g(w; X) - y)^2\right\}, \quad \text{where } g(w; X) = \frac{\gamma(X^\top w)^2}{2} + cX^\top w, \tag{4.5}$$

where γ, c are arbitrary constants. The above model, with $\gamma = 2$ and, $c = 0$ corresponds to the classical phase retrieval model (also called as a single-index model with quadratic link function). We refer to Jaganathan et al. [2016] and Fannjiang and Strohmer [2020] for an overview, importance and applications of the phase retrieval model. We would like to point out that the analysis of seemingly simple models is already non-trivial and has been done in various ways. For example,

single-data-point setting [Zhu et al., 2024, Song and Yun, 2024], simple-model setting [Lobacheva et al., 2021, Ahn et al., 2024, Kodryan et al., 2022, Zhu et al., 2023b, Chen and Bruna, 2023, Zhu et al., 2023a], etc. Different from existing works that mostly focus on asymptotic or local analysis that only hold when certain quantities are sufficiently large or small (small step-sizes [Lobacheva et al., 2021, Ahn et al., 2024, Zhu et al., 2023b], large network size [Zhu et al., 2024, 2023a]), in the following result we provide a refined global analysis on solving (4.5) that does not contain any big-O notation.

THEOREM 4.3.1. *Suppose we run gradient descent on (4.5) with step-size to be η. Define*

$$e^{(t)} := g(w^{(t)}; X) - y,\ z_t := \eta\gamma \left\|X\right\|^2 e^{(t)},\ a = \Big(\gamma y + \frac{c^2}{2}\Big)\eta \left\|X\right\|^2. \tag{4.6}$$

Then we have (i) $z_{t+1} = f_a(z_t)$ and thus Theorem 4.2.3 holds for f_a and z_t; (ii) The sharpness is given by $\lambda_{max}(\nabla^2 \ell(w^{(t)})) = \frac{3z_t+2a}{\eta}$.

Comparison with existing results. An interesting conclusion from the above theorem is that, under certain cases the step-size η should depend on the model initialization. For example when $e^{(0)} > 0$ then we should have $\eta\gamma \left\|X\right\|^2 e^{(0)} = z_0 < 2$, since for $z_0 > 2$ we have $\lim_{t\to} |z_t| = \infty$ (see, e.g., discussions under (4.3)). Note that Zhu et al. [2024] studied a related neural quadratic model (see their Eq. (3)). Here, we highlight that their results do not cover our case. Indeed, defining $\eta_{\text{crit}} = 2/\lambda_{\max}(\nabla^2 \ell(w^{(0)}))$, according to their claim, catapults happen when $\eta_{\text{crit}} < \eta < 2\eta_{\text{crit}}$. In our notation, this condition is equivalent to $2 < 3z_0 + 2a < 4$. However this cannot happen because if the initialization z_0 is sufficiently small, say $z_0 = \mathcal{O}(\epsilon)$, then we know the previous condition become $1 - \mathcal{O}(\epsilon) < a < 2 - \mathcal{O}(\epsilon)$. However, according to Lemmas 4.2.3.1 and 4.2.3.2, we have that for $1 < a < 2$ the training dynamics is in the periodic or the chaotic phase and z_t (and thus the loss function) will not converge to 0. Our theory (Lemma 4.2.3.1) suggests that catapults for quadratic regression model happens for almost every $z_0 \in (-a, 2)$ provided that $2\sqrt{2} - 2 < a \leq 1$. This intricate observation reveals that extending the current results on the catapult phenomenon from the model in Zhu et al. [2024] to our setting is not immediate and is actually highly non-trivial.

Relationship with Sharpness and EoS. We also notice that, interestingly, in the monotonic and catapult phases (i.e., $0 < a \leq 1$), we have the limiting sharpness satisfy $\lim_{t\to\infty} \lambda_{\max}(\nabla^2 \ell(w^{(t)})) = 2a/\eta = (2\gamma y + c^2) \left\|X\right\|^2$. In particular, for the catapult phase ($2\sqrt{2} - 2 < a \leq 1$) the sharpness

converges to $\frac{2a}{\eta} \in (\frac{4\sqrt{2}-4}{\eta}, \frac{2}{\eta}]$, which theoretically and quantitatively explains the empirical observations of EoS in Cohen et al. [2021]. More importantly, the notion of EoS only provides a coarse characterization of the oscillations of the limiting sharpness at the interface of the monotonic and catapult phase. For the quadratic models that we study, the limiting sharpness exhibits a more nuanced behaviour as identify in Theorem 4.3.1, while also recovering and extending existing results on EoS.

Multiple Orthogonal Data Points. We now consider gradient descent on quadratic regression on multiple data points that are mutually orthogonal. Suppose we are given a dataset $\{(X_i, y_i)\}_{i=1}^n$ with $\mathbf{X} = (X_1, ..., X_n)^\top$ satisfying $\mathbf{X}\mathbf{X}^\top = \mathrm{diag}(\|X_1\|^2, ..., \|X_n\|^2)$. Similar orthogonality conditions are widely used in the literature on sparse linear regression to understand the optimization or statistical properties [Tibshirani, 1996, Yuan and Lin, 2006]. Consider the optimization problem

$$\min_w \ell(w) := \frac{1}{n}\sum_{i=1}^n \ell_i(w) = \frac{1}{2n}\sum_{i=1}^n \left(g(w; X_i) - y_i\right)^2, \tag{4.7}$$

where $\ell_i(w)$ and $g(w; X_i)$ are as defined in (4.5).

THEOREM 4.3.2. *Define the following:*

$$\alpha^{(t)}(X_i) := c(X_i) + \gamma X_i^\top w^{(t)},\ \beta(X_i) := y_i + \frac{(c(X_i))^2}{2\gamma},\ \kappa_n(X_i) := \frac{\eta\gamma\|X_i\|^2}{n}, \tag{4.8}$$

$$e^{(t)}(X_i) := g(w^{(t)}; X_i) - y_i,\ z_i^{(t)} = \kappa_n(X_i)e^{(t)}(X_i),\ a_i = \beta(X_i)\kappa_n(X_i). \tag{4.9}$$

If we run gradient descent on solving (4.7) with step-size η, then we have (i) $z_i^{(t+1)} = f_{a_i}(z_i^{(t)})$ and thus Theorem 4.2.3 hols for f_{a_i} and $z_i^{(t)}$. (ii) The sharpness $\lambda_{\max}(\nabla^2\ell(w^{(t)})) = \max_{1\le i\le n}\frac{3z_i^{(t)}+2a_i}{\eta}$.

For this setup, the above theorem shows that the loss function is a summation of the loss on each individual data point. Recall that the training loss takes the form

$$\ell(w^{(t)}) = \frac{1}{2n}\sum_{i=1}^n \Big(g(w^{(t)}; X_i) - y_i\Big)^2 = \frac{1}{2n}\sum_{i=1}^n \frac{(z_i^{(t)})^2}{\kappa_n^2(X_i)} = \sum_{i=1}^n \frac{n(z_i^{(t)})^2}{2\eta^2\gamma^2\|X_i\|^4}. \tag{4.10}$$

Setting $\rho_i = \frac{n}{2\eta^2\gamma^2\|X_i\|^4}$, we can deduce that the dynamics of $\ell(w^{(t)})$ is given by Corollary 4.2.1. This leads to the following Corollary.

COROLLARY 4.3.1. *Under the setup in Theorem 4.3.2, for almost all $z^{(0)} \in \{z : -a_i \le z_i \le 2\}$ we have*

- *If* $0 < \max_{1\le i\le n} a_i \le 1$, *then* $\lim_{t\to\infty} \ell(w^{(t)}) = 0$. *Moreover, if* $0 < \max_{1\le i\le n} a_i \le 2\sqrt{2} - 2$, *the sequence* $\{\ell(w^{(t)})\}_{t=0}^{\infty}$ *is decreasing.*
- *If* $1 < \max_{1\le i\le n} a_i \le 2$, *then* $\{\ell(w^{(t)})\}_{t=0}^{\infty}$ *is bounded and does not converge to* 0.
- *If* $\max_{1\le i\le n} a_i > 2$, *then* $\lim_{t\to\infty} \ell(w^{(t)}) = +\infty$.

Under the orthogonality assumption, the loss functions defined on each data point exhibit a non-interacting behavior. Removing this orthogonality condition entirely is highly non-trivial. It would be interesting to extend our setting to the nearly-orthogonal one in Frei et al. [2022], and Kou et al. [2023].

4.3.2. Example 2: Neural network with quadratic activation. In this section, we consider the following two layer neural networks with its loss function on data point (X_i, y_i) defined as:

$$g(u, v; X_i) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} v_j \sigma\Big(\frac{1}{\sqrt{d}} u_j^\top X_i\Big), \ \ell_i = \frac{1}{2} \left(g(u, v; X_i) - y_i\right)^2 \tag{4.11}$$

where the hidden-layer weights $u_i \in \mathbb{R}^d$ are to be trained and outer-layer weights $v_i \in \mathbb{R}$ are held constant, which corresponds to the feature-learning setting for neural networks. Also m is the width of the hidden layer and σ is the activation function. Define $\mathbf{U} \coloneqq (u_1, ..., u_m)$. When the activation function is quadratic and $v_i = 1$ for all i, the loss function becomes

$$\min_{\mathbf{U}} \ell(\mathbf{U}) := \frac{1}{n} \sum_{j=1}^{n} \ell_j(\mathbf{U}) = \frac{1}{2n} \sum_{j=1}^{n} \Big(\frac{1}{\sqrt{md}} \sum_{i=1}^{m} (X_j^\top u_i)^2 - y_j\Big)^2. \tag{4.12}$$

As in the previous example, we assume $\mathbf{X}\mathbf{X}^\top = \mathrm{diag}(\|X_1\|^2, ..., \|X_n\|^2)$. We then have the following result on the gradient descent dynamics of the above problem.

THEOREM 4.3.3. *Define the following:*

$$e_i^{(t)} = \frac{1}{\sqrt{md}} \sum_{j=1}^{m} (X_i^\top u_j^{(t)})^2 - y_i, \ z_i^{(t)} = \frac{2\eta \|X_i\|^2 e_i^{(t)}}{\sqrt{m}dn}, \ a_i = \frac{2\eta \|X_i\|^2 y_i}{\sqrt{m}dn} \tag{4.13}$$

If we run gradient descent on solving problem (4.12) *with step-size* η, *we have* $z_i^{(t+1)} = f_{a_i}(z_i^{(t)})$ *and thus Theorem 4.2.3 and Corollary 4.3.1 hold for* $\ell(\mathbf{U}^{(t)})$.

The orthogonal assumption that $\mathbf{X}\mathbf{X}^\top = \mathrm{diag}(\|X_1\|^2, ..., \|X_n\|^2)$, helps decouple the loss function across the samples and makes the evolution of the overall loss non-interacting (across the

training samples). In order to relax this assumption, it is required to analyze bifurcation analysis of interacting dynamical systems, which is extremely challenging and not well-explored [Xu et al., 2021]. In Section A.6.2, we present empirical results showing that similar phases exists in the general non-orthogonal setting as well. Theoretically characterizing this is left as an open problem.

4.4. Experimental investigations

Before we proceed, we remark that the original PDF files for all the figures are provided as a part of the supplementary material for the sake of easier visualization. The naming convention is as follows: (i) each sub-folder correspond to the respective figure numbers and (ii) each file within a sub-folder is named according to matrix conventions. For e.g., file 1x3.pdf in sub-folder Figure 1 corresponds to Figure 4.1(c), and file 1x1.pdf in sub-folder Figure 3 corresponds to Figure 4.3.

4.4.1. Gradient descent dynamics with orthogonal data for model (4.12). Experimental setup. We now conduct experiments to evaluate the developed theory. We consider gradient descent for training the hidden layers of a two-layer neural network with orthogonal training data, described in Section 4.3.2. Recall that d, m, and n represents the dimension, hidden-layer width, and number of data points respectively. We set $d = 100, m \in \{5, 10, 25\}, n = 80$. We generate the ground-truth matrix $\mathbf{U}^* \in \mathbb{R}^{d\times m}$ where each entry is sampled from the standard normal distribution. The training data points collected in the data matrix, denoted as $\mathbf{X} \in \mathbb{R}^{n\times d}$, are the first n rows of a randomly generated orthogonal matrix. The labels are generated via the model in Section 4.3.2, i.e., $y_i = \frac{1}{\sqrt{md}} \sum_{j=1}^{m} \left(X_i^\top u_j\right)^2 + \varepsilon_i$ where ε_i is scalar noise sampled from a zero-mean normal distribution, with variances equal to $0, 0.25, 1$ in different experiments.

We set the step-size η such that $\max_{1\le i\le n} a_i$ defined in Theorem 4.3.2 belongs to the intervals of the first four phases. In particular, we choose $0.3, 0.9, 1, 1.2, 1.8$ for $m = 5, 10$ and $0.3, 0.9, 1, 1.2, 1.6$ for $m = 25$ (for each m, 0.9 and 1 are both in the catapult phase, and we pick 1 since it is the largest step-size choice allowed in the catapult phase). The numbers $0, 1, 2, 3, 4$ of the plot labels correspond to these step-size choices respectively. In Figure 4.4 we present the training loss curves in log scale and the sharpness curves for $m = 25$. The horizontal axes denote the number of steps of gradient descent. In Section A.6.1, we also provide additional simulation results for different hidden-layer

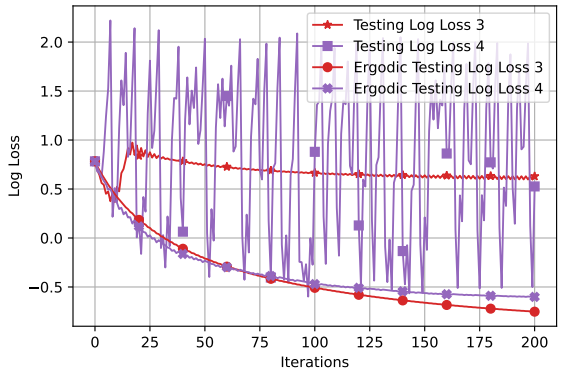


FIGURE 4.3. Test loss with and without averaging: The chaotic versions of purple and red lines correspond to the test error without averaging. The corresponding smooth versions refer to the test error with averaging. The plot demonstrates the benefit of ergodic trajectory averaging based predictions (according to Definition 4.4.1), as the averaging based predictions become more stable across the iterations. Numbers $3, 4$ denote different stepsize choices (see Section 4.4.2 for details).

widths. From the training loss curves (left column) and the sharpness curves (middle column) we can clearly observe the four phases[§] thereby confirming our theoretical results.

4.4.2. Prediction based on ergodic trajectory averaging. A main take-away from our analysis and experiments so far is that gradient descent with large step-size effectively resembles a randomized gradient descent procedure with a special type of noise, i.e., the randomness here is with respect to the orbit it converges to (in the non-monotonic phases).[¶] Recall that this viewpoint is also put-forward is several works, in particular Kong and Tao [2020]. Hence, a natural approach is to do perform ergodic trajectory averaging to reduce the fluctuations (see right column in Figure 4.4).

DEFINITION 4.4.1. *For any given point $X \in \mathbb{R}^d$, and any training iteration count t, the* ergodic trajectory averaging based prediction*, $\hat{y}$, for the point X is given by $\hat{y} := \frac{1}{t}\sum_{i=1}^{t} g(w^{(i)}; X)$, where $w^{(i)}$ corresponds to the training trajectory of the gradient descent algorithm trained with step-size η.*

Another way to think about the above prediction strategy is that the ergodic average approximates, in the limit, expectation with respect to the invariant distribution (supported on the orbit to which the trajectory converges to). In particular, Figure 4.4 right column, for the orthogonal setup, we see that as the noise increases, training in the chaotic regime and performing ergodic

[§]Here, we do not plot the divergent phase here for simplicity.

[¶]One way to show this formally is by connecting large step-size GD with slow-fast deterministic systems; see, for example, Chevyrev et al. [2020], Lim et al. [2022].

trajectory averaging provides a fast decay of training loss. A disadvantage of the ergodic averaging based prediction strategy described above is the test-time computational cost increases by $\mathcal{O}(t)$, per test point.

Figure 4.3 plots the testing loss for the model in (4.12), when trained with two values of large step-sizes ($\eta = 48, 60$). We observe that the ergodic trajectory averaging prediction smooths out the more chaotic testing loss. However, we also remark that from the plots in Figure A.7[‖], operating with slightly smaller step-size choice ($\eta = 36$) achieves the best testing error curves. See Section A.6.2 for additional observations. In the literature, ways of artificially inducing *controlled chaos* in the gradient descent trajectory has been proposed to obtain improved testing accuracy; see, for example, Lim et al. [2022]. We believe the ergodic trajectory averaging based prediction methodology discussed above may prove to be fruitful to stabilize the testing loss in such cases as well. A detailed investigation of provable benefits of the ergodic trajectory averaging predictor, is beyond the scope of the current work, and we leave it as intriguing future work.

Additional Experiments. We also provide the following additional simulation results in the appendix: (i) Section A.6.2 corresponds to non-orthogonal training data. We also include testing loss plots, and (ii) Section A.6.3 corresponds to training the hidden-layer weights of a two-layer neural network with ReLU activation functions and non-orthogonal inputs.

Take-away points from experiments. The main take-away points from the above experiments are the following: (i) in the case of orthogonal data, the experiments confirm the theoretical results in Section 4.3, (ii) in the case of non-orthogonal data, the experiments show that similar phases (including the chaotic phases) exists in the training dynamics, and (iii) ergodic averaging based prediction stabilizes the test error along the GD trajectory.

4.5. Conclusion

Unstable and chaotic behavior is frequently observed when training deep neural networks with large-order step-sizes. Motivated by this, we presented a fine-grained theoretical analysis of a cubic-map based dynamical system. We show that the gradient descent dynamics is fully captured by this dynamical system, when training the hidden layers of a two-layer neural networks with quadratic activation functions with orthogonal training data. Our analysis shows that for this class

[‖]Figure A.7 provides a detailed comparison across various step-sizes, for different noise variances.

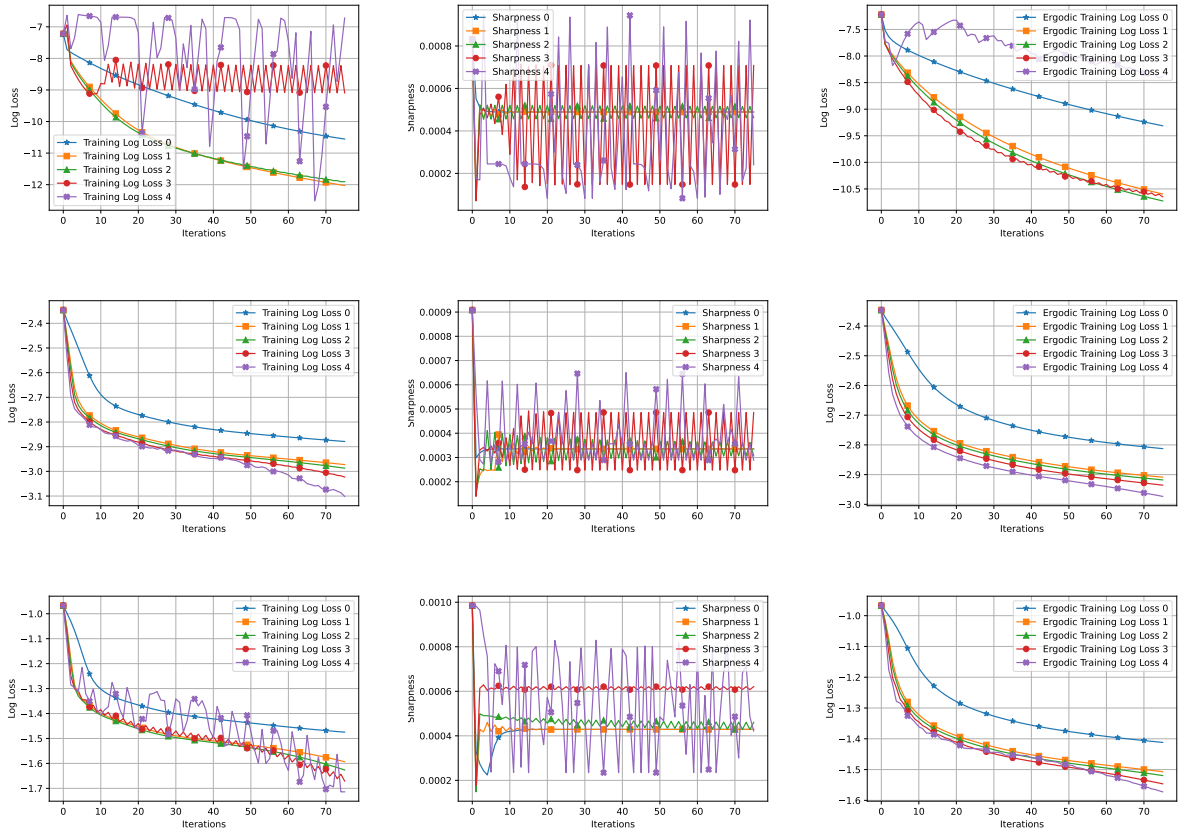


FIGURE 4.4. Hidden layer width = 25, with orthogonal data points. Rows from top to bottom represent different levels of noise – mean-zero normal distribution with variance $0, 0.25, 1$ respectively. The vertical axes are in log scale for the training loss curves. The second column is about the sharpness of the training loss functions. Numbers $0, 1, 2, 3, 4$ denote different stepsize choices (see Section 4.4.1 for details).

of models, as the step-size of the gradient descent increases, the gradient descent trajectory has five distinct phases (from being monotonic to chaotic and eventually divergent). We also provide empirical evidence that show similar behavior occurs for generic non-orthogonal data. Our results also indicate a subtle interplay on the relation between step-size and the initialization provided to the gradient descent algorithm in terms determining which phase the training trajectory will operate in. Finally, we empirically examined the impact of training in the different phases, on the generalization error.

Immediate future works include: (i) developing a theoretical characterization of the training dynamics with generic non-orthogonal training data, which involves undertaking non-trivial bifurcation analysis of interacting dynamical systems, (ii) moving beyond quadratic activation functions

and two-layer neural networks, and (iii) developing tight generalization bounds when training with large-order step-sizes. Overall, our contributions make concrete steps towards developing a fine-grained understanding of the gradient descent dynamics when training neural networks with iterative first-order optimization algorithms with large step-sizes.

APPENDIX A

Additional Experiments, Proofs, and Discussions

A.1. Proofs of Theorems in Chapter 2

We will prove Theorems 2.3.1 and 2.4.1 in Section A.1.1 and A.1.2 respectively. In each section we will first establish the relations between the optimality measure (see V_k in Section 2.3.3) and the gradient mapping, which reduce the proof of main theorems to proving the convergence of primal variables (x^k in Theorem 2.3.1 or (x^k, λ^k) in Theorem 2.4.1) and dual variables (h^k in Theorem 2.3.1 or (h_x^k, h_λ^k) in Theorem 2.4.1). Then we will prove the hypergradient estimation error, primal convergence and dual convergence separately. In our notation convention, the superscript k usually denotes the iteration number and the subscript i represents variables related to functions f_i, g_i. $L_\#$ with being a function $\#$ denotes its Lipschitz constant.

Next we state some technical lemmas that will be used in both sections.

LEMMA A.1.0.1 (Lemma 10 in Qu and Li [2017].). *Suppose $f(x)$ is μ-strongly convex and L-smooth. For any x and $\gamma < \frac{2}{\mu+L}$, define $x^+ = x - \gamma\nabla f(x)$, $x^* = \arg\min f(x)$. Then we have $\|x^+ - x^*\| \leq (1-\gamma\mu)\|x - x^*\|$.*

LEMMA A.1.0.2. *Define $\kappa = \max(L_{\nabla f}, L_{\nabla g})/\mu_g, z^*(x) = \left(\nabla_{22}^2 g(x, y^*(x))\right)^{-1} \nabla_2 f(x, y^*(x))$. Suppose Assumption 2 holds. Then $\Phi(x)$ is differentiable and $\nabla\Phi(x)$ is given by Then $\Phi(x), y^*(x)$, $z^*(x)$ are differentiable and $\nabla\Phi(x), y^*(x), z^*(x)$ are $L_{\nabla\Phi}, L_{y^*}, L_{z^*}$-Lipschitz continuous, and*

$$\nabla\Phi(x) = \nabla_1 f(x, y^*(x)) - \nabla_{12}^2 g(x, y^*(x)) \left(\nabla_{22}^2 g(x, y^*(x))\right)^{-1} \nabla_2 f(x, y^*(x)), \tag{A.1}$$

$$\nabla y^*(x) = -\nabla_{12}^2 g(x, y^*(x)) \left(\nabla_{22}^2 g(x, y^*(x))\right)^{-1}. \tag{A.2}$$

The constants are given by

$$L_{y^*} = \frac{L_{\nabla g}}{\mu_g} = \mathcal{O}(\kappa), \ L_{z^*} = \sqrt{1 + L_{y^*}^2}\left(\frac{L_{\nabla f}}{\mu_g} + \frac{L_f L_{\nabla_{22}^2 g}}{\mu_g^2}\right) = \mathcal{O}(\kappa^3),$$

$$L_{\nabla\Phi} = L_{\nabla f} + \frac{2L_{\nabla f}L_{\nabla g} + L_f^2 L_{\nabla^2 g}}{\mu_g} + \frac{2L_f L_{\nabla g} L_{\nabla^2 g} + L_{\nabla f} L_{\nabla g}^2}{\mu_g^2} + \frac{L_f L_{\nabla^2 g} L_{\nabla g}^2}{\mu_g^3} = \mathcal{O}(\kappa^3).$$

Moreover, we have

$$\|z^*(x)\| \leq \frac{L_f}{\mu_g}. \tag{A.3}$$

PROOF. See Lemma 2.2 in Ghadimi and Wang [2018] for the proof of (A.1) and (A.2), Lipschitz continuity of $\nabla\Phi$ and y^*. For the Lipschitz continuity of z^* we have for any $x, \tilde{x}$, we know

$$\begin{aligned}
&\|z^*(x) - z^*(\tilde{x})\| \\
=&\left\|\left(\nabla^2_{22}g(x, y^*(x))\right)^{-1}\nabla_2 f(x, y^*(x)) - \left(\nabla^2_{22}g(\tilde{x}, y^*(\tilde{x}))\right)^{-1}\nabla_2 f(\tilde{x}, y^*(\tilde{x}))\right\| \\
\leq&\left\|\left(\nabla^2_{22}g(x, y^*(x))\right)^{-1}\nabla_2 f(x, y^*(x)) - \left(\nabla^2_{22}g(\tilde{x}, y^*(\tilde{x}))\right)^{-1}\nabla_2 f(x, y^*(x))\right\| \\
&+\left\|\left(\nabla^2_{22}g(\tilde{x}, y^*(\tilde{x}))\right)^{-1}\nabla_2 f(x, y^*(x)) - \left(\nabla^2_{22}g(\tilde{x}, y^*(\tilde{x}))\right)^{-1}\nabla_2 f(\tilde{x}, y^*(\tilde{x}))\right\| \\
\leq&L_f\left\|\left(\nabla^2_{22}g(x, y^*(x))\right)^{-1}\right\|\left\|\nabla^2_{22}g(x, y^*(x)) - \nabla^2_{22}g(\tilde{x}, y^*(\tilde{x}))\right\|\left\|\left(\nabla^2_{22}g(x, y^*(x))\right)^{-1}\right\| \\
&+\frac{1}{\mu_g}\left\|\nabla_2 f(x, y^*(x)) - \nabla_2 f(\tilde{x}, y^*(\tilde{x}))\right\| \\
\leq&\frac{L_f L_{\nabla^2_{22}g}}{\mu_g^2}\sqrt{\|x-\tilde{x}\|^2 + \|y^*(x) - y^*(\tilde{x})\|^2} + \frac{L_{\nabla f}}{\mu_g}\sqrt{\|x-\tilde{x}\|^2 + \|y^*(x) - y^*(\tilde{x})\|^2} \\
\leq&L_{z^*}\left\|x - \tilde{x}\right\|,
\end{aligned}$$

where the first inequality uses triangle inequality, the second and third inequalities use Assumption 2, and the fourth inequality uses Lipschitz continuity of $y^*(x)$. The inequality in (A.3) holds since $g(x, \cdot)$ is μ_g-strongly convex and $\|\nabla_2 f(x, y^*(x))\| \leq L_f$ (Assumption 2). □

LEMMA A.1.0.3 (Lemma 3.2 in Ghadimi et al. [2020]). *For any closed convex set $\mathcal{X}$, and the function $\eta_{\mathcal{X}}(x, h, \tau)$ defined in Section 2.3.3 is differentiable and $\nabla\eta_{\mathcal{X}}$ is $L_{\nabla\eta_{\mathcal{X}}}$-Lipschitz continuous, with the closed form expression and constant given by*

$$\nabla_1\eta_{\mathcal{X}}(x, h, \tau) = -h + \frac{1}{\tau}(x - \bar{d}), \nabla_2\eta_{\mathcal{X}}(x, h, \tau) = \bar{d} - x, L_{\nabla\eta_{\mathcal{X}}} = 2\sqrt{(1 + 1/\tau)^2 + (1 + \tau/2)^2},$$

where $\bar{d}$ is defined as $\bar{d} = \arg\min_{d\in\mathcal{X}}\{\langle h, d - x\rangle + \frac{1}{2\tau}\|d - x\|^2\} = \Pi_{\mathcal{X}}(x - \tau h)$, which satisfies

$$\langle h + \frac{1}{\tau}(\bar{d} - x), d - \bar{d}\rangle \geq 0, \quad \textit{for all } d \in \mathcal{X}. \tag{A.4}$$

A.1.1. Proof of Theorem 2.3.1. For simplicity, we summarize the notations that will be used in Section A.1.1 as follows.

$$\begin{aligned}
&\kappa = \max(L_{\nabla f}, L_{\nabla g})/\mu_g,\ w^{k+1} = u_x^{k+1} - J^{k+1}z^k,\\
&y_*^k = y^*(x^k) = \arg\min_{y\in\mathbb{R}^{d_y}} g(x^k, y),\ z_*^k = \left(\nabla_{22}^2 g(x^k, y_*^k)\right)^{-1}\nabla_2 f(x^k, y_*^k),\\
&\Phi(x) = f(x, y^*(x)),\ \eta_{\mathcal{X}}(x, h, \tau) = \min_{d\in X}\left\{\langle h, d - x\rangle + \frac{1}{2\tau}\|d - x\|^2\right\}.
\end{aligned} \tag{A.5}$$

In this section we suppose Assumptions 2 and 3 hold. We assume stepsizes in Algorithm 1 satisfy $\beta_k = c_1\alpha_k,\ \gamma_k = c_2\alpha_k, \theta_k = c_3\alpha_k$, where $c_1, c_2, c_3 > 0$ are constants to be determined. We will utilize the following merit function in our analysis:

$$W_k = \underbrace{\Phi(x^k) - \inf_{x\in\mathcal{X}}\Phi(x) - \frac{1}{c_3}\eta_{\mathcal{X}}(x^k, h^k, \tau)}_{W_{k,1}} + \underbrace{\frac{1}{c_1}\left\|y^k - y_*^k\right\|^2 + \frac{1}{c_2}\left\|z^k - z_*^k\right\|^2}_{W_{k,2}}.$$

By definition of $\eta_{\mathcal{X}}$, we can verify that $W_{k,1} \geq 0$. Moreover, as discussed in Section 2.3.3, we consider the following optimality measure:

$$V_k = \frac{1}{\tau^2}\|x_+^k - x^k\|^2 + \|h^k - \nabla\Phi(x^k)\|^2. \tag{A.6}$$

Next we characterize the relation between V_k and gradient mapping of problem 2.1.

LEMMA A.1.0.4. *Suppose Assumptions 2 and 3 hold. In Algorithm 1 we have*

$$\frac{1}{\tau^2}\left\|x^k - \Pi_{\mathcal{X}}\left(x^k - \tau\nabla\Phi(x^k)\right)\right\|^2 \leq 2V_k.$$

PROOF. Note that we have

$$\begin{aligned}
\left\|x^k - \Pi_{\mathcal{X}}\left(x^k - \tau\nabla\Phi(x^k)\right)\right\|^2 \leq& 2\Big(\left\|x_+^k - x^k\right\|^2 + \left\|\Pi_{\mathcal{X}}\left(x^k - \tau h^k\right) - \Pi_{\mathcal{X}}\left(x^k - \tau\nabla\Phi(x^k)\right)\right\|^2\Big)\\
\leq& 2\Big(\left\|x_+^k - x^k\right\|^2 + \tau^2\left\|h^k - \nabla\Phi(x^k)\right\|^2\Big) = 2\tau^2 V_k,
\end{aligned}$$

where the first inequality uses Cauchy-Schwarz inequality and the second inequality uses the non-expansiveness of projection onto a closed convex set. This completes the proof. □

Then we bound the variance of w^{k+1} and $\left\|h^{k+1} - h^k\right\|$.

LEMMA A.1.0.5. *Suppose Assumptions 2 and 3 hold. In Algorithm 1 we have*

$$\mathbb{E}\left[\left\|w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2\right]\leq\sigma_{w,k+1}^2$$

$$\sigma_{w,k+1}^2:=\sigma_w^2+2\sigma_{g,2}^2\mathbb{E}\left[\left\|z^k-z_*^k\right\|^2\right],\ \sigma_w^2=\sigma_{f,1}^2+\frac{2\sigma_{g,2}^2L_f^2}{\mu_g^2}, \tag{A.7}$$

$$\mathbb{E}\left[\left\|h^{k+1}-h^k\right\|^2\right]\leq\sigma_{h,k}^2,$$

$$\sigma_{h,k}^2:=2\theta_k^2\mathbb{E}\left[\left\|h^k-\nabla\Phi(x^k)\right\|^2+\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\right\|^2\right]+\theta_k^2\sigma_{w,k+1}^2. \tag{A.8}$$

PROOF. We first consider w^k. Note that

$$w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k]=u_x^{k+1}-\mathbb{E}[u_x^{k+1}|\mathscr{F}_k]-\left(J^{k+1}-\mathbb{E}\left[J^{k+1}|\mathscr{F}_k\right]\right)z^k.$$

Hence we know

$$\begin{aligned}
&\mathbb{E}\left[\left\|w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2|\mathscr{F}_k\right]\\
=&\mathbb{E}\left[\left\|u_x^{k+1}-\mathbb{E}[u_x^{k+1}|\mathscr{F}_k]\right\|^2|\mathscr{F}_k\right]+\mathbb{E}\left[\left\|J^{k+1}-\mathbb{E}\left[J^{k+1}|\mathscr{F}_k\right]\right\|^2|\mathscr{F}_k\right]\left\|z^k\right\|^2\\
\leq&\sigma_{f,1}^2+2\sigma_{g,2}^2\left\|z_*^k\right\|^2+2\sigma_{g,2}^2\left\|z^k-z_*^k\right\|^2\leq\sigma_{f,1}^2+\frac{2\sigma_{g,2}^2L_f^2}{\mu_g^2}+2\sigma_{g,2}^2\left\|z^k-z_*^k\right\|^2,
\end{aligned}$$

where the first equality uses independence, the first inequality uses Cauchy-Schwarz inequality, and the second inequality uses (A.3). This proves (A.7). Next for $\left\|h^{k+1}-h^k\right\|$ we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|h^{k+1}-h^k\right\|^2|\mathscr{F}_k\right]\\
=&\theta_k^2\mathbb{E}\left[\left\|h^k-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2|\mathscr{F}_k\right]+\theta_k^2\mathbb{E}\left[\left\|w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2|\mathscr{F}_k\right]\\
\leq&2\theta_k^2\mathbb{E}\left[\left\|h^k-\nabla\Phi(x^k)\right\|^2|\mathscr{F}_k\right]+2\theta_k^2\mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\right\|^2|\mathscr{F}_k\right]+\theta_k^2\sigma_{w,k+1}^2,
\end{aligned}$$

which proves of (A.8) by taking expectation on both sides. □

REMARK. We would like to highlight that in (A.7), we explicitly characterize the upper bound of the variance of w^{k+1}, which contains $\mathbb{E}\left[\left\|z^k-z_*^k\right\|^2\right]$ and requires further analysis. In contrast, Assumption 3.7 in Dagréou et al. [2022] directly assumes the second moment of D_x^t is uniformly bounded, i.e., $\mathbb{E}\left[\left\|D_x^t\right\|^2\right]\leq B_x^2$ for some constant $B_x\geq 0$. Note that D_x^t in Dagréou et al. [2022] is

the same as our w^{k+1} (see (2.6), line 5 of Algorithm 1 and definition of w^{k+1} in (A.5)). The second moment bound can directly imply the variance bound, i.e., $\mathbb{E}\left[\left\|D_x^t - \mathbb{E}\left[D_x^t\right]\right\|^2\right] \leq \mathbb{E}\left[\left\|D_x^t\right\|^2\right] \leq B_x^2$. This implies that some stronger assumptions are needed to guarantee Assumption 3.7 in Dagréou et al. [2022], as also pointed out by the authors (see discussions right below it). Instead, our refined analysis does not require that.

A.1.1.1. *Hypergradient Estimation Error.* Note that Assumptions 3.1 and 3.2 in Dagréou et al. [2022] state that the upper-level function f is twice differentiable, the lower-level function g is three times differentiable and $\nabla^2 f, \nabla^3 g$ are Lipschitz continuous so that z_*^k, as a function of x^k (see (A.5)), is smooth, which is a crucial condition for (31) and (81) in Dagréou et al. [2022] ($v^*(x^t)$ in their notation), which follows the analysis in Equation (49) in Chen et al. [2021a]. In this section we show that, by incorporating the moving-average technique recently introduced to decentralized bilevel optimization [Chen et al., 2023b], we can remove this additional assumption. We have the following lemma characterizing the error induced by y^k and z^k.

LEMMA A.1.0.6. *Suppose Assumptions 2 and 3 hold. If the stepsizes satisfy*

$$\beta_k < \frac{2}{\mu_g + L_{\nabla g}},\ \gamma_k \leq \min\left(\frac{1}{4\mu_g},\ \frac{0.06\mu_g}{\sigma_{g,2}^2}\right), \tag{A.9}$$

then in Algorithm 1 we have

$$\begin{aligned}
&\sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\left\|y^k - y_*^k\right\|^2\right] \leq C_{yx} \sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + C_{y,0} + C_{y,1}\left(\sum_{k=0}^{K} \alpha_k^2\right) \\
&\sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\left\|z^k - z_*^k\right\|^2\right] \leq C_{zx} \sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + C_{z,0} + C_{z,1}\left(\sum_{k=0}^{K} \alpha_k^2\right).
\end{aligned} \tag{A.10}$$

where the constants are defined as

$$\begin{aligned}
C_{yx} &= \frac{2L_{y^*}^2}{c_1^2\mu_g^2},\ C_{y,0} = \frac{1}{c_1\mu_g}\mathbb{E}\left[\left\|y^0 - y_*^0\right\|^2\right],\ C_{y,1} = \frac{2c_1\sigma_{g,1}^2}{\mu_g}, \\
C_{zx} &= \frac{5L_f^2}{\mu_g^2}\left(\frac{L_{\nabla_{22}^2 g}^2}{\mu_g^2} + 1\right)\frac{2L_{y^*}^2}{c_1^2\mu_g^2} + \frac{4L_{z^*}^2}{c_2^2\mu_g^2}, \\
C_{z,0} &= \frac{5L_f^2}{\mu_g^2}\left(\frac{L_{\nabla_{22}^2 g}^2}{\mu_g^2} + 1\right) \cdot \frac{1}{c_1\mu_g}\mathbb{E}\left[\left\|y^0 - y_*^0\right\|^2\right] + \frac{1}{c_2\mu_g}\mathbb{E}\left[\left\|z^0 - z_*^0\right\|^2\right], \\
C_{z,1} &= \frac{5L_f^2}{\mu_g^2}\left(\frac{L_{\nabla_{22}^2 g}^2}{\mu_g^2} + 1\right) \cdot \frac{2c_1\sigma_{g,1}^2}{\mu_g} + \frac{2c_2\sigma_w^2}{\mu_g}.
\end{aligned}$$

PROOF. We first consider the error induced by y^k. We have

$$\begin{aligned}\left\|y^{k+1}-y_*^{k+1}\right\|^2 &\le (1+\beta_k\mu_g)\left\|y^{k+1}-y_*^k\right\|^2+\left(1+\frac{1}{\beta_k\mu_g}\right)\left\|y_*^{k+1}-y_*^k\right\|^2\\ &\le (1+\beta_k\mu_g)\left\|y^{k+1}-y_*^k\right\|^2+\left(\frac{\alpha_k^2}{\beta_k\mu_g}+\alpha_k^2\right)L_{y^*}^2\left\|x_+^k-x^k\right\|^2,\end{aligned}\tag{A.11}$$

where the first inequality uses Cauchy-Schwarz inequality: $\|u+v\|^2 \le (1+c)\big(\|u\|^2+\frac{1}{c}\|v\|^2\big)$, for any vectors u, v and constant $c>0$. Thanks to the moving-average step of x^k, our analysis of $\|y_*^{k+1}-y_*^k\|$ is simplified comparing to that in Chen et al. [2021a]. Also,

$$\begin{aligned}&\mathbb{E}\left[\left\|y^{k+1}-y_*^k\right\|^2|\mathscr{F}_k\right]=\mathbb{E}\left[\left\|y^k-\beta_k\nabla_2 g(x^k,y^k)-y_*^k-\beta_k(v^{k+1}-\nabla_2 g(x^k,y^k))\right\|^2|\mathscr{F}_k\right]\\ &\le\left\|y^k-\beta_k\nabla_2 g(x^k,y^k)-y_*^k\right\|^2+\beta_k^2\sigma_{g,1}^2\le(1-\beta_k\mu_g)^2\left\|y^k-y_*^k\right\|^2+\beta_k^2\sigma_{g,1}^2,\end{aligned}\tag{A.12}$$

where the first inequality uses Assumption (3) and Lemma A.1.0.1, and the second inequality uses Lemma A.1.0.1 (which requires strong convexity of g, Lipschtiz continuity of $\nabla_2 g$, and the first inequality in (A.9)). Combining (A.11) and (A.12), we know

$$\begin{aligned}&\mathbb{E}\left[\left\|y^{k+1}-y_*^{k+1}\right\|^2|\mathscr{F}_k\right]\\ \le&(1+\beta_k\mu_g)(1-\beta_k\mu_g)^2\left\|y^k-y_*^k\right\|^2+\left(\frac{\alpha_k^2}{\beta_k\mu_g}+\alpha_k^2\right)L_{y^*}^2\left\|x_+^k-x^k\right\|^2+(1+\beta_k\mu_g)\beta_k^2\sigma_{g,1}^2\\ \le&(1-\beta_k\mu_g)\left\|y^k-y_*^k\right\|^2+\frac{2\alpha_k^2L_{y^*}^2}{\beta_k\mu_g}\left\|x_+^k-x^k\right\|^2+2\beta_k^2\sigma_{g,1}^2.\end{aligned}$$

where the second inequality uses $\beta_k<\frac{2}{\mu_g+L_{\nabla g}}\le\frac{1}{\mu_g}$. Taking summation ($k$ from 0 to K) on both sides and taking expectation, we know

$$\sum_{k=0}^K\beta_k\mu_g\mathbb{E}\left[\left\|y^k-y_*^k\right\|^2\right]\le\mathbb{E}\left[\left\|y^0-y_*^0\right\|^2\right]+\sum_{k=0}^K\frac{2\alpha_k^2L_{y^*}^2}{\beta_k\mu_g}\mathbb{E}\left[\left\|x_+^k-x^k\right\|^2\right]+\sum_{k=0}^K2\beta_k^2\sigma_{g,1}^2,$$

which proves the first inequality in (A.10) by dividing $c_1\mu_g$ on both sides. Next we analyze the error induced by z^k. Our analysis is substantially different from Dagréou et al. [2022]:

$$\begin{aligned}\left\|z^{k+1}-z_*^{k+1}\right\|^2&\le\left(1+\frac{\gamma_k\mu_g}{3}\right)\left\|z^{k+1}-z_*^k\right\|^2+\left(1+\frac{3}{\gamma_k\mu_g}\right)\left\|z_*^{k+1}-z_*^k\right\|^2\\ &\le\left(1+\frac{\gamma_k\mu_g}{3}\right)\left\|z^{k+1}-z_*^k\right\|^2+\left(\frac{3\alpha_k^2}{\gamma_k\mu_g}+\alpha_k^2\right)L_{z^*}^2\left\|x_+^k-x^k\right\|^2\end{aligned}\tag{A.13}$$

where we use Cauchy-Schwarz inequality in the first and second inequality, we use the facts that ∇y^* is Lipschitz continuous. For $\left\|z^{k+1}-z_*^k\right\|$, we may follow the analysis of SGD under the strongly convex setting:

$$\begin{aligned}
&z^{k+1}-z_*^k\\
=&z^k-\gamma_k(H^kz^k-u_y^k)-z_*^k\\
=&z^k-\gamma_k\nabla_{22}^2g(x^k,y^k)z^k+\gamma_k\nabla_2f(x^k,y^k)-z_*^k-\gamma_k(H^{k+1}-\nabla_{22}^2g(x^k,y^k))z^k\\
&+\gamma_k(u_y^k-\nabla_2f(x^k,y^k))
\end{aligned}$$

which gives

$$\begin{aligned}
&\mathbb{E}\left[\left\|z^{k+1}-z_*^k\right\|^2|\mathscr{F}_k\right]\\
\le&\left\|z^k-\gamma_k\nabla_{22}^2g(x^k,y^k)z^k+\gamma_k\nabla_2f(x^k,y^k)-z_*^k\right\|^2+\gamma_k^2\sigma_{g,2}^2\left\|z^k\right\|^2+\gamma_k^2\sigma_{f,1}^2\\
=&\left\|(I-\gamma_k\nabla_{22}^2g(x^k,y^k))(z^k-z_*^k)-\gamma_k(\nabla_{22}^2g(x^k,y^k)z_*^k-\nabla_2f(x^k,y^k))\right\|^2+\gamma_k^2\sigma_{g,2}^2\left\|z^k\right\|^2+\gamma_k^2\sigma_{f,1}^2\\
\le&\left(1+\frac{\gamma_k\mu_g}{2}\right)\left\|(I-\gamma_k\nabla_{22}^2g(x^k,y^k))(z^k-z_*^k)\right\|^2\\
&+\left(1+\frac{2}{\gamma_k\mu_g}\right)\left\|\gamma_k\left(\nabla_{22}^2g(x^k,y^k)z_*^k-\nabla_{22}^2g(x^k,y_*^k)z_*^k+\nabla_2f(x^k,y_*^k)-\nabla_2f(x^k,y^k)\right)\right\|^2\\
&+2\gamma_k^2\sigma_{g,2}^2\left(\left\|z^k-z_*^k\right\|^2+\left\|z_*^k\right\|^2\right)+\gamma_k^2\sigma_{f,1}^2\\
\le&\left(\left(1+\frac{\gamma_k\mu_g}{2}\right)(1-\gamma_k\mu_g)^2+2\gamma_k^2\sigma_{g,2}^2\right)\left\|z^k-z_*^k\right\|^2\\
&+\left(\frac{4\gamma_k}{\mu_g}+2\gamma_k^2\right)\left(L_{\nabla_{22}^2g}^2\left\|z_*^k\right\|^2+L_{\nabla_2f}^2\right)\left\|y^k-y_*^k\right\|^2+2\gamma_k^2\sigma_{g,2}^2\left\|z_*^k\right\|^2+\gamma_k^2\sigma_{f,1}^2.
\end{aligned}\tag{A.14}$$

$$\le\left(1-\frac{4\gamma_k\mu_g}{3}\right)\left\|z^k-z_*^k\right\|^2+\left(\frac{4\gamma_k}{\mu_g}+2\gamma_k^2\right)\left(\frac{L_{\nabla_{22}^2g}^2L_f^2}{\mu_g^2}+L_f^2\right)\left\|y^k-y_*^k\right\|^2+\left(\frac{2\sigma_{g,2}^2L_f^2}{\mu_g^2}+\sigma_{f,1}^2\right)\gamma_k^2,$$

where the first inequality uses Assumption 3, the second inequality uses Cauchy-Schwarz inequality and the definition of z_*^k, the third inequality uses Cauchy-Schwarz inequality and the fact that g is μ_g-strongly convex, and the fourth inequality uses Cauchy-Schwarz inequality, (A.3) and $-\frac{\gamma_k\mu_g}{6}+2\gamma_k^2\sigma_{g,2}^2+\frac{\gamma_k^3\mu_g^3}{2}\le 0$, which is a direct result from the bound of γ_k in (A.9). It is worth noting that our estimation can be viewed as a refined version of (72) - (75) in Dagréou et al. [2022]

Combining (A.13) and (A.14) we may obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left\|z^{k+1}-z_*^{k+1}\right\|^2|\mathscr{F}_k\right]\\
\leq&\left(1+\frac{\gamma_k\mu_g}{3}\right)\mathbb{E}\left[\left\|z^{k+1}-z_*^{k}\right\|^2|\mathscr{F}_k\right]+\left(\frac{3\alpha_k^2}{\gamma_k\mu_g}+\alpha_k^2\right)L_{z^*}^2\left\|x_+^k-x^k\right\|^2\\
\leq&\left(1+\frac{\gamma_k\mu_g}{3}\right)\left[\left(1-\frac{4\gamma_k\mu_g}{3}\right)\left\|z^k-z_*^k\right\|^2+\left(\frac{4\gamma_k}{\mu_g}+2\gamma_k^2\right)\left(\frac{L_{\nabla_{22}^2g}^2L_f^2}{\mu_g^2}+L_f^2\right)\left\|y^k-y_*^k\right\|^2\right]\\
&+\left(1+\frac{\gamma_k\mu_g}{3}\right)\left(\frac{2\sigma_{g,2}^2L_f^2}{\mu_g^2}+\sigma_{f,1}^2\right)\gamma_k^2+\left(\frac{3\alpha_k^2}{\gamma_k\mu_g}+\alpha_k^2\right)L_{z^*}^2\left\|x_+^k-x^k\right\|^2\\
=&\left(1-\gamma_k\mu_g\right)\left\|z^k-z_*^k\right\|^2+\left(\frac{4\gamma_k}{\mu_g}+\frac{10\gamma_k^2}{3}+\frac{2\gamma_k^3\mu_g}{3}\right)\left(\frac{L_{\nabla_{22}^2g}^2L_f^2}{\mu_g^2}+L_f^2\right)\left\|y^k-y_*^k\right\|^2\\
&+\sigma_w^2\left(\gamma_k^2+\frac{\gamma_k^3\mu_g}{3}\right)+\left(\frac{3\alpha_k^2}{\gamma_k\mu_g}+\alpha_k^2\right)L_{z^*}^2\left\|x_+^k-x^k\right\|^2\\
\leq&\left(1-\gamma_k\mu_g\right)\left\|z^k-z_*^k\right\|^2+\frac{5\gamma_kL_f^2}{\mu_g}\left(\frac{L_{\nabla_{22}^2g}^2}{\mu_g^2}+1\right)\left\|y^k-y_*^k\right\|^2+2\sigma_w^2\gamma_k^2+\frac{4\alpha_k^2L_{z^*}^2}{\gamma_k\mu_g}\left\|x_+^k-x^k\right\|^2,
\end{aligned}
$$

where the equality uses the definition of σ_w^2 in (A.7) and the third inequality uses $\gamma_k\mu_g\leq\frac{1}{4}$. Taking summation ($k$ from 0 to K) and expectation, we know

$$
\begin{aligned}
\sum_{k=0}^K\gamma_k\mu_g\mathbb{E}\left[\left\|z^k-z_*^k\right\|^2\right]\leq&\mathbb{E}\left[\left\|z^0-z_*^0\right\|^2\right]+\sum_{k=0}^K\frac{5\gamma_kL_f^2}{\mu_g}\left(\frac{L_{\nabla_{22}^2g}^2}{\mu_g^2}+1\right)\mathbb{E}\left[\left\|y^k-y_*^k\right\|^2\right]\\
&+\sum_{k=0}^K2\sigma_w^2\gamma_k^2+\sum_{k=0}^K\frac{4\alpha_k^2L_{z^*}^2}{\gamma_k\mu_g}\mathbb{E}\left[\left\|x_+^k-x^k\right\|^2\right].
\end{aligned}
$$

This completes the proof of the second inequality in (A.10) by dividing $c_2\mu_g$ on both sides and replacing $\sum_{k=0}^K\alpha_k\mathbb{E}\left[\left\|y^k-y_*^k\right\|^2\right]$ with its upper bound in (A.10). □

Lemma A.1.0.7. *Suppose Assumptions 2 and 3 hold. We have*

$$
\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\right\|^2\leq3\left(\left(L_{\nabla f}^2+L_{\nabla^2g}^2\right)\left\|y^k-y_*^k\right\|^2+L_{\nabla g}^2\left\|z^k-z_*^k\right\|^2\right),
$$

Proof. Note that we have the following decomposition:

$$
\begin{aligned}
&\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\\
=&\mathbb{E}[u_x^{k+1}|\mathscr{F}_k]-\nabla_1f(x^k,y_*^k)-\left(\mathbb{E}\left[J^{k+1}|\mathscr{F}_k\right]z^k-\nabla_{12}^2g(x^k,y_*^k)z_*^k\right)
\end{aligned}
$$

$$=\nabla_1 f(x^k,y^k)-\nabla_1 f(x^k,y_*^k)-\nabla_{12}^2 g(x^k,y^k)\left(z^k-z_*^k\right)-\left(\nabla_{12}^2 g(x^k,y^k)-\nabla_{12}^2 g(x^k,y_*^k)\right)z_*^k.$$

which, together with Cauchy-Schwarz inequality, implies the conclusion:

$$\begin{aligned}\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\right\|^2 \leq& 3\left\|\nabla_1 f(x^k,y^k)-\nabla_1 f(x^k,y_*^k)\right\|^2+3\left\|\nabla_{12}^2 g(x^k,y^k)(z^k-z_*^k)\right\|^2\\ &+3\left\|\left(\nabla_{12}^2 g(x^k,y^k)-\nabla_{12}^2 g(x^k,y_*^k)\right)z_*^k\right\|^2\\ \leq& 3\left(\left(L_{\nabla f}^2+L_{\nabla^2 g}^2\right)\left\|y^k-y_*^k\right\|^2+L_{\nabla g}^2\left\|z^k-z_*^k\right\|^2\right).\end{aligned}$$

This completes the proof. □

A.1.1.2. *Primal Convergence.*

LEMMA A.1.0.8. *Suppose Assumptions 2 and 3 hold. If*

$$\alpha_k\leq\min\left(\frac{\tau^2}{20c_3},\ \frac{c_3}{2\tau\left(c_3L_{\nabla\Phi}+L_{\nabla\eta_{\mathcal{X}}}\right)},\ 1\right),\ \tau<1,\ c_3\leq\frac{1}{10},\tag{A.15}$$

then in Algorithm 1 we have

$$\begin{aligned}\sum_{k=0}^K\frac{\alpha_k}{\tau^2}\mathbb{E}\left[\left\|x_+^k-x^k\right\|^2\right]\leq&\frac{2}{\tau}\mathbb{E}\left[W_{0,1}\right]+3\sum_{k=0}^K\alpha_k\mathbb{E}\left[\left\|\nabla\Phi(x^k)-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2\right]\\ &+\frac{1}{2}\sum_{k=0}^K\alpha_k\mathbb{E}\left[\left\|h^k-\nabla\Phi(x^k)\right\|^2\right]+\sum_{k=0}^K\left(\alpha_k^2\sigma_{g,2}^2\mathbb{E}\left[\left\|z^k-z_*^k\right\|^2\right]+\alpha_k^2\sigma_w^2\right).\end{aligned}\tag{A.16}$$

PROOF. The smoothness of $\Phi(x)$ and $\eta_{\mathcal{X}}$ (Lemmas A.1.0.2, A.1.0.3) imply

$$\Phi(x^{k+1})-\Phi(x^k)\leq\alpha_k\langle\nabla\Phi(x^k),x_+^k-x^k\rangle+\frac{L_{\nabla\Phi}}{2}\left\|x^{k+1}-x^k\right\|^2\tag{A.17}$$

and

$$\begin{aligned}&\eta_{\mathcal{X}}(x^k,h^k,\tau)-\eta_{\mathcal{X}}(x^{k+1},h^{k+1},\tau)\\ \leq&\langle-h^k+\frac{1}{\tau}(x^k-x_+^k),x^k-x^{k+1}\rangle+\langle x_+^k-x^k,h^k-h^{k+1}\rangle+\frac{L_{\nabla\eta_{\mathcal{X}}}}{2}\left(\left\|x^{k+1}-x^k\right\|^2+\left\|h^{k+1}-h^k\right\|^2\right)\\ =&\alpha_k\langle h^k,x_+^k-x^k\rangle+\frac{\alpha_k}{\tau}\left\|x_+^k-x^k\right\|^2+\theta_k\langle h^k,x_+^k-x^k\rangle-\theta_k\langle w^{k+1},x_+^k-x^k\rangle\\ &+\frac{L_{\nabla\eta_{\mathcal{X}}}}{2}\left(\left\|x^{k+1}-x^k\right\|^2+\left\|h^{k+1}-h^k\right\|^2\right)\end{aligned}$$

(A.18)
$$\leq -\frac{\theta_k}{\tau}\left\|x_+^k - x^k\right\|^2 - \theta_k\langle w^{k+1}, x_+^k - x^k\rangle + \frac{L_{\nabla\eta_{\mathcal{X}}}}{2}\left(\left\|x^{k+1}-x^k\right\|^2 + \left\|h^{k+1}-h^k\right\|^2\right),$$

where the first inequality uses $L_{\nabla\eta_{\mathcal{X}}}$-smoothness of $\nabla\eta_{\mathcal{X}}$, and the second inequality uses the optimality condition (A.4) (with $d = x^k$). Hence by computing (A.17) + (A.18)/c_3 and taking conditional expectation with respect to $\mathscr{F}_k$ we know

$$\begin{aligned}
&\frac{\alpha_k}{\tau}\left\|x_+^k - x^k\right\|^2 \\
\leq&\frac{1}{c_3}\left(\mathbb{E}\left[\eta_{\mathcal{X}}(x^{k+1},h^{k+1},\tau)|\mathscr{F}_k\right] - \eta_{\mathcal{X}}(x^k,h^k,\tau)\right) + \Phi(x^k) - \mathbb{E}\left[\Phi(x^{k+1})|\mathscr{F}_k\right] \\
&+\alpha_k\langle\nabla\Phi(x^k) - \mathbb{E}[w^{k+1}|\mathscr{F}_k], x_+^k - x^k\rangle + \frac{(c_3L_{\nabla\Phi} + L_{\nabla\eta_{\mathcal{X}}})}{2c_3}\left\|x^{k+1}-x^k\right\|^2 \\
&+\frac{L_{\nabla\eta_{\mathcal{X}}}}{2c_3}\mathbb{E}\left[\left\|h^{k+1}-h^k\right\|^2|\mathscr{F}_k\right] \\
=&W_{k,1} - \mathbb{E}\left[W_{k+1,1}|\mathscr{F}_k\right] + \alpha_k\langle\nabla\Phi(x^k) - \mathbb{E}[w^{k+1}|\mathscr{F}_k], x_+^k - x^k\rangle \\
&+\frac{(c_3L_{\nabla\Phi} + L_{\nabla\eta_{\mathcal{X}}})}{2c_3}\left\|x^{k+1}-x^k\right\|^2 + \frac{L_{\nabla\eta_{\mathcal{X}}}}{2c_3}\mathbb{E}\left[\left\|h^{k+1}-h^k\right\|^2|\mathscr{F}_k\right] \\
\leq&W_{k,1} - \mathbb{E}\left[W_{k+1,1}|\mathscr{F}_k\right] + \alpha_k\Big(\tau\left\|\nabla\Phi(x^k) - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2 + \frac{1}{4\tau}\left\|x_+^k - x^k\right\|^2\Big) \\
&+\frac{\alpha_k}{4\tau}\left\|x_+^k - x^k\right\|^2 + \frac{5}{2c_3\tau}\mathbb{E}\left[\left\|h^{k+1}-h^k\right\|^2|\mathscr{F}_k\right],
\end{aligned} \tag{A.19}$$

where the second inequality uses Young's inequality and $\frac{\alpha_k^2(c_3L_{\nabla\Phi}+L_{\nabla\eta_{\mathcal{X}}})}{2c_3} \leq \frac{\alpha_k}{4\tau}$, $L_{\nabla\eta_{\mathcal{X}}} < \frac{5}{\tau}$ when (A.15) holds. Note that by (A.8) we know

(A.20)
$$\begin{aligned}
&\frac{5}{c_3\tau^2}\mathbb{E}\left[\left\|h^{k+1}-h^k\right\|^2\right] \\
\leq&\frac{10c_3\alpha_k^2}{\tau^2}\mathbb{E}\left[\left\|h^k - \nabla\Phi(x^k)\right\|^2 + \left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Phi(x^k)\right\|^2\right] + \frac{5c_3\alpha_k^2}{\tau^2}\sigma_w^2 \\
+&\frac{10c_3\alpha_k^2\sigma_{g,2}^2}{\tau^2}\mathbb{E}\left[\left\|z^k - z_*^k\right\|^2\right] \\
\leq&\frac{\alpha_k}{2}\mathbb{E}\left[\left\|h^k - \nabla\Phi(x^k)\right\|^2\right] + \alpha_k\mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Phi(x^k)\right\|^2\right] + \alpha_k^2\sigma_w^2 + \alpha_k^2\sigma_{g,2}^2\mathbb{E}\left[\left\|z^k - z_*^k\right\|^2\right]
\end{aligned}$$

where the second inequality uses (A.15). Taking summation and expectation on both sides of (A.19) and using (A.20), we obtain (A.16). □

A.1.1.3. *Dual Convergence.*

LEMMA A.1.0.9. *Suppose Assumptions 2 and 3 hold. In Algorithm 1 we have*

$$\begin{aligned}
&\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h^k-\nabla\Phi(x^k)\right\|^2\right]\\
\leq&\frac{1}{c_3}\mathbb{E}\left[\left\|h^0-\nabla\Phi(x^0)\right\|^2\right]+2\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\right\|^2\right]\\
&+\frac{2L_{\nabla\Phi}^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|x_+^k-x^k\right\|^2\right]+2c_3\sigma_{g,2}^2\sum_{k=0}^{K}\alpha_k^2\mathbb{E}\left[\left\|z^k-z_*^k\right\|^2\right]+\sum_{k=0}^{K}c_3\alpha_k^2\sigma_w^2. \qquad \text{(A.21)}
\end{aligned}$$

PROOF. Note that by moving-average update of h^k, we have

$$\begin{aligned}
&h^{k+1}-\nabla\Phi(x^{k+1})\\
=&(1-\theta_k)h^k+\theta_k(w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k])+\theta_k\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^{k+1})\\
=&(1-\theta_k)(h^k-\nabla\Phi(x^k))+\theta_k(\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k))+\nabla\Phi(x^k)-\nabla\Phi(x^{k+1})\\
&+\theta_k(w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k])
\end{aligned}$$

Hence we know

(A.22)

$$\begin{aligned}
&\mathbb{E}\left[\left\|h^{k+1}-\nabla\Phi(x^{k+1})\right\|^2|\mathscr{F}_k\right]\\
=&\left\|(1-\theta_k)(h^k-\nabla\Phi(x^k))+\theta_k(\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k))+\nabla\Phi(x^k)-\nabla\Phi(x^{k+1})\right\|^2\\
&+\theta_k^2\mathbb{E}\left[\left\|w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2|\mathscr{F}_k\right]\\
\leq&(1-\theta_k)\left\|h^k-\nabla\Phi(x^k)\right\|^2+\theta_k\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)+\frac{1}{\theta_k}(\nabla\Phi(x^k)-\nabla\Phi(x^{k+1}))\right\|^2+\theta_k^2\sigma_{w,k+1}^2\\
\leq&(1-\theta_k)\left\|h^k-\nabla\Phi(x^k)\right\|^2+2\theta_k\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\right\|^2\\
&+\frac{2}{\theta_k}\left\|\nabla\Phi(x^k)-\nabla\Phi(x^{k+1})\right\|^2+\theta_k^2\sigma_{w,k+1}^2\\
\leq&(1-\theta_k)\left\|h^k-\nabla\Phi(x^k)\right\|^2+2\theta_k\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Phi(x^k)\right\|^2+\frac{2\alpha_k^2L_{\nabla\Phi}^2}{\theta_k}\left\|x_+^k-x^k\right\|^2+\theta_k^2\sigma_{w,k+1}^2,
\end{aligned}$$

where the first equality uses the fact that x^k, h^k, x^{k+1}, are all $\mathscr{F}_k$-measurable and are independent of w^{k+1} given $\mathscr{F}_k$, the first inequality uses the convexity of $\|\cdot\|^2$ and (A.7), the second inequality

uses Cauchy-Schwarz inequality, the third inequality uses the Lipschitz continuity of $\nabla\Phi$ in Lemma A.1.0.10, and the update rules of x^{k+1}. Taking summation, expectation on both sides of (A.22), dividing c_3 and using (A.7), we know (A.21) holds. □

A.1.1.4. *Proof of Theorem 2.3.1.* Now we are ready to prove Theorem 2.3.1. From Lemma A.1.0.4 we know it suffices to bound V_k. By definition of V_k in (A.6), (A.16) and (A.21) we have

$$
\begin{aligned}
(\text{A.23}) \quad & \sum_{k=0}^{K}\alpha_k\mathbb{E}\left[V_k\right] = \sum_{k=0}^{K}\left(\frac{\alpha_k}{\tau^2}\mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + \alpha_k\mathbb{E}\left[\left\|h^k - \nabla\Phi(x^k)\right\|^2\right]\right)\\
\leq & \frac{2L_{\nabla\Phi}^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + \frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h^k - \nabla\Phi(x^k)\right\|^2\right]\\
& + 5\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|\nabla\Phi(x^k) - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2\right] + (1+2c_3)\sigma_{g,2}^2\sum_{k=0}^{K}\alpha_k^2\mathbb{E}\left[\left\|z^k - z_*^k\right\|^2\right]\\
& + \frac{2}{\tau}\mathbb{E}\left[W_{0,1}\right] + \frac{1}{c_3}\mathbb{E}\left[\left\|h^0 - \nabla\Phi(x^0)\right\|^2\right] + (1+c_3)\sigma_w^2\left(\sum_{k=0}^{K}\alpha_k^2\right),\\
\leq & \frac{2L_{\nabla\Phi}^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + \frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h^k - \nabla\Phi(x^k)\right\|^2\right]\\
& + 15\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left(L_{\nabla f}^2 + L_{\nabla^2 g}^2\right)\left\|y^k - y_*^k\right\|^2 + L_{\nabla g}^2\left\|z^k - z_*^k\right\|^2\right] + L_{\nabla g}^2\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|z^k - z_*^k\right\|^2\right]\\
& + \frac{2}{\tau}\mathbb{E}\left[W_{0,1}\right] + \frac{1}{c_3}\mathbb{E}\left[\left\|h^0 - \nabla\Phi(x^0)\right\|^2\right] + (1+c_3)\sigma_w^2\left(\sum_{k=0}^{K}\alpha_k^2\right)\\
\leq & C_{vx}\tau^2\sum_{k=0}^{K}\frac{\alpha_k}{\tau^2}\mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + C_{vh}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h^k - \nabla\Phi(x^k)\right\|^2\right] + C_{v,0} + C_{v,1}\left(\sum_{k=0}^{K}\alpha_k^2\right),
\end{aligned}
$$

where we assume

$$
(\text{A.24}) \qquad (1+2c_3)\sigma_{g,2}^2\alpha_k \leq L_{\nabla g}^2,
$$

in the second inequality. The constants are defined as

$$
\begin{aligned}
& C_{vx} = 15\left(L_{\nabla f}^2 + L_{\nabla^2 g}^2\right)C_{yx} + 16L_{\nabla g}^2 C_{zx} + \frac{2L_{\nabla\Phi}^2}{c_3^2},\ C_{vh} = \frac{1}{2},\\
& C_{v,0} = 15\left(L_{\nabla f}^2 + L_{\nabla^2 g}^2\right)C_{y,0} + 16L_{\nabla g}^2 C_{z,0} + \frac{2}{\tau}\mathbb{E}\left[W_{0,1}\right] + \frac{1}{c_3}\mathbb{E}\left[\left\|h^0 - \nabla\Phi(x^0)\right\|^2\right],\\
& C_{v,1} = 15\left(L_{\nabla f}^2 + L_{\nabla^2 g}^2\right)C_{y,1} + 16L_{\nabla g}^2 C_{z,1} + (1+c_3)\sigma_w^2.
\end{aligned}
$$

Using constants defined in Lemma A.1.0.6, we know

$$C_{vx} = \mathcal{O}\left(\frac{\kappa^8}{c_1^2} + \frac{\kappa^4}{c_2^2} + \frac{\kappa^6}{c_3^2}\right), C_{vh} = \mathcal{O}(1), C_{v,0} = \mathcal{O}\left(\frac{\kappa^5}{c_1} + \frac{\kappa^2}{c_2} + \frac{1}{\tau}\right), C_{v,1} = \mathcal{O}\big(c_1\kappa^5 + c_2\kappa^2\big).$$

Hence we can pick $\alpha_k \equiv \Theta(1/\sqrt{K})$, $\tau = \Theta\left(\kappa^{-4}\right)$, $c_1 = \mathcal{O}(1)$, $c_2 = \mathcal{O}(1)$, $c_3 = \mathcal{O}(1)$ so that the conditions ((A.9), (A.15) and (A.24)) in previous lemmas hold, and $\tau = \Theta(\kappa^{-4})$ such that

$$C_{vx}\tau^2 = \mathcal{O}(\kappa^8\tau^2) \leq \frac{1}{2}.$$

Plugging in all the constants in (A.23), we have

$$\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}[V_k] \leq \frac{1}{2K}\Big(\sum_{k=0}^{K}\frac{1}{\tau^2}\mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + \sum_{k=0}^{K}\mathbb{E}\left[\left\|h^k - \nabla\Phi(x^k)\right\|^2\right]\Big) + \mathcal{O}\big(\frac{\kappa^5}{\sqrt{K}}\big).$$

Then we have $\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\left[V_k\right] = \mathcal{O}(\kappa^5/\sqrt{K})$. which, together with Lemma A.1.0.4, proves Theorem 2.3.1.

A.1.2. Proof of Theorem 2.4.1. In this section we present our proof of Theorem 2.4.1. For simplicity, we summarize the notations that will be used in our proof as follows.

$$L_{\nabla f} = \max_{1\leq i\leq n} L_{\nabla f_i},\ L_{\nabla g} = \max_{1\leq i\leq n} L_{\nabla g_i},\ L_{\nabla^2 g_i} = \max_{1\leq i\leq n} L_{\nabla^2 g_i},\ \mu_g = \max_{1\leq i\leq n}\mu_{g_i},$$

$$\kappa = \max(L_{\nabla f}, L_{\nabla g})/\mu_g,\ u_x^{k+1} = \sum_{i=1}^{n} u_{x,i}^{k+1},\ w^{k+1} = \sum_{i=1}^{n}\lambda_i^k\big(u_{x,i}^{k+1} - J_i^{k+1}z_i^k\big),$$

$$\lambda_*^k = \lambda_*(x^k) = \arg\max_{\lambda\in\Delta_n}\Phi_{\mu_\lambda}(x^k,\lambda),\ y_{*,i}^k = y_i^*(x^k) = \arg\min_{y\in\mathbb{R}^{d_y}} g_i(x^k,y),$$

$$\Phi_i(x) = f_i(x, y_i^*(x)),\ \Phi^k = \Big(\Phi_1(x^k),...,\Phi_n(x^k)\Big)^\top, z_{*,i}^k = \big(\nabla_{22}^2 g_i(x^k, y_{*,i}^k)\big)^{-1}\nabla_2 f_i(x^k, y_{*,i}^k),$$

$$\Psi(x) = \max_{\lambda\in\Delta_n}\Phi_{\mu_\lambda}(x,\lambda) = \max_{\lambda\in\Delta_n}\left(\sum_{i=1}^{n}\lambda_i\Phi_i(x) - \frac{\mu_\lambda}{2}\left\|\lambda - \frac{\mathbf{1}_n}{n}\right\|^2\right),$$

$$\eta_X(x,h,\tau) = \min_{d\in X}\left\{\langle h, d-x\rangle + \frac{1}{2\tau}\left\|d-x\right\|^2\right\},\ \text{where } X = \mathcal{X} \text{ or } \Delta_n.$$

In this subsection we suppose Assumptions 2, 3 hold for all f_i, g_i and Assumption 4 holds. We suppose stepsizes in Algorithm 2 satisfy $\beta_k = c_1\alpha_k$, $\gamma_k = c_2\alpha_k, \theta_k = c_3\alpha_k$, where $c_1, c_2, c_3 > 0$ are constants to be determined. We will utilize the following merit function in our analysis:

$$\tilde{W}_k = \tilde{W}_{k,1} + \tilde{W}_{k,2}, \tilde{W}_{k,1} = \tilde{W}_{k,1}^{(1)} + \tilde{W}_{k,1}^{(2)}, \tilde{W}_{k,1}^{(1)} = \Psi(x^k) - \Phi_{\mu_\lambda}(x^k,\lambda^k) - \frac{1}{c_3}\eta_{\Delta_n}(\lambda^k, -h_\lambda^k, \tau_\lambda)$$

$$\tilde{W}_{k,1}^{(2)} = \Psi(x^k) - \inf_{x\in\mathcal{X}} \Psi(x) - \frac{1}{c_3}\eta_{\mathcal{X}}(x^k, h_x^k, \tau_x), \tilde{W}_{k,2} = \sum_{i=1}^{n} \left(\frac{1}{c_1}\left\|y_i^k - y_{*,i}^k\right\|^2 + \frac{1}{c_2}\left\|z_i^k - z_{*,i}^k\right\|^2\right).$$

By definition of $\Psi, \eta_{\mathcal{X}}, \eta_{\Delta_n}$, we can verify that $\tilde{W}_{k,1}^{(1)} \geq 0, \tilde{W}_{k,1}^{(2)} \geq 0$. Moreover, as discussed in Section 2.4.2, we consider the following optimality measure:

(A.25)
$$\tilde{V}_k = \underbrace{\frac{1}{\tau_x^2}\left\|x_+^k - x^k\right\|^2 + \left\|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k, \lambda^k)\right\|^2}_{\tilde{V}_{k,1}\text{: Optimality of min problem}} + \underbrace{\frac{1}{\tau_\lambda^2}\left\|\lambda_+^k - \lambda^k\right\|^2 + \left\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k, \lambda^k)\right\|^2}_{\tilde{V}_{k,2}\text{: Optimality of max problem}}.$$

The following lemma provides some smoothness of functions that we will use in our proof.

LEMMA A.1.0.10. *Functions* $\nabla\Psi(\cdot), \nabla_1\Phi_{\mu_\lambda}(\cdot,\lambda), \nabla_1\Phi(\cdot,\lambda), \nabla_1\Phi_{\mu_\lambda}(x,\cdot), \nabla_1\Phi(x,\cdot), \nabla_2\Phi_{\mu_\lambda}(\cdot,\lambda), \nabla_2\Phi_{\mu_\lambda}(x,\cdot)$ *are* $L_{\nabla\Psi}, L_{\nabla\Phi}, L_{\nabla\Phi}, L_{\nabla_1\Phi_{\mu_\lambda}}, L_{\nabla_1\Phi_{\mu_\lambda}}, L_{\nabla_2\Phi_{\mu_\lambda}}, \mu_\lambda$*-Lipschitz continuous respectively, with constants given by* $L_{\nabla\Psi} = \frac{n}{\mu_\lambda}\left(L_\Phi^2 + b_\Phi L_{\nabla\Phi}\right) + L_{\nabla\Phi}$, $L_{\nabla_1\Phi_{\mu_\lambda}} = L_{\nabla_2\Phi_{\mu_\lambda}} = \sqrt{n}L_\Phi$.

PROOF. For $\nabla\Psi$ we first notice that the nonconvex-strongly-concave problem in (2.10) can be reformulated as a bilevel problem:

$$\min_{x\in\mathcal{X}} \Psi(x) = \Phi_{\mu_\lambda}(x, \lambda^*(x)) \text{ s.t. } \lambda^*(x) = \arg\min_{\lambda\in\Delta_n} \left(-\Phi_{\mu_\lambda}(x,\lambda)\right) = \frac{\mu_\lambda}{2}\left\|\lambda - \frac{\mathbf{1}_n}{n}\right\|^2 - \sum_{i=1}^{n}\lambda_i\Phi_i(x).$$

By Lemma A.1.0.2 we know

$$\begin{aligned}
\nabla\Psi(x) =& \nabla_1\Phi_{\mu_\lambda}(x,\lambda^*(x)) - \nabla_{12}^2\Phi_{\mu_\lambda}(x,\lambda^*(x))\left(\nabla_{22}^2\Phi_{\mu_\lambda}(x,\lambda^*(x))\right)^{-1}\nabla_2\Phi_{\mu_\lambda}(x,\lambda^*(x)) \\
=& \sum_{i=1}^{n}\lambda_i^*(x)\nabla\Phi_i(x) + \frac{1}{\mu_\lambda}\left(\nabla\Phi_1(x), ..., \nabla\Phi_n(x)\right)\left[\begin{pmatrix}\Phi_1(x)\\ \vdots \\ \Phi_n(x)\end{pmatrix} - \mu_\lambda\left(\lambda^*(x) - \frac{\mathbf{1}_n}{n}\right)\right] \\
=& \frac{1}{\mu_\lambda}\sum_{i=1}^{n}\Phi_i(x)\nabla\Phi_i(x) + \frac{1}{n}\sum_{i=1}^{n}\nabla\Phi_i(x),
\end{aligned}$$

from which we know $\nabla\Psi(\cdot)$ is $L_{\nabla\Psi}$-Lipschitz continuous since

$$\begin{aligned}
&\|\Phi_i(x)\nabla\Phi_i(x) - \Phi_i(\tilde{x})\nabla\Phi_i(\tilde{x})\| \\
\leq& \|\Phi_i(x)\nabla\Phi_i(x) - \Phi_i(x)\nabla\Phi_i(\tilde{x})\| + \|\Phi_i(x)\nabla\Phi_i(\tilde{x}) - \Phi_i(\tilde{x})\nabla\Phi_i(\tilde{x})\| \\
\leq& \left(L_\Phi^2 + b_\Phi L_{\nabla\Phi}\right)\|x - y\|.
\end{aligned}$$

Note that for any fixed $\lambda \in \Delta_n$ and $x, \tilde{x} \in \mathcal{X}$, we have

$$\nabla_1 \Phi_{\mu_\lambda}(x, \lambda) = \nabla_1 \Phi(x, \lambda) = \sum_{i=1}^{n} \lambda_i \nabla \Phi_i(x), \tag{A.26}$$

$$\|\nabla_1 \Phi_{\mu_\lambda}(x, \lambda) - \nabla_1 \Phi_{\mu_\lambda}(\tilde{x}, \lambda)\| = \left\| \sum_{i=1}^{n} \lambda_i \left(\nabla \Phi_i(x) - \nabla \Phi_i(\tilde{x})\right) \right\| \leq L_{\nabla\Phi} \|x - \tilde{x}\|. \tag{A.27}$$

Similarly, for any fixed $x \in \mathcal{X}$ and $\lambda, \tilde{\lambda} \in \Delta_n$ we know

$$\left\|\nabla_1 \Phi_{\mu_\lambda}(x, \lambda) - \nabla_1 \Phi_{\mu_\lambda}(x, \tilde{\lambda})\right\| = \left\| \sum_{i=1}^{n} (\lambda_i - \tilde{\lambda}_i) \nabla \Phi_i(x) \right\| \leq \sqrt{n} L_\Phi \left\|\lambda - \tilde{\lambda}\right\|. \tag{A.28}$$

(A.26), (A.27) and (A.28) imply $\nabla_1 \Phi_{\mu_\lambda}(\cdot, \lambda), \nabla_1 \Phi(\cdot, \lambda)$ are $L_{\nabla\Phi}$-Lipschitz continuous and $\nabla_1 \Phi(x, \cdot)$, $\nabla_1 \Phi_{\mu_\lambda}(x, \cdot)$ are $L_{\nabla_1 \Phi_{\mu_\lambda}}$-Lipschitz continuous. Finally, for $\nabla_2 \Phi_{\mu_\lambda}(x, \lambda)$ we have $\nabla_2 \Phi_{\mu_\lambda}(x, \lambda)$ $= (\Phi_1(x), ..., \Phi_n(x))^\top - \mu_\lambda \left(\lambda - \frac{\mathbf{1}_n}{n}\right)$, and thus functions $\nabla_2 \Phi_{\mu_\lambda}(\cdot, \lambda), \nabla_2 \Phi_{\mu_\lambda}(x, \cdot)$ are $\sqrt{n} L_\Phi$, μ_λ-Lipschitz continuous respectively. □

Next we present a technical lemma that will be used in analyzing the strongly convex function over a closed convex set.

LEMMA A.1.0.11. *Suppose $f(x)$ is μ-strongly convex and L-smooth over a closed convex set $\mathcal{X}$. For any $\tau \leq \frac{1}{L}$ define $x_+ = \Pi_{\mathcal{X}}(x - \tau \nabla f(x))$ and $x_* = \arg\min_{x \in \mathcal{X}} f(x)$, we have $\left(1 - \sqrt{1 - \tau\mu}\right) \|x - x_*\| \leq \|x - x_+\|$.*

PROOF. By Corollary 2.2.4 in Nesterov [2018] we know

$$\begin{aligned} \frac{1}{\tau}\langle x - x_+, x - x_* \rangle &\geq \frac{1}{2\tau} \|x - x_+\|^2 + \frac{\mu}{2} \|x - x_*\|^2 + \frac{\mu}{2} \|x_+ - x_*\|^2 \\ &= \left(\frac{1}{2\tau} + \frac{\mu}{2}\right) \|x - x_+\|^2 + \mu \|x - x_*\|^2 - \mu \langle x - x_+, x - x_* \rangle \end{aligned}$$

which implies $\|x - x_+\| \|x - x_*\| \geq \langle x - x_+, x - x_* \rangle \geq \frac{1}{2} \|x - x_+\|^2 + r \|x - x_*\|^2$, where $r = \frac{\mu}{\frac{1}{\tau} + \mu} \leq \frac{1}{2}$. Applying Young's inequality to the left hand side of the above inequality, we know $\frac{1 + \sqrt{1-2r}}{4r} \|x - x_+\|^2 + \frac{r}{1 + \sqrt{1-2r}} \|x - x_*\|^2 \geq \frac{1}{2} \|x - x_+\|^2 + r \|x - x_*\|^2$, which gives $\|x - x_+\| \geq \left(1 - \sqrt{1 - 2r}\right) \|x - x_*\| \geq \left(1 - \sqrt{1 - \tau\mu}\right) \|x - x_*\|$. This completes the proof. □

The next lemma shows the relation between the stationarity used in Theorem 2.4.1 and our measure of optimality $\tilde{V}_k$ in (A.25).

LEMMA A.1.0.12. *Suppose Assumptions 2, 3 hold for all f_i, g_i and Assumption 4 holds. If $\tau_\lambda \mu_\lambda = 1$, then in Algorithm 2 we have*

$$\frac{1}{\tau_x^2}\left\|x^k - \Pi_{\mathcal{X}}\big(x^k - \tau_x \nabla_1 \Phi_{\mu_\lambda}(x^k, \lambda^k)\big)\right\|^2 \le 2\left(\frac{1}{\tau_x^2}\left\|x_+^k - x^k\right\|^2 + \left\|h_x^k - \nabla_1 \Phi_{\mu_\lambda}(x^k, \lambda^k)\right\|^2\right),$$
$$\left\|\lambda^k - \lambda_*^k\right\|^2 \le \frac{2}{\mu_\lambda^2}\left(\frac{1}{\tau_\lambda^2}\left\|\lambda_+^k - \lambda^k\right\|^2 + \left\|h_\lambda^k - \nabla_2 \Phi_{\mu_\lambda}(x^k, \lambda^k)\right\|^2\right),$$

which imply $\left\|\frac{1}{\tau_x}\big(x^k - \Pi_{\mathcal{X}}\big(x^k - \tau_x \nabla_1 \Phi_{\mu_\lambda}(x^k, \lambda^k)\big)\big)\right\|^2 + \left\|\lambda^k - \lambda_*^k\right\|^2 \le \max\left(2, \frac{2}{\mu_\lambda^2}\right) \tilde{V}_k$.

PROOF. The first inequality follows (A.1.0.4):

$$\begin{aligned}
&\left\|x^k - \Pi_{\mathcal{X}}\big(x^k - \tau_x \nabla_1 \Phi_{\mu_\lambda}(x^k, \lambda^k)\big)\right\|^2 \\
\le& 2\Big(\left\|x_+^k - x^k\right\|^2 + \left\|\Pi_{\mathcal{X}}\big(x^k - \tau_x h_x^k\big) - \Pi_{\mathcal{X}}\big(x^k - \tau_x \nabla_1 \Phi_{\mu_\lambda}(x^k, \lambda^k)\big)\right\|^2\Big) \\
\le& 2\Big(\left\|x_+^k - x^k\right\|^2 + \tau_x^2\left\|h_x^k - \nabla_1 \Phi_{\mu_\lambda}(x^k, \lambda^k)\right\|^2\Big),
\end{aligned}$$

where the first inequality uses Cauchy-Schwarz inequality and the second inequality uses the non-expansiveness of projection onto a closed convex set. Note $\lambda_*^k = \arg\min_{\lambda \in \Delta_n} \Phi_{\mu_\lambda}(x^k, \lambda)$ is a minimizer (over the probability simplex) of a μ_λ-smooth and μ_λ-strongly convex function $\Phi_{\mu_\lambda}(x^k, \cdot)$. Hence we know from Lemma A.1.0.11 that

$$\begin{aligned}
&\mu_\lambda^2\left\|\lambda_*^k - \lambda^k\right\|^2 \\
\le& \tau_\lambda^{-2}\big(1 + \sqrt{1 - \tau_\lambda \mu_\lambda}\big)^2\left\|\lambda^k - \Pi_{\Delta_n}\big(\lambda^k + \tau_\lambda \nabla_2 \Phi_{\mu_\lambda}(x^k, \lambda^k)\big)\right\|^2 \\
\le& 2\tau_\lambda^{-2}\big(1 + \sqrt{1 - \tau_\lambda \mu_\lambda}\big)^2\Big(\left\|\lambda_+^k - \lambda^k\right\|^2 + \left\|\Pi_{\Delta_n}(\lambda^k + \tau_\lambda h_\lambda^k) - \Pi_{\Delta_n}(\lambda^k + \tau_\lambda \nabla_2 \Phi_{\mu_\lambda}(x^k, \lambda^k))\right\|^2\Big) \\
\le& 2\tau_\lambda^{-2}\big(1 + \sqrt{1 - \tau_\lambda \mu_\lambda}\big)^2\Big(\left\|\lambda_+^k - \lambda^k\right\|^2 + \tau_\lambda^2\left\|h_\lambda^k - \nabla_2 \Phi_{\mu_\lambda}(x^k, \lambda^k)\right\|^2\Big),
\end{aligned}$$

where the second inequality uses Cauchy-Schwarz inequality and the third inequality uses non-expansiveness of the projection operator. Setting $\tau_\lambda \mu_\lambda = 1$ completes the proof. □

LEMMA A.1.0.13. *Suppose Assumptions 2, 3 hold for all f_i, g_i and Assumption 4 holds. In Algorithm 2 we have*

$$\mathbb{E}\left[\left\|w^{k+1} - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2\right] \le \sigma_{w,k+1}^2 \tag{A.29}$$

$$
\begin{aligned}
(\text{A.30}) \qquad & \mathbb{E}\left[\left\|h_x^{k+1}-h_x^k\right\|^2\right] \le \sigma_{h_x,k}^2, \quad \mathbb{E}\left[\left\|h_\lambda^{k+1}-h_\lambda^k\right\|^2\right] \le \sigma_{h_\lambda,k}^2, \\
& \sigma_{w,k+1}^2 := \sigma_w^2 + 2\sigma_{g,2}^2 \sum_{i=1}^n \mathbb{E}[\lambda_i^k \left\|z_i^k - z_{*,i}^k\right\|^2], \sigma_w^2 = \sigma_{f,1}^2 + \frac{2\sigma_{g,2}^2 L_f^2}{\mu_g^2} \\
& \sigma_{h_x,k}^2 := 2\theta_k^2 \mathbb{E}[\left\|h_x^k - \nabla_1 \Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2 + \left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1 \Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2] + \theta_k^2 \sigma_{w,k+1}^2 \\
& \sigma_{h_\lambda,k}^2 := \theta_k^2 \mathbb{E}\left[\left\|h_\lambda^k - \nabla_2 \Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right] + n\theta_k^2 \sigma_{f,0}^2.
\end{aligned}
$$

Proof. We first consider w^k. Note that

$$
w^{k+1} - \mathbb{E}[w^{k+1}|\mathscr{F}_k] = \sum_{i=1}^n \lambda_i^k \left(u_{x,i}^{k+1} - \mathbb{E}[u_{x,i}^{k+1}|\mathscr{F}_k] - \left(J_i^{k+1} - \mathbb{E}[J_i^{k+1}|\mathscr{F}_k]\right) z_i^k\right).
$$

Hence we know

$$
\begin{aligned}
& \mathbb{E}\left[\left\|w^{k+1} - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2 |\mathscr{F}_k\right] \\
=& \sum_{i=1}^n (\lambda_i^k)^2 \left(\mathbb{E}\left[\left\|u_{x,i}^k - \mathbb{E}[u_{x,i}^k|\mathscr{F}_k]\right\|^2 |\mathscr{F}_k\right] + \mathbb{E}\left[\left\|J_i^{k+1} - \mathbb{E}[J_i^{k+1}|\mathscr{F}_k]\right\|^2 |\mathscr{F}_k\right] \left\|z_i^k\right\|^2\right) \\
\le& \sum_{i=1}^n \lambda_i^k \left(\sigma_{f,1}^2 + 2\sigma_{g,2}^2 \left\|z_{*,i}^k\right\|^2 + 2\sigma_{g,2}^2 \left\|z_i^k - z_{*,i}^k\right\|^2\right) \\
\le& \sigma_{f,1}^2 + \frac{2\sigma_{g,2}^2 L_f^2}{\mu_g^2} + 2\sigma_{g,2}^2 \sum_{i=1}^n \lambda_i^k \left\|z_i^k - z_{*,i}^k\right\|^2.
\end{aligned}
$$

Taking expectation on both sides proves (A.29). Next for $\left\|h_x^{k+1} - h_x^k\right\|$ we have

$$
\begin{aligned}
& \mathbb{E}\left[\left\|h_x^{k+1} - h_x^k\right\|^2 |\mathscr{F}_k\right] = \theta_k^2 \mathbb{E}\left[\left\|h_x^k - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2 |\mathscr{F}_k\right] + \theta_k^2 \mathbb{E}\left[\left\|w^{k+1} - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2 |\mathscr{F}_k\right] \\
& \le 2\theta_k^2 \mathbb{E}\left[\left\|h_x^k - \nabla_1 \Phi(x^k,\lambda^k)\right\|^2 |\mathscr{F}_k\right] + 2\theta_k^2 \mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1 \Phi(x^k,\lambda^k)\right\|^2 |\mathscr{F}_k\right] + \theta_k^2 \sigma_{w,k+1}^2,
\end{aligned}
$$

which proves the first inequality of (A.30). Similarly we have

$$
\begin{aligned}
& \mathbb{E}\left[\left\|h_\lambda^{k+1} - h_\lambda^k\right\|^2 |\mathscr{F}_k\right] \\
=& \theta_k^2 \mathbb{E}\left[\left\|h_\lambda^k - \mathbb{E}[s^{k+1}|\mathscr{F}_k] + \mu_\lambda\big(\lambda^k - \frac{\mathbf{1}_n}{n}\big)\right\|^2 |\mathscr{F}_k\right] + \theta_k^2 \mathbb{E}\left[\left\|s^{k+1} - \mathbb{E}[s^{k+1}|\mathscr{F}_k]\right\|^2 |\mathscr{F}_k\right] \\
\le& \theta_k^2 \mathbb{E}\left[\left\|h_\lambda^k - \nabla_2 \Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2 |\mathscr{F}_k\right] + n\theta_k^2 \sigma_{f,0}^2,
\end{aligned}
$$

which proves the second inequality of (A.30). □

A.1.2.1. *Hypergradient Estimation Error.*

LEMMA A.1.0.14. *Suppose Assumptions 2, 3 hold for all f_i, g_i and Assumption 4 holds. In Algorithm 2 if the stepsizes satisfy*

$$\beta_k < \frac{2}{\mu_g + L_{\nabla g}}, \ \gamma_k \leq \min\left(\frac{1}{4\mu_g}, \ \frac{0.06\mu_g}{\sigma_{g,2}^2}\right), \tag{A.31}$$

then we have

$$\sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\sum_{i=1}^{n} \left\|y_i^k - y_{*,i}^k\right\|^2\right] \leq nC_{yx} \sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + \sum_{i=1}^{n} C_{y_i,0} + nC_{y,1}\left(\sum_{k=0}^{K} \alpha_k^2\right)$$
$$\sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\sum_{i=1}^{n} \left\|z_i^k - z_{*,i}^k\right\|^2\right] \leq nC_{zx} \sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + \sum_{i=1}^{n} C_{z_i,0} + nC_{z,1}\left(\sum_{k=0}^{K} \alpha_k^2\right)$$

where constants $C_{yx}, C_{y,1}, C_{zx}, C_{z,1}$ are defined in Lemma A.1.0.6. $C_{y_i,0}, C_{z_i,0}$ are defined as

$$C_{y_i,0} = \frac{1}{c_1\mu_g}\mathbb{E}\left[\left\|y_i^0 - y_{*,i}^0\right\|^2\right], \ C_{z_i,0} = \frac{5L_f^2}{\mu_g^2}\left(\frac{L_{\nabla_{22}^2 g}^2}{\mu_g^2} + 1\right)C_{y_i,0} + \frac{1}{c_2\mu_g}\mathbb{E}\left[\left\|z_i^0 - z_{*,i}^0\right\|^2\right].$$

PROOF. Note that the proof follows almost the same reasoning in Lemma A.1.0.6. Since Assumptions 2 and 3 hold for all f_i, g_i, by replacing y^k, y_*^k, z^k, z_*^k with $y_i^k, y_{*,i}^k, z_i^k, z_{*,i}^k$ respectively, we have similar results hold for each $1 \leq i \leq n$,

$$\begin{aligned}
&\sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\left\|y_i^k - y_{*,i}^k\right\|^2\right] \leq C_{yx} \sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + C_{y_i,0} + C_{y,1}\left(\sum_{k=0}^{K} \alpha_k^2\right), \\
&\sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\left\|z_i^k - z_{*,i}^k\right\|^2\right] \leq C_{zx} \sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + C_{z_i,0} + C_{z,1}\left(\sum_{k=0}^{K} \alpha_k^2\right).
\end{aligned} \tag{A.32}$$

Taking summation on both sides of (A.32), we complete the proof. □

The next lemma shows that the inequalities above will be used in the error analysis of $\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi(x^k, \lambda^k)\right\|$.

LEMMA A.1.0.15. *Suppose Assumptions 2, 3 hold for all f_i, g_i and Assumption 4 holds. We have*

$$\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi_{\mu_\lambda}(x^k, \lambda^k)\right\|^2 \leq \sum_{i=1}^{n} 3\lambda_i^k \left\{\left(L_{\nabla f}^2 + L_{\nabla^2 g}^2\right)\left\|y_i^k - y_{*,i}^k\right\|^2 + L_{\nabla g}^2\left\|z_i^k - z_{*,i}^k\right\|^2\right\},$$

$$\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Psi(x^k)\right\|^2 \le \sum_{i=1}^n 4\lambda_i^k\left\{\left(L_{\nabla f}^2+L_{\nabla^2 g}^2\right)\left\|y_i^k-y_{*,i}^k\right\|^2+L_{\nabla g}^2\left\|z_i^k-z_{*,i}^k\right\|^2\right\}$$
$$+8nL_\Phi^2\left\{\left\|\lambda_+^k-\lambda^k\right\|^2+\frac{1}{\mu_\lambda^2}\left\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right\}.$$

Proof. Note that we have the following decomposition:

$$\begin{aligned}
&\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\\
=&\mathbb{E}[u_x^{k+1}|\mathscr{F}_k]-\sum_{i=1}^n\lambda_i^k\nabla_1 f_i(x^k,y_{*,i}^k)-\sum_{i=1}^n\lambda_i^k\left(\mathbb{E}[J_i^{k+1}|\mathscr{F}_k]z_i^k-\nabla_{12}^2 g_i(x^k,y_{*,i}^k)z_{*,i}^k\right)\\
=&\sum_{i=1}^n\lambda_i^k\Big\{\nabla_1 f_i(x^k,y_i^k)-\nabla_1 f_i(x^k,y_{*,i}^k)-\nabla_{12}^2 g_i(x^k,y_i^k)\left(z_i^k-z_{*,i}^k\right)\\
&\qquad\qquad-\left[\nabla_{12}^2 g_i(x^k,y_i^k)-\nabla_{12}^2 g_i(x^k,y_{*,i}^k)\right]z_{*,i}^k\Big\}. \qquad \text{(A.33)}
\end{aligned}$$

which, together with Cauchy-Schwarz inequality, implies

$$\begin{aligned}
&\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\\
\le&3\left\|\sum_{i=1}^n\lambda_i^k\big(\nabla_1 f_i(x^k,y_i^k)-\nabla_1 f_i(x^k,y_{*,i}^k)\big)\right\|^2+3\left\|\sum_{i=1}^n\lambda_i^k\nabla_{12}^2 g_i(x^k,y_i^k)\big(z_i^k-z_{*,i}^k\big)\right\|^2\\
&+3\left\|\sum_{i=1}^n\big(\nabla_{12}^2 g_i(x^k,y_i^k)-\nabla_{12}^2 g_i(x^k,y_{*,i}^k)\big)z_{*,i}^k\right\|^2\\
\le&\sum_{i=1}^n 3\lambda_i^k\big(\left(L_{\nabla f}^2+L_{\nabla^2 g}^2\right)\left\|y_i^k-y_{*,i}^k\right\|^2+L_{\nabla g}^2\left\|z_i^k-z_{*,i}^k\right\|^2\big).
\end{aligned}$$

Similarly we have

$$\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Psi(x^k)=\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)+\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda_*^k).$$

Applying Cauchy-Schwarz inequality, Assumption 2 and Lemma A.1.0.10 to the above equation and (A.33), we know

$$\begin{aligned}
&\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Psi(x^k)\right\|^2\\
\le&4\left\|\sum_{i=1}^n\lambda_i^k\big(\nabla_1 f_i(x^k,y_i^k)-\nabla_1 f_i(x^k,y_{*,i}^k)\big)\right\|^2+4\left\|\sum_{i=1}^n\lambda_i^k\nabla_{12}^2 g_i(x^k,y_i^k)\big(z_i^k-z_{*,i}^k\big)\right\|^2
\end{aligned}$$

$$+4\left\|\sum_{i=1}^{n}\left(\nabla_{12}^2 g_i(x^k,y_i^k)-\nabla_{12}^2 g_i(x^k,y_{*,i}^k)\right)z_{*,i}^k\right\|^2+4\left\|\nabla_1\Phi(x^k,\lambda^k)-\nabla_1\Phi(x^k,\lambda_*^k)\right\|^2$$
$$\leq\sum_{i=1}^{n}4\lambda_i^k\left\{\left(L_{\nabla f}^2+L_{\nabla^2 g}^2\right)\left\|y_i^k-y_{*,i}^k\right\|^2+L_{\nabla g}^2\left\|z_i^k-z_{*,i}^k\right\|^2\right\}+4nL_\Phi^2\left\|\lambda^k-\lambda_*^k\right\|^2,$$

which together with Lemma A.1.0.12 completes the proof. □

A.1.2.2. *Primal Convergence.*

LEMMA A.1.0.16. *Suppose Assumptions 2, 3 hold for all f_i, g_i and Assumption 4 holds. If*

$$\alpha_k\leq\min\left(\frac{\tau_x^2}{20c_3},\ \frac{c_3}{2\tau_x\left(c_3L_{\nabla\Phi}+L_{\nabla\eta_\mathcal{X}}\right)},\ \frac{c_3}{4\tau_\lambda(L_{\nabla\eta_{\Delta_n}}+c_3\mu_\lambda)},\ \frac{n\tau_\lambda L_\Phi^2}{L_\Psi+L_{\nabla\Phi}},\ 1\right),$$
$$\tau_x<1,\ \tau_\lambda\mu_\lambda=1,\ c_3\leq\min\left(\frac{1}{10},\ \frac{1}{8(\mu_\lambda+1)^2}\right), \tag{A.34}$$

then in Algorithm 2 we have

$$\sum_{k=0}^{K}\frac{\alpha_k}{\tau_x^2}\mathbb{E}\left[\left\|x_+^k-x^k\right\|^2\right]$$
$$\leq\frac{2}{\tau_x}\mathbb{E}\left[\tilde{W}_{0,1}^{(1)}\right]+2\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla\Psi(x^k)\right\|^2\right]$$
$$+\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]+\frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h_x^k-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]$$
$$+\sigma_{g,2}^2\sum_{k=0}^{K}\alpha_k^2\mathbb{E}\left[\sum_{i=1}^{n}\lambda_i^k\left\|z_i^k-z_{*,i}^k\right\|^2\right]+\sigma_w^2\sum_{k=0}^{K}\alpha_k^2,$$
$$\sum_{k=0}^{K}\frac{\alpha_k}{\tau_\lambda^2}\mathbb{E}\left[\left\|\lambda_+^k-\lambda^k\right\|^2\right]$$
$$\leq\frac{2}{\tau_\lambda}\mathbb{E}\left[\tilde{W}_{0,1}^{(2)}\right]+\frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]+4L_f^2\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\sum_{i=1}^{n}\left\|y_i^k-y_{*,i}^k\right\|^2\right]$$
$$+13nL_\Phi^2\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|x_+^k-x^k\right\|^2\right]+n\sigma_{f,0}^2\sum_{k=0}^{K}\alpha_k^2. \tag{A.35}$$

PROOF. The proof of the first inequality in (A.35) is almost the same as that in (A.1.0.8). Note that by replacing $\Phi, h^k, W_{k,1}$ with $\Psi, h_x^k, \tilde{W}_{k,1}$, we know

$$\frac{\alpha_k}{\tau_x}\left\|x_+^k-x^k\right\|^2$$

$$
\begin{aligned}
&\leq \tilde{W}_{k,1}^{(1)} - \mathbb{E}[\tilde{W}_{k+1,1}^{(1)}|\mathscr{F}_k] + \alpha_k\big(\tau_x \left\|\nabla\Psi(x^k) - \mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2 + \frac{1}{4\tau_x}\left\|x_+^k - x^k\right\|^2\big) \\
&\quad + \frac{\alpha_k}{4\tau_x}\left\|x_+^k - x^k\right\|^2 + \frac{5}{2c_3\tau_x}\mathbb{E}\left[\left\|h_x^{k+1} - h_x^k\right\|^2 |\mathscr{F}_k\right], \qquad \text{(A.36)}
\end{aligned}
$$

Similar to (A.20), from (A.30) we have that

$$
\begin{aligned}
&\frac{5}{c_3\tau_x^2}\mathbb{E}\left[\left\|h_x^{k+1} - h_x^k\right\|^2\right] \\
\leq &\frac{10c_3\alpha_k^2}{\tau_x^2}\mathbb{E}\left[\left\|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2 + \left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right] + \frac{5c_3\alpha_k^2}{\tau_x^2}\sigma_w^2 \\
&+ \frac{10c_3\alpha_k^2\sigma_{g,2}^2}{\tau_x^2}\mathbb{E}\Big[\sum_{i=1}^n \lambda_i^k\left\|z_i^k - z_{*,i}^k\right\|^2\Big]. \\
\leq &\frac{\alpha_k}{2}\mathbb{E}\left[\left\|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right] + \alpha_k\mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right] + \alpha_k^2\sigma_w^2 \\
&+ \alpha_k^2\sigma_{g,2}^2\mathbb{E}\Big[\sum_{i=1}^n \lambda_i^k\left\|z_i^k - z_{*,i}^k\right\|^2\Big], \qquad \text{(A.37)}
\end{aligned}
$$

where the second inequality uses (A.34). Taking summation and expectation on both sides of (A.36) and using (A.37), we obtain the first inequality in (A.35). For the second inequality in (A.35), the $L_{\nabla\Psi}$-smoothness of $\Psi(x)$ and $L_{\nabla\eta_{\mathcal{X}}}$-smoothness of $\eta_{\mathcal{X}}$ in Lemma A.1.0.10 imply

$$
\begin{aligned}
&\Psi(x^{k+1}) - \Psi(x^k) \leq \alpha_k\langle\nabla\Psi(x^k), x_+^k - x^k\rangle + \frac{L_{\nabla\Psi}}{2}\left\|x^{k+1} - x^k\right\|^2, \qquad \text{(A.38)} \\
&\eta_{\Delta_n}(\lambda^k, -h_\lambda^k, \tau_\lambda) - \eta_{\Delta_n}(\lambda^{k+1}, -h_\lambda^{k+1}, \tau_\lambda) \\
\leq &\langle h_\lambda^k + \frac{1}{\tau_\lambda}(\lambda^k - \lambda_+^k), \lambda^k - \lambda^{k+1}\rangle + \langle\lambda_+^k - \lambda^k, -h_\lambda^k + h_\lambda^{k+1}\rangle \\
&+ \frac{L_{\nabla\eta_{\Delta_n}}}{2}\left(\left\|\lambda^{k+1} - \lambda^k\right\|^2 + \left\|-h_\lambda^{k+1} + h_\lambda^k\right\|^2\right) \\
= &\alpha_k\langle -h_\lambda^k, \lambda_+^k - \lambda^k\rangle + \frac{\alpha_k}{\tau_\lambda}\left\|\lambda_+^k - \lambda^k\right\|^2 + \theta_k\langle\lambda_+^k - \lambda^k, s^{k+1} - h_\lambda^k - \mu_\lambda(\lambda^k - \frac{\mathbf{1}_n}{n})\rangle \\
&+ \frac{L_{\nabla\eta_{\Delta_n}}}{2}\left(\left\|\lambda^{k+1} - \lambda^k\right\|^2 + \left\|h_\lambda^{k+1} - h_\lambda^k\right\|^2\right) \\
\leq &-\frac{\theta_k}{\tau_\lambda}\left\|\lambda_+^k - \lambda^k\right\|^2 + \theta_k\langle s^{k+1} - \mu_\lambda(\lambda^k - \frac{\mathbf{1}_n}{n}), \lambda_+^k - \lambda^k\rangle \\
&+ \frac{L_{\nabla\eta_{\Delta_n}}}{2}\left(\left\|\lambda^{k+1} - \lambda^k\right\|^2 + \left\|h_\lambda^{k+1} - h_\lambda^k\right\|^2\right). \qquad \text{(A.39)}
\end{aligned}
$$

We also have

$$
\begin{aligned}
&\Phi_{\mu_\lambda}(x^k,\lambda^k)-\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1})\\
=&\sum_{i=1}^n\left(\lambda_i^k\Phi_i(x^k)-\lambda_i^{k+1}\Phi_i(x^{k+1})\right)+\frac{\mu_\lambda}{2}\left\|\lambda^{k+1}-\frac{\mathbf{1}_n}{n}\right\|^2-\frac{\mu_\lambda}{2}\left\|\lambda^k-\frac{\mathbf{1}_n}{n}\right\|^2\\
=&\langle\lambda^k,\Phi^k\rangle-\langle\lambda^{k+1},\Phi^{k+1}\rangle+\frac{\mu_\lambda}{2}\left(\left\|\lambda^{k+1}-\lambda^k+\lambda^k-\frac{\mathbf{1}_n}{n}\right\|^2-\left\|\lambda^k-\frac{\mathbf{1}_n}{n}\right\|^2\right)\\
=&\langle\lambda^k-\lambda^{k+1},\Phi^k\rangle+\langle\lambda^{k+1},\Phi^k-\Phi^{k+1}\rangle+\mu_\lambda\alpha_k\langle\lambda^k-\frac{\mathbf{1}_n}{n},\lambda_+^k-\lambda^k\rangle+\frac{\mu_\lambda}{2}\left\|\lambda^{k+1}-\lambda^k\right\|^2\\
=&\alpha_k\langle\lambda^k-\lambda_+^k,\mathbb{E}[s^{k+1}|\mathscr{F}_k]-\mu_\lambda\big(\lambda^k-\frac{\mathbf{1}_n}{n}\big)\rangle+\alpha_k\langle\lambda^k-\lambda_+^k,\Phi^k-\mathbb{E}[s^{k+1}|\mathscr{F}_k]\rangle\\
&+\frac{\mu_\lambda}{2}\left\|\lambda^{k+1}-\lambda^k\right\|^2+\langle\lambda^{k+1},\Phi^k-\Phi^{k+1}\rangle\\
\le&\alpha_k\langle\lambda^k-\lambda_+^k,\mathbb{E}[s^{k+1}|\mathscr{F}_k]-\mu_\lambda\big(\lambda^k-\frac{\mathbf{1}_n}{n}\big)\rangle+\alpha_k\langle\lambda^k-\lambda_+^k,\Phi^k-\mathbb{E}[s^{k+1}|\mathscr{F}_k]\rangle\\
&+\frac{\mu_\lambda}{2}\left\|\lambda^{k+1}-\lambda^k\right\|^2-\alpha_k\langle\nabla_1\Phi(x^k,\lambda^k),x_+^k-x^k\rangle+\sqrt{n}L_\Phi\left\|\lambda^{k+1}-\lambda^k\right\|\left\|x_+^k-x^k\right\|\\
&+\frac{L_{\nabla\Phi}}{2}\left\|x^{k+1}-x^k\right\|^2. \qquad \text{(A.40)}
\end{aligned}
$$

where the inequality uses Lemma A.1.0.10 and (c) in Assumption 2 to obtain

$$
\begin{aligned}
&\langle\lambda^{k+1},\Phi^k-\Phi^{k+1}\rangle=\sum_{i=1}^n\lambda_i^{k+1}(\Phi_i(x^k)-\Phi_i(x^{k+1}))\\
\le&\sum_{i=1}^n\lambda_i^{k+1}\big(\langle\nabla\Phi_i(x^k),x^k-x^{k+1}\rangle+\frac{L_{\nabla\Phi}}{2}\left\|x^k-x^{k+1}\right\|^2\big)\\
\le&-\alpha_k\langle\nabla_1\Phi(x^k,\lambda^k),x_+^k-x^k\rangle+\sqrt{n}L_\Phi\left\|\lambda^{k+1}-\lambda^k\right\|\left\|x_+^k-x^k\right\|+\frac{L_{\nabla\Phi}}{2}\left\|x^{k+1}-x^k\right\|^2.
\end{aligned}
$$

Taking conditional expectation with respect to $\mathscr{F}_k$ on (A.38) + (A.39)/c_3 + (A.40), we know

$$
\begin{aligned}
&\frac{\alpha_k}{\tau_\lambda}\left\|\lambda_+^k-\lambda^k\right\|^2\\
\le&\tilde{W}_{k,1}^{(2)}-\mathbb{E}\left[\tilde{W}_{k+1,1}^{(2)}|\mathscr{F}_k\right]+\alpha_k\langle\nabla\Psi(x^k)-\nabla_1\Phi(x^k,\lambda^k),x_+^k-x^k\rangle\\
&+\alpha_k\langle\lambda^k-\lambda_+^k,\Phi^k-\mathbb{E}[s^{k+1}|\mathscr{F}_k]\rangle+\frac{(L_{\nabla\Psi}+L_{\nabla\Phi})}{2}\left\|x^{k+1}-x^k\right\|^2\\
&+\frac{(L_{\nabla\eta_{\Delta_n}}+c_3\mu_\lambda)}{2c_3}\left\|\lambda^{k+1}-\lambda^k\right\|^2+\sqrt{n}L_\Phi\left\|\lambda^{k+1}-\lambda^k\right\|\left\|x_+^k-x^k\right\|+\frac{L_{\nabla\eta_{\Delta_n}}}{2c_3}\mathbb{E}\left[\left\|h_\lambda^{k+1}-h_\lambda^k\right\|^2|\mathscr{F}_k\right]\\
\le&\tilde{W}_{k,1}^{(2)}-\mathbb{E}\left[\tilde{W}_{k+1,1}^{(2)}|\mathscr{F}_k\right]+\alpha_k\sqrt{n}L_\Phi\left\|\lambda^k-\lambda_*^k\right\|\left\|x_+^k-x^k\right\|
\end{aligned}
$$

$$
\begin{aligned}
&+\alpha_k L_f \left\|\lambda_+^k-\lambda^k\right\|\left(\sum_{i=1}^n\left\|y_i^k-y_{*,i}^k\right\|^2\right)^{\frac{1}{2}}+\alpha_k\sqrt{n}L_\Phi\left\|\lambda_+^k-\lambda^k\right\|\left\|x_+^k-x^k\right\|\\
&+\frac{\alpha_k^2(L_{\nabla\Psi}+L_{\nabla\Phi})}{2}\left\|x_+^k-x^k\right\|^2+\frac{\alpha_k^2(L_{\nabla\eta_{\Delta_n}}+c_3\mu_\lambda)}{2c_3}\left\|\lambda_+^k-\lambda^k\right\|^2+\frac{L_{\nabla\eta_{\Delta_n}}}{2c_3}\mathbb{E}\left[\left\|h_\lambda^{k+1}-h_\lambda^k\right\|^2|\mathscr{F}_k\right]\\
\leq&\tilde{W}_{k,1}^{(2)}-\mathbb{E}\left[\tilde{W}_{k+1,1}^{(2)}|\mathscr{F}_k\right]+\alpha_k\left(\frac{1}{16\tau_\lambda}\left\|\lambda^k-\lambda_*^k\right\|^2+4n\tau_\lambda L_\Phi^2\left\|x_+^k-x^k\right\|^2\right)\\
&+\alpha_k\left(\frac{1}{8\tau_\lambda}\left\|\lambda_+^k-\lambda^k\right\|^2+2\tau_\lambda L_f^2\sum_{i=1}^n\left\|y_i^k-y_{*,i}^k\right\|^2\right)+\alpha_k\left(\frac{1}{8\tau_\lambda}\left\|\lambda_+^k-\lambda^k\right\|^2+2n\tau_\lambda L_\Phi^2\left\|x_+^k-x^k\right\|^2\right)\\
&+\frac{\alpha_k n\tau_\lambda L_\Phi^2}{2}\left\|x_+^k-x^k\right\|^2+\frac{\alpha_k}{8\tau_\lambda}\left\|\lambda_+^k-\lambda^k\right\|^2+\frac{L_{\nabla\eta_{\Delta_n}}}{2c_3}\mathbb{E}\left[\left\|h_\lambda^{k+1}-h_\lambda^k\right\|^2|\mathscr{F}_k\right],
\end{aligned}
\tag{A.41}
$$

where the second inequality uses Lemma A.1.0.10, and the third inequality uses Young's inequality and the conditions on α_k (see (A.34)): $\frac{\alpha_k}{8\tau_\lambda}-\frac{\alpha_k^2(L_{\nabla\eta_{\Delta_n}}+c_3\mu_\lambda)}{2c_3}\geq 0$, $\alpha_k^2(L_{\nabla\Psi}+L_{\nabla\Phi})\leq\alpha_k n\tau_\lambda L_\Phi^2$. Recall that in Lemma A.1.0.12 we have

$$
\left\|\lambda^k-\lambda_*^k\right\|^2\leq 2\left\|\lambda_+^k-\lambda^k\right\|^2+\frac{2}{\mu_\lambda^2}\left\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2,
\tag{A.42}
$$

and by (A.30) we know

$$
\begin{aligned}
\frac{L_{\nabla\eta_{\Delta_n}}}{c_3\tau_\lambda}\mathbb{E}\left[\left\|h_\lambda^{k+1}-h_\lambda^k\right\|^2\right]&\leq 2c_3\alpha_k^2(\mu_\lambda+1)^2\left(\mathbb{E}\left[\left\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]+n\sigma_{f,0}^2\right)\\
&\leq\frac{\alpha_k}{4}\mathbb{E}\left[\left\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]+n\alpha_k^2\sigma_{f,0}^2.
\end{aligned}
\tag{A.43}
$$

where the second inequality uses $2c_3(\mu_\lambda+1)^2\leq\frac{1}{4}$, $\alpha_k\leq 1$ in (A.34). By (A.41), (A.42), and (A.43):

$$
\begin{aligned}
\frac{\alpha_k}{\tau_\lambda^2}\mathbb{E}\left[\left\|\lambda_+^k-\lambda^k\right\|^2\right]\leq&\frac{2}{\tau_\lambda}\mathbb{E}\left[\tilde{W}_{k,1}^{(2)}-\tilde{W}_{k+1,1}^{(2)}\right]+\frac{\alpha_k}{2}\mathbb{E}\left[\left\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]\\
&+4\alpha_k L_f^2\mathbb{E}\left[\sum_{i=1}^n\left\|y_i^k-y_{*,i}^k\right\|^2\right]+13\alpha_k nL_\Phi^2\mathbb{E}\left[\left\|x_+^k-x^k\right\|^2\right]+n\alpha_k^2\sigma_{f,0}^2,
\end{aligned}
$$

which implies the second inequality in (A.35) by taking summation. □

A.1.2.3. *Dual Convergence.*

LEMMA A.1.0.17. *Suppose Assumptions 2, 3 hold for all f_i, g_i and Assumption 4 holds. In Algorithm 2 we have*

$$
\sum_{k=0}^K\alpha_k\mathbb{E}\left[\left\|h_x^k-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]
$$

$$
\begin{aligned}
\leq & \frac{1}{c_3}\mathbb{E}\left[\left\|h_x^0-\nabla_1\Phi_{\mu_\lambda}(x^0,\lambda^0)\right\|^2\right]+3\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]\\
&+\frac{3L_{\nabla\Phi}^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|x_+^k-x^k\right\|^2\right]+\frac{3nL_\Phi^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|\lambda_+^k-\lambda^k\right\|^2\right]\\
&+2c_3\sigma_{g,2}^2\sum_{k=0}^{K}\alpha_k^2\mathbb{E}\Big[\sum_{i=1}^{n}\lambda_i^k\left\|z_i^k-z_{*,i}^k\right\|^2\Big]+c_3\sigma_w^2\sum_{k=0}^{K}\alpha_k^2,\\
&\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]\\
\leq & \frac{1}{c_3}\mathbb{E}\left[\left\|h_\lambda^0-\nabla_2\Phi_{\mu_\lambda}(x^0,\lambda^0)\right\|^2\right]+3\alpha_k L_f^2\sum_{i=1}^{n}\mathbb{E}\left[\left\|y_i^k-y_{*,i}^k\right\|^2\right]\\
&+\frac{3nL_\Phi^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|x_+^k-x^k\right\|^2\right]+\frac{3\mu_\lambda^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|\lambda_+^k-\lambda^k\right\|^2\right]+nc_3\sigma_{f,0}^2\sum_{k=0}^{K}\alpha_k^2. \qquad \text{(A.44)}
\end{aligned}
$$

Proof. The proof is similar to that of Lemma A.1.0.9, except that we now have another λ^k to handle. Since $\nabla_1\Phi(x,\lambda)=\nabla_1\Phi_{\mu_\lambda}(x,\lambda)$ for all (x,λ) (see (2.10)), for simplicity we omit subscript μ_λ in $\nabla_1\Phi_{\mu_\lambda}(x,\lambda)$ in proof. Note that by moving-average update of h_x^k, we have

$$
\begin{aligned}
&h_x^{k+1}-\nabla_1\Phi(x^{k+1},\lambda^{k+1})\\
=&(1-\theta_k)h_x^k+\theta_k(w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k])+\theta_k\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi(x^{k+1},\lambda^{k+1})\\
=&(1-\theta_k)(h_x^k-\nabla_1\Phi(x^k,\lambda^k))+\theta_k(\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi(x^k,\lambda^k))\\
&+\nabla_1\Phi(x^k,\lambda^k)-\nabla_1\Phi(x^{k+1},\lambda^{k+1})+\theta_k(w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k])
\end{aligned}
$$

Hence we know

$$
\begin{aligned}
&\mathbb{E}\left[\left\|h_x^{k+1}-\nabla_1\Phi(x^{k+1},\lambda^{k+1})\right\|^2|\mathscr{F}_k\right]\\
=&\Big\|(1-\theta_k)(h_x^k-\nabla_1\Phi(x^k,\lambda^k))+\theta_k(\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi(x^k,\lambda^k))\\
&+\nabla_1\Phi(x^k,\lambda^k)-\nabla_1\Phi(x^{k+1},\lambda^{k+1})\Big\|^2+\theta_k^2\mathbb{E}\left[\left\|w^{k+1}-\mathbb{E}[w^{k+1}|\mathscr{F}_k]\right\|^2|\mathscr{F}_k\right]\\
\leq&(1-\theta_k)\left\|h_x^k-\nabla_1\Phi(x^k,\lambda^k)\right\|^2+\theta_k^2\sigma_{w,k+1}^2\\
&+\theta_k\left\|(\mathbb{E}[w^{k+1}|\mathscr{F}_k]-\nabla_1\Phi(x^k,\lambda^k))+\frac{1}{\theta_k}(\nabla_1\Phi(x^k,\lambda^k)-\nabla_1\Phi(x^{k+1},\lambda^{k+1}))\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq (1-\theta_k)\left\|h_x^k - \nabla_1\Phi(x^k,\lambda^k)\right\|^2 + 3\theta_k\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi(x^k,\lambda^k)\right\|^2 + \theta_k^2\sigma_{w,k+1}^2 \\
&\quad + \frac{3}{\theta_k}\left\|\nabla_1\Phi(x^k,\lambda^k) - \nabla_1\Phi(x^{k+1},\lambda^k)\right\|^2 + \frac{3}{\theta_k}\left\|\nabla_1\Phi(x^{k+1},\lambda^k) - \nabla_1\Phi(x^{k+1},\lambda^{k+1})\right\|^2 \\
&\leq (1-\theta_k)\left\|h_x^k - \nabla_1\Phi(x^k,\lambda^k)\right\|^2 + 3\theta_k\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi(x^k,\lambda^k)\right\|^2 \\
&\quad + \frac{3\alpha_k^2}{\theta_k}\left(L_{\nabla\Phi}^2\left\|x_+^k - x^k\right\|^2 + nL_\Phi^2\left\|\lambda_+^k - \lambda^k\right\|^2\right) + \theta_k^2\sigma_{w,k+1}^2, \qquad \text{(A.45)}
\end{aligned}
$$

where the first equality uses the fact that $x^k, \lambda^k, h_x^k, x^{k+1}, \lambda^{k+1}$ are all $\mathscr{F}_k$-measurable and are independent of w^{k+1} given $\mathscr{F}_k$, the first inequality uses the convexity of $\|\cdot\|^2$ and (A.29), the second inequality uses Cauchy-Schwarz inequality, the third inequality uses the Lipschitz continuity of $\nabla_1\Phi$ in Lemma A.1.0.10, and the update rules of x^{k+1} and λ^{k+1}. Taking summation, expectation on both sides of (A.45), dividing c_3, and applying (A.29), we know the first inequality in (A.44) holds. Similarly we have

$$
\begin{aligned}
&h_\lambda^{k+1} - \nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1}) \\
=&(1-\theta_k)h_\lambda^k + \theta_k\big(s^{k+1} - \mu_\lambda\lambda^k + \mu_\lambda\frac{\mathbf{1}_n}{n}\big) - \nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1}) \\
=&(1-\theta_k)(h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)) + \theta\left(\mathbb{E}[s^{k+1}|\mathscr{F}_k] - \nabla_2\Phi(x^k,\lambda^k)\right) \\
&+ \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k) - \nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1}) + \theta_k(s^{k+1} - \mathbb{E}[s^{k+1}|\mathscr{F}_k]).
\end{aligned}
$$

where the second equality uses $\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k) = \nabla_2\Phi(x^k,\lambda^k) - \mu_\lambda\big(\lambda^k - \frac{\mathbf{1}_n}{n}\big)$. So we know

$$
\begin{aligned}
&\mathbb{E}\left[\left\|h_\lambda^{k+1} - \nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1})\right\|^2|\mathscr{F}_k\right] \\
=&\left\|(1-\theta_k)(h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)) + \theta\left(\mathbb{E}[s^{k+1}|\mathscr{F}_k] - \nabla_2\Phi(x^k,\lambda^k)\right)\right. \\
&\left.+ \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k) - \nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1})\right\|^2 + \theta_k^2\mathbb{E}\left[\left\|s^{k+1} - \mathbb{E}[s^{k+1}|\mathscr{F}_k]\right\|^2|\mathscr{F}_k\right] \\
\leq&(1-\theta_k)\left\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2 + n\theta_k^2\sigma_{f,0}^2 \\
&+ \theta_k\left\|\mathbb{E}[s^{k+1}|\mathscr{F}_k] - \nabla_2\Phi(x^k,\lambda^k) + \frac{1}{\theta_k}(\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k) - \nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1}))\right\|^2 \\
\leq&(1-\theta_k)\left\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2 + 3\theta_k\left\|\mathbb{E}[s^{k+1}|\mathscr{F}_k] - \nabla_2\Phi(x^k,\lambda^k)\right\|^2 + n\theta_k^2\sigma_{f,0}^2 \\
&+ \frac{3}{\theta_k}(\left\|\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k) - \nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^k)\right\|^2 + \left\|\nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^k) - \nabla_2\Phi_{\mu_\lambda}(x^{k+1},\lambda^{k+1})\right\|^2)
\end{aligned}
$$

$$\leq (1-\theta_k)\left\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2 + 3\theta_k L_f^2 \sum_{i=1}^n \left\|y_i^k - y_{*,i}^k\right\|^2$$

$$+\frac{3\alpha_k^2}{\theta_k}\left(nL_\Phi^2\left\|x_+^k - x^k\right\|^2 + \mu_\lambda^2\left\|\lambda_+^k - \lambda^k\right\|^2\right) + n\theta_k^2\sigma_{f,0}^2, \tag{A.46}$$

where the third inequality uses Lemma A.1.0.10 and the fact that

$$\mathbb{E}[s^{k+1}|\mathscr{F}_k] = \left(f_1(x^k,y_1^k),...,f_n(x^k,y_n^k)\right)^\top, \nabla_2\Phi(x^k,\lambda^k) = \left(f_1(x^k,y_{*,1}^k),...,f_n(x^k,y_{*,n}^k)\right)^\top$$

Taking summation, expectation on both sides of (A.46), and dividing c_3, we know the second inequality in (A.44) holds. □

A.1.2.4. *Proof of Theorem 2.4.1 and Corollary 2.4.1.* Now we are ready to present our main convergence results. Note that by Lemmas (A.1.0.16) and (A.1.0.17), for $\tilde{V}_{k,1}$ we have

$$\sum_{k=0}^K \alpha_k\mathbb{E}[\tilde{V}_{k,1}] = \sum_{k=0}^K \frac{\alpha_k}{\tau_x^2}\mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + \sum_{k=0}^K \alpha_k\mathbb{E}\left[\left\|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]$$

$$\leq \frac{3L_{\nabla\Phi}^2}{c_3^2}\sum_{k=0}^K \alpha_k\mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + \frac{1}{2}\sum_{k=0}^K \alpha_k\mathbb{E}\left[\left\|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]$$

$$+\frac{2}{\tau_x}\mathbb{E}\left[\tilde{W}_{0,1}^{(1)}\right] + \frac{1}{c_3}\mathbb{E}\left[\left\|h_x^0 - \nabla_1\Phi_{\mu_\lambda}(x^0,\lambda^0)\right\|^2\right] + 2\sum_{k=0}^K \alpha_k\mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Psi(x^k)\right\|^2\right]$$

$$+4\sum_{k=0}^K \alpha_k\mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right] + \frac{3nL_\Phi^2}{c_3^2}\sum_{k=0}^K \alpha_k\mathbb{E}\left[\left\|\lambda_+^k - \lambda^k\right\|^2\right]$$

$$+(1+2c_3)\,\sigma_{g,2}^2\sum_{k=0}^K \alpha_k^2\mathbb{E}\left[\sum_{i=1}^n \lambda_i^k\left\|z_i^k - z_{*,i}^k\right\|^2\right] + (1+c_3)\,\sigma_w^2\left(\sum_{k=0}^K \alpha_k^2\right). \tag{A.47}$$

By Lemma A.1.0.15 we know

$$4\sum_{k=0}^K \alpha_k\mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right] + 2\sum_{k=0}^K \alpha_k\mathbb{E}\left[\left\|\mathbb{E}[w^{k+1}|\mathscr{F}_k] - \nabla\Psi(x^k)\right\|^2\right]$$

$$\leq \sum_{k=0}^K \alpha_k\mathbb{E}\left[\sum_{i=1}^n 20\left(\left(L_{\nabla f}^2 + L_{\nabla^2 g}^2\right)\left\|y_i^k - y_{*,i}^k\right\|^2 + L_{\nabla g}^2\left\|z_i^k - z_{*,i}^k\right\|^2\right)\right]$$

$$+\sum_{k=0}^K 16nL_\Phi^2\alpha_k\mathbb{E}\left[\left\|\lambda_+^k - \lambda^k\right\|^2 + \frac{1}{\mu_\lambda^2}\left\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]. \tag{A.48}$$

Choosing

$$(1+2c_3)\,\sigma_{g,2}^2\alpha_k \le L_{\nabla g}^2 \tag{A.49}$$

in (A.47), and using (A.48), we know

$$\begin{aligned}\sum_{k=0}^{K}\alpha_k\mathbb{E}[\tilde{V}_{k,1}] \le & C_{v_1,x}\tau_x^2\sum_{k=0}^{K}\frac{\alpha_k}{\tau_x^2}\mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + C_{v_1,h_x}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h_x^k - \nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]\\ & + C_{v_1,\lambda}\tau_\lambda^2\sum_{k=0}^{K}\frac{\alpha_k}{\tau_\lambda^2}\mathbb{E}\left[\left\|\lambda_+^k - \lambda^k\right\|^2\right] + C_{v_1,h_\lambda}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]\\ & + C_{v_1,0} + C_{v_1,1}\left(\sum_{k=0}^{K}\alpha_k^2\right),\end{aligned} \tag{A.50}$$

where the constants are defined as

$$\begin{aligned}C_{v_1,x} =& 20n\big(L_{\nabla f}^2 + L_{\nabla^2 g}^2\big)C_{yx} + 21nL_{\nabla g}^2C_{zx} + \frac{3L_{\nabla\Phi}^2}{c_3^2},\ C_{v_1,h_x} = \frac{1}{2},\\ C_{v_1,\lambda} =& \left(16 + \frac{3}{c_3^2}\right)nL_\Phi^2,\ C_{v_1,h_\lambda} = \frac{16nL_\Phi^2}{\mu_\lambda^2},\\ C_{v_1,0} =& 20\big(L_{\nabla f}^2 + L_{\nabla^2 g}^2\big)\left(\sum_{i=1}^{n}C_{y_i,0}\right) + 21L_{\nabla g}^2\left(\sum_{i=1}^{n}C_{z_i,0}\right) + \frac{2}{\tau_x}\mathbb{E}\left[\tilde{W}_{0,1}^{(1)}\right]\\ & + \frac{1}{c_3}\mathbb{E}\left[\left\|h_x^0 - \nabla_1\Phi_{\mu_\lambda}(x^0,\lambda^0)\right\|^2\right],\\ C_{v_1,1} =& 20n\big(L_{\nabla f}^2 + L_{\nabla^2 g}^2\big)C_{y,1} + 21nL_{\nabla g}^2C_{z,1}.\end{aligned}$$

For $\tilde{V}_{k,2}$ we have

$$\begin{aligned}&\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,2}\big] = \sum_{k=0}^{K}\frac{\alpha_k}{\tau_\lambda^2}\mathbb{E}\left[\left\|\lambda_+^k - \lambda^k\right\|^2\right] + \sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]\\ \le& \frac{3\mu_\lambda^2}{c_3^2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|\lambda_+^k - \lambda^k\right\|^2\right] + \frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h_\lambda^k - \nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]\\ &+ \frac{2}{\tau_\lambda}\mathbb{E}\left[\tilde{W}_{0,1}^{(2)}\right] + \frac{1}{c_3}\mathbb{E}\left[\left\|h_\lambda^0 - \nabla_2\Phi_{\mu_\lambda}(x^0,\lambda^0)\right\|^2\right] + 7L_f^2\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\sum_{i=1}^{n}\left\|y_i^k - y_{*,i}^k\right\|^2\right]\\ &+ \left(13 + \frac{3}{c_3^2}\right)nL_\Phi^2\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|x_+^k - x^k\right\|^2\right] + n(1+c_3)\sigma_{f,0}^2\left(\sum_{k=0}^{K}\alpha_k^2\right),\end{aligned}$$

which implies

$$\begin{aligned}\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,2}\big] \leq & C_{v_2,x}\tau_x^2\sum_{k=0}^{K}\frac{\alpha_k}{\tau_x^2}\mathbb{E}\left[\left\|x_+^k-x^k\right\|^2\right]+C_{v_2,h_x}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h_x^k-\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]\\ &+C_{v_2,\lambda}\tau_\lambda^2\sum_{k=0}^{K}\frac{\alpha_k}{\tau_\lambda^2}\mathbb{E}\left[\left\|\lambda_+^k-\lambda^k\right\|^2\right]+C_{v_2,h_\lambda}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\left\|h_\lambda^k-\nabla_2\Phi_{\mu_\lambda}(x^k,\lambda^k)\right\|^2\right]\\ &+C_{v_2,0}+C_{v_2,1}\left(\sum_{k=0}^{K}\alpha_k^2\right)\end{aligned}\tag{A.51}$$

where the constants are defined as

$$\begin{aligned}C_{v_2,x}=&7nL_f^2C_{yx}+\left(13+\frac{3}{c_3^2}\right)nL_\Phi^2,\ C_{v_2,h_x}=0,\ C_{v_2,\lambda}=\frac{3\mu_\lambda^2}{c_3^2},\ C_{v_2,h_\lambda}=\frac{1}{2},\\ C_{v_2,0}=&7L_f^2\left(\sum_{i=1}^{n}C_{y_i,0}\right)+\frac{2}{\tau_\lambda}\mathbb{E}\left[\tilde{W}_{0,1}^{(2)}\right]+\frac{1}{c_3}\mathbb{E}\left[\left\|h_\lambda^0-\nabla_2\Phi_{\mu_\lambda}(x^0,\lambda^0)\right\|^2\right]\\ C_{v_2,1}=&7nL_f^2C_{y,1}+n(1+c_3)\sigma_{f,0}^2.\end{aligned}$$

According to the definition of the constants in Lemmas A.1.0.6 and A.1.0.14, we could obtain (for simplicity we omit the dependency on κ here)

$$\begin{aligned}&C_{v_1,x}=\mathcal{O}\left(\frac{n}{c_1^2}+\frac{n}{c_2^2}+\frac{1}{c_3^2}\right),C_{v_1,h_x}=\frac{1}{2}=\mathcal{O}\left(1\right),C_{v_1,\lambda}=\mathcal{O}\left(n+\frac{n}{c_3^2}\right),C_{v_1,h_\lambda}=\mathcal{O}\left(\frac{n}{\mu_\lambda^2}\right),\\ &C_{v_1,0}=\mathcal{O}\left(\frac{n}{c_1}+\frac{n}{c_2}+\frac{1}{c_3}+\frac{1}{\tau_x}+\frac{1}{c_3\tau_x}\right),\ C_{v_1,1}=\mathcal{O}\left(nc_1+nc_2\right),\\ &C_{v_2,x}=\mathcal{O}\left(\frac{n}{c_1^2}+n+\frac{n}{c_3^2}\right),\ C_{v_2,h_x}=0,\ C_{v_2,\lambda}=\mathcal{O}\left(\frac{1}{c_3^2}\right),\ C_{v_2,h_\lambda}=\frac{1}{2}=\mathcal{O}\left(1\right),\\ &C_{v_2,0}=\mathcal{O}\left(\frac{n}{c_1}+\frac{1}{c_3}\right),\ C_{v_2,1}=\mathcal{O}\left(nc_1+n+nc_3\right).\end{aligned}$$

Hence we can pick $\alpha_k, c_1, c_2, c_3, \tau_x, \tau_\lambda$ such that $\alpha_k\equiv\Theta(1/\sqrt{nK})$, $c_1=c_2=\sqrt{n}$, $c_3=\Theta(1)$, $\tau_x=\mathcal{O}\left(\mu_\lambda/n\right)$, $\tau_\lambda=1/\mu_\lambda$, which leads to $C_{v_1,x}\tau_x^2\leq\frac{1}{2}$, $C_{v_2,x}C_{v_1,\lambda}\tau_x^2\tau_\lambda^2\leq\frac{1}{8}$, $C_{v_2,\lambda}\tau_\lambda^2\leq\frac{1}{2}$, and the conditions ((A.31), (A.34), and (A.49)) in previous lemmas hold. Moreover, using the above conditions in (A.50) and (A.51), we can get

$$\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,1}\big]\leq\frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,1}\big]+C_{v_1,\lambda}\tau_\lambda^2\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,2}\big]+\mathcal{O}\left(n\right)$$

$$\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,2}\big]\leq\frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,2}\big]+C_{v_2,x}\tau_x^2\sum_{k=0}^{K}\alpha_k\mathbb{E}\big[\tilde{V}_{k,1}\big]+\mathcal{O}\left(\sqrt{n}\right).$$

Combining the above two inequalities, we have

$$\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\big[\tilde{V}_{k,1}\big]=\mathcal{O}\bigg(\frac{n^2}{\mu_\lambda^2\sqrt{K}}\bigg),\ \frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\big[\tilde{V}_{k,2}\big]=\mathcal{O}\bigg(\frac{n}{\sqrt{K}}\bigg),$$

which completes the proof of Theorem 2.4.1 since we have

$$\begin{aligned}
&\frac{1}{\tau_x^2}\mathbb{E}[\left\|x^k-\Pi_{\mathcal{X}}\big(x^k-\tau_x\nabla\Psi_{\mu_\lambda}(x^k)\big)\right\|^2]\\
\leq&\frac{2}{\tau_x^2}\mathbb{E}[\left\|x^k-\Pi_{\mathcal{X}}\big(x^k-\tau_x\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\big)\right\|^2]\\
&+\frac{2}{\tau_x^2}\mathbb{E}[\left\|\Pi_{\mathcal{X}}\big(x^k-\tau_x\nabla\Psi_{\mu_\lambda}(x^k)\big)-\Pi_{\mathcal{X}}\big(x^k-\tau_x\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\big)\right\|^2]\\
\leq&\frac{2}{\tau_x^2}\mathbb{E}[\left\|x^k-\Pi_{\mathcal{X}}\big(x^k-\tau_x\nabla_1\Phi_{\mu_\lambda}(x^k,\lambda^k)\big)\right\|^2]+2nL_\Phi^2\mathbb{E}[\left\|\lambda^k-\lambda_*^k\right\|^2]\leq 4\mathbb{E}[\tilde{V}_{k,1}]+\frac{4nL_\Phi^2}{\mu_\lambda^2}\mathbb{E}[\tilde{V}_{k,2}]
\end{aligned}$$

where the second inequality uses non-expansiveness of projection operator and $\sqrt{n}L_\Phi$-Lipschitz continuity of $\nabla_1\Phi_{\mu_\lambda}(x,\cdot)$ in Lemma A.1.0.10. Note that we have n^2 in the numerator since we explicitly write out the Lipschitz constant $L_{\nabla_1\Phi_{\mu_\lambda}}$.

To prove Corollary 2.4.1, we notice that by choosing $\mu_\lambda=\mathcal{O}(\sqrt{\epsilon})$, we have

$$\|\nabla\Phi_{\mu_\lambda}(x,\lambda)-\nabla\Phi(x,\lambda)\|^2\leq\mu_\lambda^2\left\|\lambda-\frac{\mathbf{1}_n}{n}\right\|^2\leq\mu_\lambda^2=\mathcal{O}(\epsilon),$$

and thus under the same setup of Theorem 2.4.1, we know from Section D.2 of Lin et al. [2020b] that any ϵ-stationary point of Problem (2.10) is an ϵ-stationary point of Problem (2.9). Hence the corresponding sample complexity is $\mathcal{O}(n^5\epsilon^{-4})$.

A.2. Discussions on the Prior Works Related to Chapter 2

A.2.1. Regarding Gu et al. [2023]. In this section, we discuss several issues in the current form of Gu et al. [2023], which introduces a **M**ulti-**O**bjective **R**obust **Bi**level **T**wo-timescale optimization algorithm (`MORBiT`).

The primary issue in the current analysis of `MORBiT` arises from the *ambiguity* and *inconsistency* regarding the *expectation and filtration*. As a consequence, the current form of the paper was unable to demonstrate $\mathbb{E}[\max_{i\in[n]}\|y_i^k-y_i^*(x^{(k-1)})\|^2]\leq\tilde{\mathcal{O}}(\sqrt{n}K^{-2/5})$ claimed in Theorem 1 (10b) of Gu

et al. [2023]. The subsequent arguments are incorrect. We discuss some mistakes made in Gu et al. [2023] as follows.

We start by looking at Lemma 8 (informal) and Lemma 14 (formal) in Gu et al. [2023] that characterize the upper bound of $\mathcal{L}^{(k+1)}-\mathcal{L}^{(k)}$ where $\mathcal{L}^{(k)}=\mathbb{E}[\sum_{i=1}^n \lambda_i^{(k)}\ell_i(x^{(k)})]$. Here, the function ℓ_i is the function $\Phi_i(x)$ in our notation. The paper incorrectly asserted that $\mathcal{L}^k=\sum_{i=1}^n \lambda_i^{(k)}\mathbb{E}[\ell_i(x^{(k)})]$. To see why, let $\mathscr{F}_k$ denote the sigma algebra generated by all iterates (x,y,λ) with superscripts not greater than k. It is important to note that both $\{\lambda_i^{(k)}\}$ and $x^{(k)}$ are random objectives given the filtration $\mathscr{F}_k$. The ambiguity lies in the lack of clarity regarding the randomness over which the expectation operation is performed. In fact, we can rewrite the claim of Lemma 14 in Gu et al. [2023] without hiding the randomness. Let $\mathcal{L}^{(k)}=\sum_{i=1}^n \lambda_i^{(k)}\ell_i(x^{(k)})$. Then, we have

$$\begin{aligned}\mathcal{L}^{(k+1)}-\mathcal{L}^{(k)} \leq & \mathcal{O}(\alpha) \underbrace{\left(\sum_{i=1}^n \lambda_i^k\left\|y_i^{k+1}-y_i^*(x^{(k)})\right\|\right)^2}_{\leq \max_{i\in[n]}\left\|y_i^{k+1}-y_i^*(x^{(k)})\right\|^2} \\ & -\frac{1}{\alpha}\left\|x^{k+1}-x^k\right\|^2+\mathcal{O}(\gamma n)+\mathcal{O}(\alpha)\left\|h_x^{(k)}-\mathbb{E}[h_x^{(k)}\mid \mathscr{F}_k]\right\|^2,\end{aligned} \tag{A.52}$$

where α,β,γ are step sizes for x, y, and λ respectively. We hide the dependency for constants in their assumptions for simplicity. In addition, we want to emphasize that, unlike our notation, $h_x^{(k)}$ and $h_\lambda^{(k)}$ are stochastic gradients at step k. Therefore, $h_x^{(k)}$ and $h_\lambda^{(k)}$ are random objects given $\mathscr{F}_k$. By taking expectations over all the randomness above, we can see that Lemma 14 in Gu et al. [2023] is incorrect because it writes in the form of $\max\mathbb{E}[\cdot]$ instead of $\mathbb{E}[\max(\cdot)]$. Therefore, the subsequent arguments regarding the convergence of x,y,λ are incorrect, at least in the current form.

Regardless of the error, one may be able to proceed with the proof by utilizing Eq.(A.52) since our ultimate goal is to demonstrate the convergence of $\mathbb{E}[\max_{i\in[n]}\|y_i^k-y_i^*(x^{(k-1)})\|^2]$. One possible direction is to utilize the basic recursive inequality of $\max_{i\in[n]}\|y_i^{k+1}-y_i^*(x^{(k)})\|^2$. Observe that for each $i\in[n]$, we can establish the following inequality similar to Lemma 13 in Gu et al. [2023] without hiding the randomness:

$$\begin{aligned}&\left\|y_i^{(k+1)}-y_i^*(x^{(k)})\right\|^2 \leq (1-\mathcal{O}(\mu_g\beta))\left\|y_i^{(k)}-y_i^*(x^{(k-1)})\right\|^2+\mathcal{O}\left(\frac{1}{\mu_g\beta}\right)\left\|x^k-x^{k-1}\right\|^2 \\ &+\mathcal{O}(\beta^2)\left\|h_{y,i}^{(k)}-\mathbb{E}[h_{y,i}^{(k)}|\mathscr{F}_k]\right\|^2+\mathcal{O}(\beta)\langle y_i^{(k)}-y_i^*(x^{(k-1)}),h_{y,i}^{(k)}-\mathbb{E}[h_{y,i}^{(k)}|\mathscr{F}_k]\rangle\end{aligned}$$

However, the order of taking the expectation over the randomness and the maximum over $i \in [n]$ adds complexity to the problem. The last inner-product term can only be zero when first taking expectation given $\mathscr{F}_k$. When applying Young's inequality to bound this term, it inevitably introduces terms such as $\mathcal{O}(\beta)\|h_{y,i}^{(k)} - \mathbb{E}[h_{y,i}^{(k)}|\mathscr{F}_k]\|^2$ or $\mathcal{O}(1)\|y_i^{(k)} - y_i^*(x^{(k-1)})\|^2$, which make it challenging to proceed further with the convergence analysis.

Finally, we remark about the choice of the stationarity condition used in Gu et al. [2023]. Although the algorithmic aspect in Gu et al. [2023] is motivated by Lin et al. [2020a], the notion of stationarity for λ in Gu et al. [2023] is different from Lin et al. [2020a]. Under the notion of stationarity in Lin et al. [2020a] (Definition 3.7) $\Phi_{1/2\ell}(\cdot)$ is the Moreau envelope of $\Phi(\cdot)$, which is defined after taking the max over y (i.e., λ in our notation) in Definition 3.5 in Lin et al. [2020a], and a point x is ϵ-stationarity when $\left\|\nabla\Phi_{1/2\ell}(x)\right\| \leq \epsilon$. It is unclear if (10a) and (10c) in Gu et al. [2023] will imply similar convergence results under the notion of stationarity in Definition 3.7 in Lin et al. [2020a].

A.2.2. Regarding Hu et al. [2022]. Hu et al. [2022] considered a multi-block min-max bilevel optimization, which shares similarity with Problem (2.10) we consider. However, we note that their Assumption 2.2 on the LL function g_i requires $\nabla_{22}^2 g_i(x, y; \zeta) \succeq \mu_g I$, and is much stronger than ours and that in Gu et al. [2023]. For example, for any $0 < \mu_g < L_g$ and

$$\nabla_{22}^2 g_i(x, y_i; \zeta) = \begin{pmatrix} 2L_g & 0 \\ 0 & 0 \end{pmatrix} \text{ or } \begin{pmatrix} 0 & 0 \\ 0 & 2\mu_g \end{pmatrix} \text{ with equal probability}$$

indicates that $\nabla_{22}^2 g_i(x, y_i) = \operatorname{diag}(L_g, \mu_g) \succeq \mu_g I$ can hold even if $\nabla_{22}^2 g_i(x, y_i; \zeta) \succeq \mu_g I$ does not hold for any ζ. Further, they do not characterize the dependence on μ_g in the final complexity. Hence we omit a detailed comparison with Hu et al. [2022].

A.3. Additional Experiments on Heterogeneous Data of Chapter 3

To introduce heterogeneity, we set r as the heterogeneity rate, and the data distribution of x_e in Section 3.4.1 on node i is $\mathcal{N}(0, i^2 \cdot r^2)$. In Figure A.1(a), A.1(b) and A.1(c) (and similarly for A.1(d), A.1(e), and A.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 heterogeneity rates.

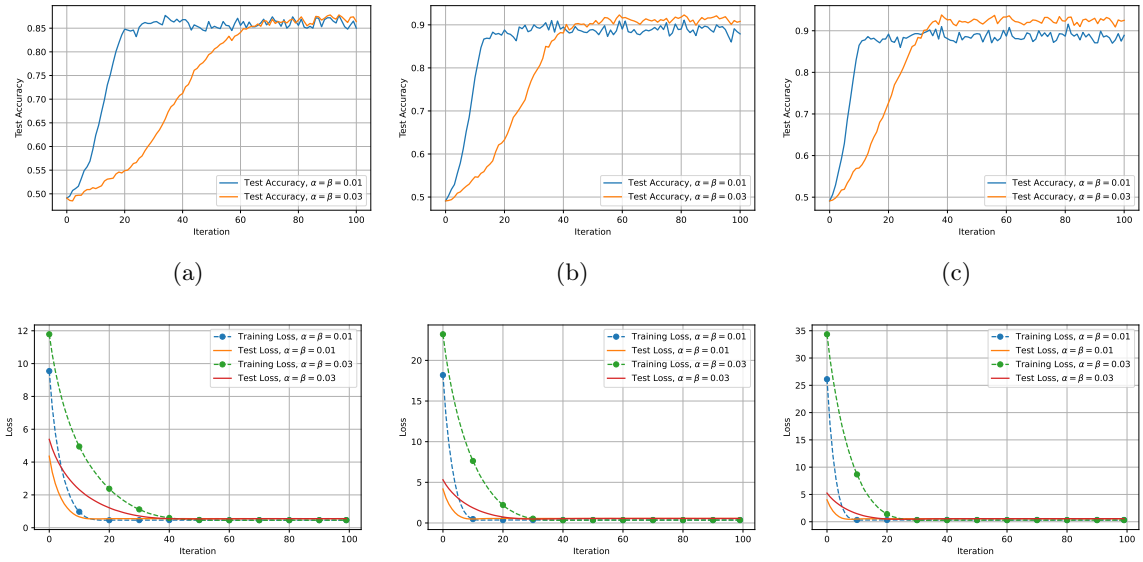


FIGURE A.1. ℓ^2-regularized logistic regression on synthetic data.

A.4. Proofs of Theorems in Chapter 3

Figure A.2 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 $X_k = (x_{1,k}, ..., x_{n,k})$, $Y_k^{(t)} = \left(y_{1,k}^{(t)}, ..., y_{n,k}^{(t)}\right)$.
- 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

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

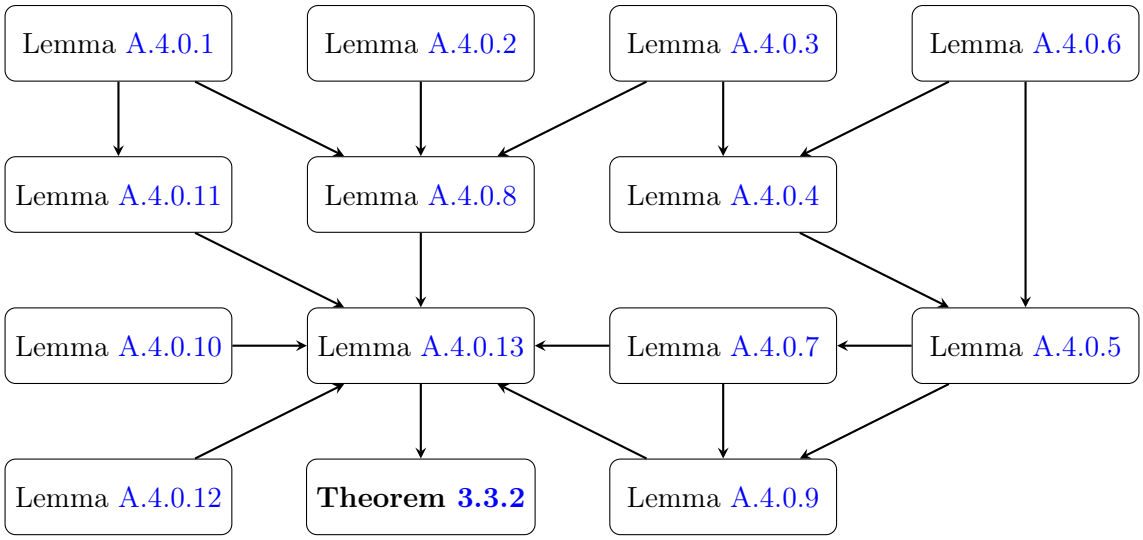


FIGURE A.2. Structure of the proof

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

LEMMA A.4.0.1. *Suppose Assumptions 5 and 8 hold, we know* $\nabla\Phi(x)$ *and* $y^*(x)$ *defined in* (3.2) *are* L_Φ *and* L_{y^*}*-Lipschitz continuous respectively with the constants given by*

$$L_\Phi = L_{f,1} + \frac{2L_{f,1}L_{g,1} + L_{g,2}L_{f,0}^2}{\mu_g} + \frac{2L_{g,1}L_{f,0}L_{g,2} + L_{g,1}^2 L_{f,1}}{\mu_g^2} + \frac{L_{g,2}L_{g,1}^2 L_{f,0}}{\mu_g^3}, \quad L_{y^*} = \frac{L_{g,1}}{\mu_g}. \tag{A.53}$$

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

LEMMA A.4.0.2. *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.4.0.2. 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.4.0.3. *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 3 is essentially decentralized stochastic gradient descent with gradient tracking on a strongly convex quadratic function.

LEMMA A.4.0.4. *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 6. 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

$$\tilde{\sigma}_1^2 = \sigma_1^2 + L^2,\ \tilde{\sigma}_2^2 = \sigma_2^2 + \max_i \|b_i\|^2,\ x^* := \left(\frac{1}{n}\sum_{i=1}^{n} A_i\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} b_i\right),$$

$$C_1 = 9\sigma_1^2 + 6\alpha^2\tilde{\sigma}_1^2 + \frac{18\alpha^2\sigma_1^2\tilde{\sigma}_1^2}{n},\ C_2 = 12\tilde{\sigma}_1^2 + 9\sigma_1^2 + 12\alpha^2 L^2\tilde{\sigma}_1^2 + \frac{18\alpha^2\sigma_1^2\tilde{\sigma}_1^2}{n},$$

$$C_3 = 6\rho^2\tilde{\sigma}_1^2,\ C_4 = 2\sigma_2^2 + \frac{6\alpha^2\sigma_2^2\tilde{\sigma}_1^2}{n} + \left(9\sigma_1^2 + \frac{18\alpha^2\sigma_1^2\tilde{\sigma}_1^2}{n}\right)\|x^*\|^2,$$

$$c = \left(\frac{\alpha^2}{n}(3\sigma_1^2\|x^*\|^2+\sigma_2^2),\ 0,\ \frac{(1+\rho^2)}{1-\rho^2}C_4\right)^\top,\ M = \begin{pmatrix} M_{11} & M_{12} & 0 \\ 0 & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33}\end{pmatrix},$$

$$M_{11} = 1-\alpha\mu,\ M_{12} = \left(\frac{2\alpha}{\mu}+2\alpha^2\right)\tilde{\sigma}_1^2,\ M_{22} = \frac{1+\rho^2}{2},\ M_{23} = \alpha^2\frac{1+\rho^2}{1-\rho^2}$$

$$M_{31} = \frac{1+\rho^2}{1-\rho^2}C_1,\ M_{32} = \frac{1+\rho^2}{1-\rho^2}C_2,\ M_{33} = \frac{1+\rho^2}{2}+\frac{1+\rho^2}{1-\rho^2}C_3\alpha^2.$$

If α satisfies

$$\left(1+\frac{\alpha\mu}{2}\right)(1-\alpha\mu)^2+\frac{3\alpha^2\sigma_1^2}{n} < 1-\alpha\mu,\ 0<\alpha_1\le\alpha\le\alpha_2 \text{ for some } 0<\alpha_1<\alpha_2,$$

$$\rho(M) < 1-\frac{2\alpha\mu}{3}, \quad \text{and } M \text{ has 3 different positive eigenvalues,} \tag{A.54}$$

then we have

$$\begin{aligned}
\mathbb{E}\left[\|\bar{x}_{k+1}-x^*\|^2\right] &\le (1-\alpha\mu)\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]+\left(\frac{2\alpha}{\mu}+2\alpha^2\right)\frac{\tilde{\sigma}_1^2}{n}\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right] \\
&\quad+\frac{\alpha^2}{n}(3\sigma_1^2\|x^*\|^2+\sigma_2^2), \\
\|X_{k+1}-\bar{x}_{k+1}\mathbf{1}^\top\|^2 &\le \frac{(1+\rho^2)}{2}\|X_k-\bar{x}_k\mathbf{1}^\top\|^2+\alpha^2\frac{1+\rho^2}{1-\rho^2}\|S_k-\bar{s}_k\mathbf{1}^\top\|^2, \\
\mathbb{E}\left[\frac{\|S_{k+1}-\bar{s}_{k+1}\mathbf{1}^\top\|^2}{n}\right] &\le \frac{1+\rho^2}{1-\rho^2}C_1\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]+\frac{1+\rho^2}{1-\rho^2}C_2\mathbb{E}\left[\frac{\|X_k-\bar{x}_k\mathbf{1}^\top\|^2}{n}\right] \\
&\quad+\left(\frac{1+\rho^2}{2}+\frac{1+\rho^2}{1-\rho^2}C_3\alpha^2\right)\mathbb{E}\left[\frac{\|S_k-\bar{s}_k\mathbf{1}^\top\|^2}{n}\right]+\frac{1+\rho^2}{1-\rho^2}C_4.
\end{aligned} \tag{A.55}$$

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

$$\max\left(\frac{1}{n}\mathbb{E}\left[\|X_k-x^*\mathbf{1}^\top\|^2\right],\frac{1}{n}\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right]\right)$$

(A.56)

$$\le 3C_M\left(1-\frac{2\alpha\mu}{3}\right)^k\left(\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right]+\mathbb{E}\left[\frac{\|X_0\|^2+\|S_0\|^2}{n}\right]\right)+\frac{5C_M\|c\|}{\alpha\mu},$$

(A.57)

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

Proof of Lemma A.4.0.4. Note that by definition of $\tilde{\sigma}_1^2$ and $\tilde{\sigma}_2^2$ we have

$$\mathbb{E}\left[\|A_{i,k}\|^2\right] = \mathbb{E}\left[\|A_{i,k} - A_i\|^2\right] + \|A_i\|_2^2 \leq \sigma_1^2 + L^2 = \tilde{\sigma}_1^2,$$

$$\mathbb{E}\left[\|b_{i,k}\|^2\right] = \mathbb{E}\left[\|b_{i,k} - b_i\|^2\right] + \|b_i\|^2 \leq \sigma_2^2 + \max_i \|b_i\|^2 = \tilde{\sigma}_2^2. \tag{A.58}$$

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

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

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

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

$$
\begin{aligned}
&+\frac{\alpha^2}{n}\left(\frac{3\sigma_1^2}{2}\mathbb{E}\left[\|\bar{x}_k\|^2\right]+\sigma_2^2\right)\\
&\leq\left[\left(1+\frac{\alpha\mu}{2}\right)(1-\alpha\mu)^2+\frac{3\alpha^2\sigma_1^2}{n}\right]\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]+\left(\frac{2\alpha}{\mu}+2\alpha^2\right)\frac{\tilde{\sigma}_1^2}{n}\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right]\\
&+\frac{\alpha^2}{n}(3\sigma_1^2\|x^*\|^2+\sigma_2^2)\\
&\leq(1-\alpha\mu)\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]+\left(\frac{2\alpha}{\mu}+2\alpha^2\right)\frac{\tilde{\sigma}_1^2}{n}\mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right]+\frac{\alpha^2}{n}(3\sigma_1^2\|x^*\|^2+\sigma_2^2).
\end{aligned}
$$

The first inequality holds because we have

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

the second inequality uses $\|\bar{x}_k\|^2\leq 2\|\bar{x}_k-x^*\|^2+2\|x^*\|^2$, and the third inequality uses (A.54). For $\|X_{k+1}-\bar{x}_{k+1}\mathbf{1}^\top\|^2$ we know

$$
\begin{aligned}
&\|X_{k+1}-\bar{x}_{k+1}\mathbf{1}^\top\|^2=\|X_kW-\bar{x}_k\mathbf{1}^\top-\alpha(S_k-\bar{s}_k\mathbf{1}^\top)\|^2\\
&\leq\left(1+\frac{1-\rho^2}{2\rho^2}\right)\rho^2\|X_k-\bar{x}_k\mathbf{1}^\top\|^2+\left(1+\frac{2\rho^2}{1-\rho^2}\right)\alpha^2\|S_k-\bar{s}_k\mathbf{1}^\top\|^2.
\end{aligned}\tag{A.59}
$$

The inequality uses Cauchy-Schwarz inequality and the fact that

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

where the last inequality uses Assumption 6. For $\|S_k-\bar{s}_k\mathbf{1}^\top\|^2$ we know

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

$$= \frac{1+\rho^2}{2}\|S_k - \bar{s}_k \mathbf{1}^\top\|^2 + \frac{1+\rho^2}{1-\rho^2}\|V_{k+1} - V_k\|^2. \tag{A.60}$$

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

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

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

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

where the second inequality holds since

$$\mathbb{E}\left[\|\bar{s}_k\|^2\right] = \mathbb{E}\left[\|\frac{1}{n}\sum_{i=1}^{n}(A_{i,k}x_{i,k} - b_{i,k})\|^2\right]$$

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

Hence we know

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

where the second inequality uses

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

The above inequalities and (A.60) imply

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

Now if we define

$$
\Gamma_k=\left(\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right],\ \mathbb{E}\left[\frac{\|X_k-\bar{x}_k\mathbf{1}^\top\|^2}{n}\right],\ \mathbb{E}\left[\frac{\|S_k-\bar{s}_k\mathbf{1}^\top\|^2}{n}\right]\right)^\top,
$$

$$c = \left(\frac{\alpha^2}{n}(3\sigma_1^2\|x^*\|^2+\sigma_2^2),\ 0,\ \frac{(1+\rho^2)}{1-\rho^2}C_4\right)^\top,\ M = \begin{pmatrix} M_{11} & M_{12} & 0 \\ 0 & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{pmatrix},$$

$$M_{11} = 1-\alpha\mu,\ M_{12} = \left(\frac{2\alpha}{\mu}+2\alpha^2\right)\tilde{\sigma}_1^2,\ M_{22} = \frac{1+\rho^2}{2},\ M_{23} = \alpha^2\frac{1+\rho^2}{1-\rho^2}$$

$$M_{31} = \frac{1+\rho^2}{1-\rho^2}C_1,\ M_{32} = \frac{1+\rho^2}{1-\rho^2}C_2,\ M_{33} = \frac{1+\rho^2}{2}+\frac{1+\rho^2}{1-\rho^2}C_3\alpha^2,$$

then by (A.55) 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 (A.54) we know there exists an invertible matrix $P \in \mathbb{R}^{3\times 3}$ 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

$$\|M^k\|_2 = \|P\cdot\mathrm{diag}(\lambda_1^k,\lambda_2^k,\lambda_3^k)P^{-1}\|_2 \le \left(1-\frac{2\alpha\mu}{3}\right)^k\|P\|_2\|P^{-1}\|_2 = C_M\cdot\left(1-\frac{2\alpha\mu}{3}\right)^k, \tag{A.61}$$

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

$$\begin{aligned}\|\sum_{i=0}^{k} M^i\|_2 &= \|\sum_{i=0}^{k} P\cdot\mathrm{diag}(\lambda_1^i,\lambda_2^i,\lambda_3^i)P^{-1}\|_2 = \|P\cdot\mathrm{diag}\left(\sum_{i=0}^{k}\lambda_1^i,\sum_{i=0}^{k}\lambda_2^i,\sum_{i=0}^{k}\lambda_3^i\right)P^{-1}\|_2 \\ &\le C_M\cdot\max_i\frac{1}{1-\lambda_i} < \frac{3C_M}{2\alpha\mu},\end{aligned} \tag{A.62}$$

where the last inequality uses the upper bound of the eigenvalues. For (A.56) we have

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

$$
\begin{aligned}
\leq & 2\sqrt{2}C_M\left(1-\frac{2\alpha\mu}{3}\right)^k\|\Gamma_0\|+2\sqrt{2}\cdot\frac{3C_M}{2\alpha\mu}\|c\| \\
\leq & 2\sqrt{2}C_M\left(1-\frac{2\alpha\mu}{3}\right)^k(\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right]+\mathbb{E}\left[\frac{\|X_0-\bar{x}_0\mathbf{1}^\top\|^2}{n}\right]+\mathbb{E}\left[\frac{\|S_0-\bar{s}_0\mathbf{1}^\top\|^2}{n}\right])+\frac{3\sqrt{2}C_M\|c\|}{\alpha\mu} \\
\leq & 3C_M\left(1-\frac{2\alpha\mu}{3}\right)^k\left(\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right]+\mathbb{E}\left[\frac{\|X_0\|^2+\|S_0\|^2}{n}\right]\right)+\frac{5C_M\|c\|}{\alpha\mu},
\end{aligned}
$$

where the fifth inequality uses (A.61) and (A.62), and the seventh inequality uses the fact that $\|X_0-\bar{x}_0\mathbf{1}^\top\|=\|X_0\left(I-\frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\|\leq\|X_0\|$ (same for S_0). (A.57) can be viewed as a corollary of the above inequality by noticing that

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

□

Remark:

- Lemma A.4.0.4 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 (A.57), 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 (A.56) vanishes (i.e., linear convergence). We will utilize this important observation in the next lemma.

LEMMA A.4.0.5. *Suppose Assumptions 5, 6 and 8 hold. In Algorithm 3 define*

$$
\begin{aligned}
C_1 &= 9\sigma_{g,2}^2+6\gamma^2(\sigma_{g,2}^2+L_{g,1}^2)+\frac{18\gamma^2\sigma_{g,2}^2(\sigma_{g,2}^2+L_{g,1}^2)}{n}, \\
C_2 &= 12(\sigma_{g,2}^2+L_{g,1}^2)+9\sigma_{g,2}^2+12\gamma^2L_{g,1}^2(\sigma_{g,2}^2+L_{g,1}^2)+\frac{18\gamma^2\sigma_{g,2}^2(\sigma_{g,2}^2+L_{g,1}^2)}{n}, \\
C_3 &= 6\rho^2(\sigma_{g,2}^2+L_{g,1}^2),\ C_4=2\sigma_f^2+\frac{6\gamma^2\sigma_f^2(\sigma_{g,2}^2+L_{g,1}^2)}{n}+(9\sigma_{g,2}^2+\frac{18\gamma^2\sigma_{g,2}^2(\sigma_{g,2}^2+L_{g,1}^2)}{n})\|x^*\|^2,
\end{aligned}
$$

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

$$M_{11} = 1-\gamma\mu_g,\ M_{12} = \left(\frac{2\gamma}{\mu_g}+2\gamma^2\right)(\sigma_{g,2}^2+L_{g,1}^2),\ M_{22} = \frac{1+\rho^2}{2},\ M_{23} = \gamma^2\frac{1+\rho^2}{1-\rho^2},$$

$$M_{31} = \frac{1+\rho^2}{1-\rho^2}C_1,\ M_{32} = \frac{1+\rho^2}{1-\rho^2}C_2,\ M_{33} = \frac{1+\rho^2}{2}+\frac{1+\rho^2}{1-\rho^2}C_3\gamma^2.$$

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*

$$\left(1+\frac{\gamma\mu_g}{2}\right)(1-\gamma\mu_g)^2+\frac{3\gamma^2\sigma_{g,2}^2}{n} < 1-\gamma\mu_g,\ 0<\gamma_1\le\gamma\le\gamma_2 \text{ for some } 0<\gamma_1<\gamma_2,$$

$$\max\left(\rho(\tilde{M}),\rho(M)\right) < 1-\frac{2\gamma\mu_g}{3},\ \text{both } M \text{ and } \tilde{M} \text{have 3 different positive eigenvalues,} \tag{A.63}$$

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

(A.64)

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

(A.65)

$$\frac{1}{n}\|\mathbb{E}\left[Z_t^{(k)}-\bar{z}_t^{(k)}\mathbf{1}^\top|\mathcal{F}_k\right]\|^2 \le 3C_{\tilde{M}}\left(1-\frac{2\gamma\mu_g}{3}\right)^t\left(\frac{L_{f,0}^2}{\mu_g^2}+L_{f,0}^2\right).$$

Proof of Lemma A.4.0.5. Note that (A.64) is a direct results of Lemma A.4.0.4 by noticing that

$$z_{i,t+1}^{(k)} = \sum_{j=1}^n w_{ij}z_{j,t}^{(k)}-\gamma d_{i,t}^{(k)},\ Z_0^{(k)}=0,$$

$$d_{i,t+1}^{(k)} = \sum_{i=1}^n w_{ij}d_{j,t}^{(k)}+s_{i,t+1}^{(k)}-s_{i,t}^{(k)},\ s_{i,t}^{(k)} = H_{i,t}^{(k)}z_{i,t}^{(k)}-b_{i,t}^{(k)},$$

$$\mathbb{E}\left[H_{i,t}^{(k)}|\mathcal{F}_k\right] = \nabla_y^2 g_i(x_{i,k},y_{i,k}^{(T)}),\ \mathbb{E}\left[\|H_{i,t}^{(k)}-\nabla_y^2 g_i(x_{i,k},y_{i,k}^{(T)})\|^2|\mathcal{F}_k\right]\le\sigma_{g,2}^2,$$

$$\mathbb{E}\left[b_{i,t}^{(k)}|\mathcal{F}_k\right] = \nabla_y f_i(x_{i,k},y_{i,k}^{(T)}),\ \mathbb{E}\left[\|b_{i,t}^{(k)}-\nabla_y f_i(x_{i,k},y_{i,k}^{(T)})\|^2|\mathcal{F}_k\right]\le\sigma_f^2,$$

for any $k\ge 0, 1\le i\le n$, and $t\ge 0$. Assumption 5 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 (A.57),

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

which proves (A.64). To prove (A.65), we notice that in expectation, the updates can be written as

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

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.4.0.4. Using this observation, (A.56), and the definition of $\|c\|$ and $\tilde{M}$, we know (A.65) holds. □

Now we provide a technical lemma that guarantees (A.54) and (A.63). For simplicity we can just consider (A.54).

LEMMA A.4.0.6. *Let M be the matrix defined in Lemma A.4.0.4. There exist $0<\alpha_1<\alpha_2$ such that $\alpha\in(\alpha_1,\alpha_2)$ and*

$$\left(1+\frac{\alpha\mu}{2}\right)(1-\alpha\mu)^2+\frac{3\alpha^2\sigma_1^2}{n}<1-\alpha\mu, \tag{A.66}$$

$$\rho(M)<1-\frac{2\alpha\mu}{3},\quad \textit{and}\quad M\ \textit{has 3 different positive eigenvalues.} \tag{A.67}$$

PROOF OF LEMMA A.4.0.6. Note that (A.66) is equivalent to

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

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

$$0<\alpha_1<\alpha_2<\frac{\sqrt{9\sigma_1^4+n^2\mu^4}-3\sigma_1^2}{n\mu^3} \tag{A.68}$$

will ensure (A.66). Next we consider (A.67). 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 (A.67) is

$$\varphi\left(1 - \frac{2\alpha\mu}{3}\right) > 0,\ \varphi(M_{11}) < 0,\ \varphi(M_{22}) > 0,\ \varphi(0) < 0,\ M_{11} > M_{22}, \tag{A.69}$$

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

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

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 (A.69) we have

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

Hence a sufficient condition for (A.69) 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 $1 \le i \le 3$. (A.68). The above expression implies (A.68) and (A.69), and hence (A.66) and (A.67) 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

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

  and thus the existence of γ_1 and γ_2 in (A.63) is also guaranteed.

Using Lemma A.4.0.5 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.4.0.7. *Suppose Assumptions 5, 6, 7, and 8 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

$$\frac{(1+\rho^2)}{2}+\beta_k^2\frac{1+\rho^2}{1-\rho^2}\cdot 6L_{g,1}^2\leq\frac{3+\rho^2}{4}<1, \tag{A.70}$$

then in Algorithm 5, for any $k\geq 0$ and $0\leq t\leq T-1$ we have

$$\begin{aligned}&\mathbb{E}\left[\|U_k\|^2\right]\leq n\sigma_u^2,\ \mathbb{E}\left[\|X_k-\bar{x}_k\mathbf{1}^\top\|^2\right]\leq n\sigma_x^2\tilde{\alpha}_k^2,\\ &\frac{1}{n}\mathbb{E}\left[\|Y_k^{(t)}-\bar{y}_k^{(t)}\mathbf{1}^\top\|^2\right]\leq\left[\left(\frac{3+\rho^2}{4}\right)^t T-t\left(\frac{3+\rho^2}{4}\right)\right]\tilde{\beta}_k^2+t\tilde{\beta}_{k+1}^2.\end{aligned} \tag{A.71}$$

Proof of Lemma A.4.0.7. Note that the inner and outer loop updates satisfy

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

which gives

$$\|X_{k+1}-\bar{x}_{k+1}\mathbf{1}^\top\|^2\leq\frac{(1+\rho^2)}{2}\|X_k-\bar{x}_k\mathbf{1}^\top\|^2+\alpha_k^2\frac{1+\rho^2}{1-\rho^2}\|R_k-\bar{r}_k\mathbf{1}^\top\|^2, \tag{A.72}$$

$$\|Y_k^{(t+1)}-\bar{y}_k^{(t+1)}\mathbf{1}^\top\|^2\leq\frac{(1+\rho^2)}{2}\|Y_k^{(t)}-\bar{y}_k^{(t)}\mathbf{1}^\top\|^2+\beta_k^2\frac{1+\rho^2}{1-\rho^2}\|V_k^{(t)}-\bar{v}_k^{(t)}\mathbf{1}^\top\|^2. \tag{A.73}$$

The inequalities hold similarly as the inequality in (A.59). Notice that we have

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

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

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

The fourth inequality uses (A.64). Using the above two inequaities in (A.72) 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.4.0.2 and $X_0 = 0$, we can obtain the first two results of (A.71). 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 (A.71). The above inequality together with (A.73) imply

(A.74)

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

where the second inequality uses Lemma A.4.0.2 and (A.70). 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.4.0.2. Using the above estimation in (A.74), we know

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

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

Now we are ready to analyze the convergence of the inner loop of Algorithm 5.

LEMMA A.4.0.8. *Suppose Assumptions 5 and 8 hold. For any $0 \leq t \leq T-1$ define*

(A.75)
$$C_{k,t+1} = \sum_{l=0}^{t}\left[\left(\frac{\beta_k}{\mu_g} + \beta_k^2\right)L_{g,1}^2\left(\sigma_x^2\tilde{\alpha}_k^2 + \left[\left(\frac{3+\rho^2}{4}\right)^l T - l\left(\frac{3+\rho^2}{4}\right)\right]\tilde{\beta}_k^2 + l\tilde{\beta}_{k+1}^2\right) + \frac{\beta_k^2\sigma_{g,1}^2}{n}\right].$$

If $T \geq 1$ and $0 < \beta_k \leq \min\{1, \frac{1}{\mu_g}\}$, then in Algorithm 5, we have

(A.76)
$$\frac{\mu_g}{2}\sum_{k=1}^{K}\beta_k\mathbb{E}\left[\|\bar{y}_k^{(0)} - y_{k-1}^*\|^2\right] \leq \mathbb{E}\left[\|\bar{y}_1^{(0)} - y_0^*\|^2\right] + L_{y^*}^2\sum_{k=1}^{K}\left(\frac{2\alpha_{k-1}^2}{\beta_k\mu_g} + \alpha_{k-1}^2\right)\mathbb{E}\left[\|\bar{r}_{k-1}\|^2\right] + \sum_{k=1}^{K}C_{k,T},$$

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

PROOF OF LEMMA A.4.0.8. For any $k \geq 0,\ 1 \leq t \leq T-1$, define

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

We know

(A.77)

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

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.4.0.3 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 (A.71) we know

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

where the second inequality uses Lemma A.4.0.2. Observe that we also have

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

where the third inequality holds since $(1+\frac{a}{2})(1-a)^T \leq (1-\frac{a}{2})$ for any $a > 0$ and $T \geq 1$, and $y^*(x)$ is L_{y^*}-smooth. This implies

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

Taking summation on both sides, we have

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

□

Lemma A.4.0.9. *Suppose Assumptions 5, 6, 7, and 8 hold. In Algorithm 3 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 (A.63), then we have

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbb{E}\left[\bar{z}_t^{(k)}|\mathcal{F}_k\right] - z_*^{(k)}\|^2\right] \\
\leq &(1-\gamma\mu_g)^N \cdot \frac{L_{f,0}^2}{\mu_g^2} + 5\left(\frac{1}{\mu_g^2} + \frac{\gamma}{\mu_g}\right)\left(L_{g,2}^2\sigma_z^2 + L_{f,1}^2\right)\left(\mathbb{E}\left[\|\bar{y}_{k+1}^{(0)} - y_k^*\|^2\right] + \sigma_x^2\tilde{\alpha}_k^2 + T\tilde{\beta}_{k+1}^2\right) \\
&+90C_{\tilde{M}}L_{g,1}^2\left(\frac{1}{\mu_g^2} + \frac{\gamma}{\mu_g}\right)\left(\frac{L_{f,0}^2}{\mu_g^2} + L_{f,0}^2\right)\left(1-\frac{2\gamma\mu_g}{3}\right)^{N-1}. \qquad \text{(A.78)}
\end{aligned}$$

Proof of Lemma A.4.0.9. 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

$$\begin{aligned}
&\|\dot{z}_{t+1,k} - z_*^{(k)}\|^2\\
=&\|\dot{z}_{t,k} - \gamma\left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right) - z_*^{(k)} - \gamma\left(\dot{s}_{t,k} - \left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right)\right)\|^2\\
\leq&(1+\gamma\mu_g)\|\dot{z}_{t,k} - \gamma\left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right) - z_*^{(k)}\|^2 + (1+\frac{1}{\gamma\mu_g})\gamma^2\|\dot{s}_{t,k} - \left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right)\|^2\\
\leq&(1+\gamma\mu_g)(1-\gamma\mu_g)^2\|\dot{z}_{t,k} - z_*^{(k)}\|^2 + \left(\frac{\gamma}{\mu_g} + \gamma^2\right)\|\dot{s}_{t,k} - \left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right)\|^2\\
\leq&(1-\gamma\mu_g)\|\dot{z}_{t,k} - z_*^{(k)}\|^2 + \left(\frac{\gamma}{\mu_g} + \gamma^2\right)\|\dot{s}_{t,k} - \left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right)\|^2, \qquad \text{(A.79)}
\end{aligned}$$

where the second inequality uses Lemma A.4.0.3. For $\dot{s}_{t,k} - \left(H^{(k)}\dot{z}_{t,k} - b^{(k)}\right)$ we have

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

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

The above inequality and (A.79) imply

$$
\begin{aligned}
\text{(A.80)}\ \ & \mathbb{E}\left[\|\dot{z}_{N,k} - z_*^{(k)}\|^2\right] \\
\leq & (1-\gamma\mu_g)\mathbb{E}\left[\|\dot{z}_{N-1,k} - z_*^{(k)}\|^2\right] + \left(\frac{\gamma}{\mu_g} + \gamma^2\right)\mathbb{E}\left[\|\dot{s}_{N-1,k} - \left(H^{(k)}\dot{z}_{N-1,k} - b^{(k)}\right)\|^2\right] \\
\leq & (1-\gamma\mu_g)^N\mathbb{E}\left[\|z_*^{(k)}\|^2\right] + \frac{5L_{g,1}^2\left(\frac{\gamma}{\mu_g} + \gamma^2\right)}{n}\sum_{t=0}^{N-1}(1-\gamma\mu_g)^{N-1-t}\mathbb{E}\left[\|\mathbb{E}\left[Z_t^{(k)} - \bar{z}_t^{(k)}\mathbf{1}^\top|\mathcal{F}_k\right]\|^2\right] \\
& + \frac{\frac{\gamma}{\mu_g} + \gamma^2}{1-(1-\gamma\mu_g)} \cdot \frac{5\left(L_{g,2}^2\sigma_z^2 + L_{f,1}^2\right)}{n}\mathbb{E}\left[n\|\bar{y}_k^{(T)} - y_k^*\|^2 + \|X_k - \bar{x}_k\mathbf{1}^\top\|^2 + \|Y_k^{(T)} - \bar{y}_k^{(T)}\mathbf{1}^\top\|^2\right] \\
\leq & (1-\gamma\mu_g)^N \cdot \frac{L_{f,0}^2}{\mu_g^2} + 5\left(\frac{1}{\mu_g^2} + \frac{\gamma}{\mu_g}\right)\left(L_{g,2}^2\sigma_z^2 + L_{f,1}^2\right)\left(\mathbb{E}\left[\|\bar{y}_k^{(T)} - y_k^*\|^2\right] + \sigma_x^2\tilde{\alpha}_k^2 + T\tilde{\beta}_{k+1}^2\right) \\
& + 5L_{g,1}^2\left(\frac{\gamma}{\mu_g} + \gamma^2\right)\sum_{t=0}^{N-1}\left(1 - \frac{\gamma\mu_g}{2}\right)^{N-1-t}\left(3C_{\tilde{M}}\left(1 - \frac{2\gamma\mu_g}{3}\right)^t\left(\frac{L_{f,0}^2}{\mu_g^2} + L_{f,0}^2\right)\right) \\
\leq & (1-\gamma\mu_g)^N \cdot \frac{L_{f,0}^2}{\mu_g^2} + 5\left(\frac{1}{\mu_g^2} + \frac{\gamma}{\mu_g}\right)\left(L_{g,2}^2\sigma_z^2 + L_{f,1}^2\right)\left(\mathbb{E}\left[\|\bar{y}_k^{(T)} - y_k^*\|^2\right] + \sigma_x^2\tilde{\alpha}_k^2 + T\tilde{\beta}_{k+1}^2\right) \\
& + 90C_{\tilde{M}}L_{g,1}^2\left(\frac{1}{\mu_g^2} + \frac{\gamma}{\mu_g}\right)\left(\frac{L_{f,0}^2}{\mu_g^2} + L_{f,0}^2\right)\left(1 - \frac{2\gamma\mu_g}{3}\right)^{N-1},
\end{aligned}
$$

where the second inequality uses Lemma A.4.0.2, the third inequality uses (A.65), and the fourth inequality holds since

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

□

LEMMA A.4.0.10. *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.4.0.7 and A.4.0.8 satisfy*

$$\sum_{k=0}^{K} \tilde{\alpha}_k^2 \leq \frac{2}{1-\rho^2} \sum_{i=0}^{K-1} \alpha_i^2 = \mathcal{O}\left(\sum_{k=0}^{K} \alpha_k^2\right)$$

$$\sum_{k=0}^{K} \tilde{\beta}_{k+1}^2 \leq \frac{4(1+\rho^2)}{(1-\rho^2)^2}\left[(10\sigma_{g,1}^2 + 5\delta^2)\sum_{i=0}^{K}\beta_i^2 + \frac{20L_{g,1}^2\sigma_x^2}{1-\rho^2}\sum_{i=0}^{K-1}\alpha_i^2\right] = \mathcal{O}\left(\sum_{k=0}^{K}(\alpha_k^2+\beta_k^2)\right)$$

$$\sum_{k=1}^{K} C_{k,T} \leq \left(\frac{1}{\mu_g}+1\right) L_{g,1}^2\left[T\sigma_x^2\sum_{k=1}^{K}\tilde{\alpha}_k^2 + 2T^2\sum_{k=0}^{K}\tilde{\beta}_{k+1}^2\right] + \frac{T\sigma_{g,1}^2}{n}\sum_{k=1}^{K}\beta_k^2 = \mathcal{O}\left(\sum_{k=0}^{K}(\alpha_k^2+\beta_k^2)\right).$$

PROOF OF LEMMA A.4.0.10. 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

$$\begin{aligned}
\sum_{k=0}^{K}\tilde{\beta}_{k+1}^2 =& \frac{1+\rho^2}{1-\rho^2}\sum_{k=0}^{K}\sum_{i=0}^{k}\beta_i^2(10\sigma_{g,1}^2 + 10L_{g,1}^2\sigma_x^2\tilde{\alpha}_i^2 + 5\delta^2)\left(\frac{3+\rho^2}{4}\right)^{k-i} \\
\leq& \frac{4(1+\rho^2)}{(1-\rho^2)^2}\sum_{i=0}^{K}\beta_i^2(10\sigma_{g,1}^2 + 10L_{g,1}^2\sigma_x^2\tilde{\alpha}_i^2 + 5\delta^2) \\
\leq& \frac{4(1+\rho^2)}{(1-\rho^2)^2}\left[(10\sigma_{g,1}^2+5\delta^2)\sum_{i=0}^{K}\beta_i^2 + \frac{20L_{g,1}^2\sigma_x^2}{1-\rho^2}\sum_{i=0}^{K-1}\alpha_i^2\right].
\end{aligned}$$

Lastly, we know

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

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

Now we are ready to give the proof of Theorem 3.3.2.

LEMMA A.4.0.11. *Suppose Assumptions 5, 6, 7, and 8 hold. For Algorithm 5 we have*

$$\begin{aligned}(\text{A.81}) \qquad &\sum_{k=0}^{K}\left(\frac{\alpha_k}{2}-\frac{L_\Phi\alpha_k^2}{2}\right)\mathbb{E}\left[\|\bar{r}_k\|^2\right]\\ \leq&\frac{1}{2}\sum_{k=0}^{K}\alpha_k\mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2\right]+2\sigma_u^2\sum_{k=0}^{K}\alpha_k^2+\Phi(0)-\inf_x\Phi(x)+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_0\|^2\right].\end{aligned}$$

PROOF OF LEMMA A.4.0.11. The L_Φ-smoothness of Φ indicates that

$$(\text{A.82}) \qquad \Phi(\bar{x}_{k+1})-\Phi(\bar{x}_k)\leq\nabla\Phi(\bar{x}_k)^\top(-\alpha_k\bar{r}_k)+\frac{L_\Phi\alpha_k^2}{2}\|\bar{r}_k\|^2.$$

Notice that we also have

$$(\text{A.83}) \qquad \frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}\|^2|\mathcal{F}_k\right]-\frac{1}{2}\|\bar{r}_k\|^2=-\alpha_k\|\bar{r}_k\|^2+\alpha_k\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]^\top\bar{r}_k+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}-\bar{r}_k\|^2|\mathcal{F}_k\right].$$

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

$$\begin{aligned}(\text{A.84})\quad&\left(\frac{\alpha_k}{2}-\frac{L_\Phi\alpha_k^2}{2}\right)\mathbb{E}\left[\|\bar{r}_k\|^2\right]\\ \leq&\frac{\alpha_k}{2}\mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2\right]+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}-\bar{r}_k\|^2\right]+\mathbb{E}\left[\Phi(\bar{x}_k)-\Phi(\bar{x}_{k+1})\right]\\ +&\frac{1}{2}\mathbb{E}\left[\|\bar{r}_k\|^2\right]-\frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}\|^2\right]\\ \leq&\frac{\alpha_k}{2}\mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2\right]+2\alpha_k^2\sigma_u^2+\mathbb{E}\left[\Phi(\bar{x}_k)-\Phi(\bar{x}_{k+1})\right]+\frac{1}{2}\mathbb{E}\left[\|\bar{r}_k\|^2\right]-\frac{1}{2}\mathbb{E}\left[\|\bar{r}_{k+1}\|^2\right],\end{aligned}$$

where the second inequality holds since we know

$$\begin{aligned}&\mathbb{E}\left[\|\bar{r}_k\|^2\right]\leq\max\left(\mathbb{E}\left[\|\bar{r}_{k-1}\|^2\right],\mathbb{E}\left[\|\bar{u}_k\|^2\right]\right)\leq\max_{0\leq i\leq k}\mathbb{E}\left[\|\bar{u}_i\|^2\right]\leq\sigma_u^2,\\ &\mathbb{E}\left[\|\bar{r}_{k+1}-\bar{r}_k\|^2\right]=\alpha_k^2\mathbb{E}\left[\|\bar{r}_k-\bar{u}_k\|^2\right]\leq2\alpha_k^2\mathbb{E}\left[\|\bar{r}_k\|^2+\|\bar{u}_k\|^2\right]\leq4\sigma_u^2.\end{aligned}$$

In these two conclusions $\mathbb{E}\left[\|\bar{u}_i\|^2\right] \leq \sigma_u^2$ is due to the first inequality in (A.71). Taking summation on both sides of (A.84), we have (A.81). □

LEMMA A.4.0.12. *For Algorithm 5 we have*

$$\begin{aligned}
&\sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\|\bar{r}_k - \nabla\Phi(\bar{x}_k)\|^2\right] \\
\leq &\mathbb{E}\left[\|\bar{r}_0 - \nabla\Phi(0)\|^2\right] + 2\sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2\right] + 2\sum_{k=0}^{K} \alpha_k \mathbb{E}\left[\|\bar{r}_k\|^2\right] + \sigma_u^2 \sum_{k=0}^{K} \alpha_k^2.
\end{aligned} \tag{A.85}$$

PROOF OF LEMMA A.4.0.12. Recall that in Algorithm 5 we know

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

which implies

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

Hence we know

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

Taking summation on both sides, we obtain (A.85). □

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 3.3.2.

LEMMA A.4.0.13. *In Algorithm 5 if we define*

$$\alpha_k = \frac{\mu_g^4}{3L_{g,1}^2 C_y} \cdot \beta_k \equiv \frac{1}{\sqrt{K}},\ \gamma \text{ such that (A.63) holds},\ N = \Theta(\log K),\ T \geq 1,$$
$$C_y = 5\left(L_{f,1}^2 + \frac{L_{g,2}^2 L_{f,0}^2}{\mu_g^2}\right) + 50L_{g,1}^2\left(\frac{1}{\mu_g^2} + \frac{\gamma}{\mu_g}\right)\left(L_{g,2}^2\sigma_z^2 + L_{f,1}^2\right).$$

we have

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

PROOF OF LEMMA A.4.0.13. Notice that we have

$$\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right] = \frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}) - \frac{1}{n}\sum_{i=1}^{n}\nabla_{xy}^2 g_i(x_{i,k}, y_{i,k}^{(T)})\mathbb{E}\left[z_{i,N}^{(k)}|\mathcal{F}_k\right]$$

and

$$\begin{aligned}
&\nabla\Phi(\bar{x}_k)\\
=&\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\bar{x}_k, y_k^*) - \left(\frac{1}{n}\sum_{i=1}^{n}\nabla_{xy}^2 g_i(\bar{x}_k, y_k^*)\right)\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_y^2 g_i(\bar{x}_k, y_k^*)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_y f_i(\bar{x}_k, y_k^*)\right)\\
=&\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\bar{x}_k, y_k^*) - \frac{1}{n}\left(\sum_{i=1}^{n}\nabla_{xy}^2 g_i(\bar{x}_k, y_k^*)\right)\left(\sum_{i=1}^{n}\nabla_y^2 g_i(\bar{x}_k, y_k^*)\right)^{-1}\left(\sum_{i=1}^{n}\nabla_y f_i(\bar{x}_k, y_k^*)\right)\\
=&\frac{1}{n}\sum_{i=1}^{n}\nabla_x f_i(\bar{x}_k, y_k^*) - \frac{1}{n}\left(\sum_{i=1}^{n}\nabla_{xy}^2 g_i(\bar{x}_k, y_k^*)\right)z_*^{(k)}.
\end{aligned}$$

Hence we know

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

which implies

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

where the third inequality uses (A.71), (A.65) and (A.78). Taking summation on both sides, we have

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

Setting for all k that

$$
\alpha_k=C_{\alpha,\beta}\cdot\beta_k\equiv\frac{1}{\sqrt{K}},\ C_{\alpha,\beta}=\frac{\mu_g}{2\sqrt{3C_y}L_{y^*}},
$$

and using (A.76) and Lemma A.4.0.10, we know

$$
\begin{aligned}
&\frac{1}{\sqrt{K}}\sum_{k=0}^{K}\mathbb{E}\left[\|\mathbb{E}\left[\bar{u}_k|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2\right]\\
=&C_yC_{\alpha,\beta}\sum_{k=0}^{K}\beta_k\|\bar{y}_k^{(T)}-y_k^*\|^2+\mathcal{O}\left(\frac{1}{\sqrt{K}}+\sqrt{K}\left(1-\frac{\gamma\mu_g}{2}\right)^{N-1}\right)\\
=&C_yC_{\alpha,\beta}L_{y^*}^2\sum_{k=0}^{K}\left(\frac{4C_{\alpha,\beta}}{\sqrt{K}\mu_g^2}+\frac{2}{K\mu_g}\right)\mathbb{E}\left[\|\bar{r}_k\|^2\right]+\mathcal{O}\left(1+\sqrt{K}\left(1-\frac{\gamma\mu_g}{2}\right)^{N-1}\right)
\end{aligned}
\tag{A.86}
$$

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

which together with (A.81) and (A.53) imply

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

Hence we know

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

Using the above expression, (A.86) and Lemma A.4.0.12, we know

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

for sufficiently large K. Note that γ is in a constant interval by (A.63), hence $\left(1 - \frac{\gamma\mu_g}{2}\right)$ is a constant that is independent of K. Picking $N = \Theta(\log K)$ such that $\left(1 - \frac{\gamma\mu_g}{2}\right)^{N-1} = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$, 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 (A.71) 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.4.0.10 The above two equalities prove Theorem 3.3.2. To find an ϵ-stationary point, we may set $K = \Theta(\epsilon^{-2})$ and we know from $T \geq 1$, $N = \log K$ that the sample complexity will be $\tilde{\mathcal{O}}(\epsilon^{-2})$. □

A.5. Discussions on the Prior Works Related to Chapter 3

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

A.5.1. Assumption 3.4 (iv) and (v) in Yang et al. [2022].

- Assumption 3.4 (iv) assumes bounded second moment of $\nabla_y g_i(x, y; \xi)$. It is stronger than our Assumption 7 as discussed right after Assumption 7.

  As pointed out by one reviewer during the discussion period, bounded moment condition on $\nabla_y g_i(x, y; \xi)$ is also restrictive especially when g_i is strongly convex in y. To see this, we notice that the unbiasedness of $\nabla_y g_i(x, y; \xi)$ and its bounded second moment imply

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

  for all x, y. Here $\nabla_y g(x, y; \xi) := \frac{1}{n}\sum_{i=1}^n \nabla_y g_i(x, y; \xi_i)$. Then for any y_1, y_2

  $$2C_g \geq \|\nabla_y g(x, y_1) - \nabla_y g(x, y_2)\| \geq \mu_g \|y_1 - y_2\|$$

  where the second inequality uses the fact that $g(x, y)$ is μ_g-strongly convex in y for any x. However $\sup_{y_1, y_2} \|y_1 - y_2\| = +\infty$, which leads to the contradiction, meaning that there does not exist a function g satisfying all the assumptions above. In short, a function cannot be strongly convex and have bounded gradient at the same time , but both assumptions are used in Yang et al. [2022].

- 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.

A.5.2. MDBO. Although Gao et al. [2022] claims that they solve the G-DSBO problem, their hypergradient (see equations (2) and (3) of their paper accessed from arXiv at the time of the submission of our manuscript to ICML: https://arxiv.org/abs/2206.15025v1) is defined as

$$\nabla F(x) := \frac{1}{K}\sum_{k=1}^{K} \nabla F^{(k)}(x),$$

where

$$\nabla F^{(k)}(x) := \nabla_x f^{(k)}(x, y^*(x)) - \nabla^2_{xy} g^{(k)}(x, y^*(x))(\nabla^2_y g^{(k)}(x, y^*(x)))^{-1}\nabla_y f^{(k)}(x, y^*(x)).$$

Clearly, this is not the hypergradient of G-DSBO, unless $g^{(i)}(x, y) = g^{(j)}(x, y)$ for any $1 \le i < j \le n$, which requires an additional assumption that the data distributions that generate the lower level function $g^{(i)}$ are the same. Note that their algorithm cannot be classified as P-DSBO either, because $y^*(x)$ in the above expression is defined globally. Therefore, their algorithm is not designed for neither G-DSBO nor P-DSBO. It is not clear what problem that their algorithm is designed for.

While we are preparing our camera-ready version, we find the latest version of Gao et al. [2022] (which is Gao et al. [2023]), which implicitly uses the condition that all lower level functions are the same. See equation (2) on page 3 of Gao et al. [2023] and the description right above it: "Then, according to Lemma 1 of (Gao, 2022a), we can compute the gradient of $F^{(k)}(x)$ as follows:", where "(Gao, 2022a)" represents Gao [2022], in which their Lemma 1 explicitly states "When the data distributions across all devices are homogeneous". However, all assumptions about MDBO in Gao et al. [2022] do not mention anything about the data distributions of the lower level functions $g^{(i)}$. It should be noted that once all lower level functions $g^{(i)}$ are the same then their problem setup is

one special case of ours in (3.2) (i.e., when $g^{(i)} = g^{(j)}$ for any $i \neq j$), and it does not need to tackle the major challenge discussed in (3.5).

A.5.3. Computational complexity. Assume that computing a stochastic derivative with size m requires $\mathcal{O}(m)$ computational complexity. For example the complexity of computing a stochastic Hessian matrix $\nabla_y^2 g_i(x, y; \xi)$ is $\mathcal{O}(q^2)$ and the complexity of computing a stochastic gradient $\nabla_x f(x, y; \phi)$ is $\mathcal{O}(p)$. Note that computing a Hessian-vector product (or Jacobian-vector product) is as cheap as computing a gradient [Pearlmutter, 1994, Bottou et al., 2018]. FEDNEST [Tarzanagh et al., 2022], SPDB [Lu et al., 2022], and our Algorithm 5 MA-DSBO only require stochastic first order and matrix-vector product oracles and thus the computational complexity is $\tilde{\mathcal{O}}(d\epsilon^{-2})$, where $d := \max(p, q)$. Note that DSBO-JHIP [Chen et al., 2022b] requires computing full Jacobian matrices which lead to $\tilde{\mathcal{O}}(pq\epsilon^{-3})$ complexity. GBDSBO [Yang et al., 2022] computes full Hessian matrices in the Hessian inverse estimation inner loop (Line 10-13 of Algorithm 1 in Yang et al. [2022]), and full Jacobian matrices in the outer loop (Line 8 of Algorithm 1 in Yang et al. [2022]), and thus their computational cost is $\mathcal{O}((q^2 \log(\frac{1}{\epsilon}) + pq)n^{-1}\epsilon^{-2})$.

A.6. Experimental Investigations of Chapter 4

A.6.1. Additional experiments for the orthogonal case. For this section, we follow the same experimental setup as described in Section 4.4.1. Only the hidden-layer width is changed. Specifically, in Figures A.3 and A.4 we plot the training loss, sharpness of training loss and the trajectory-averaging training in various phases.

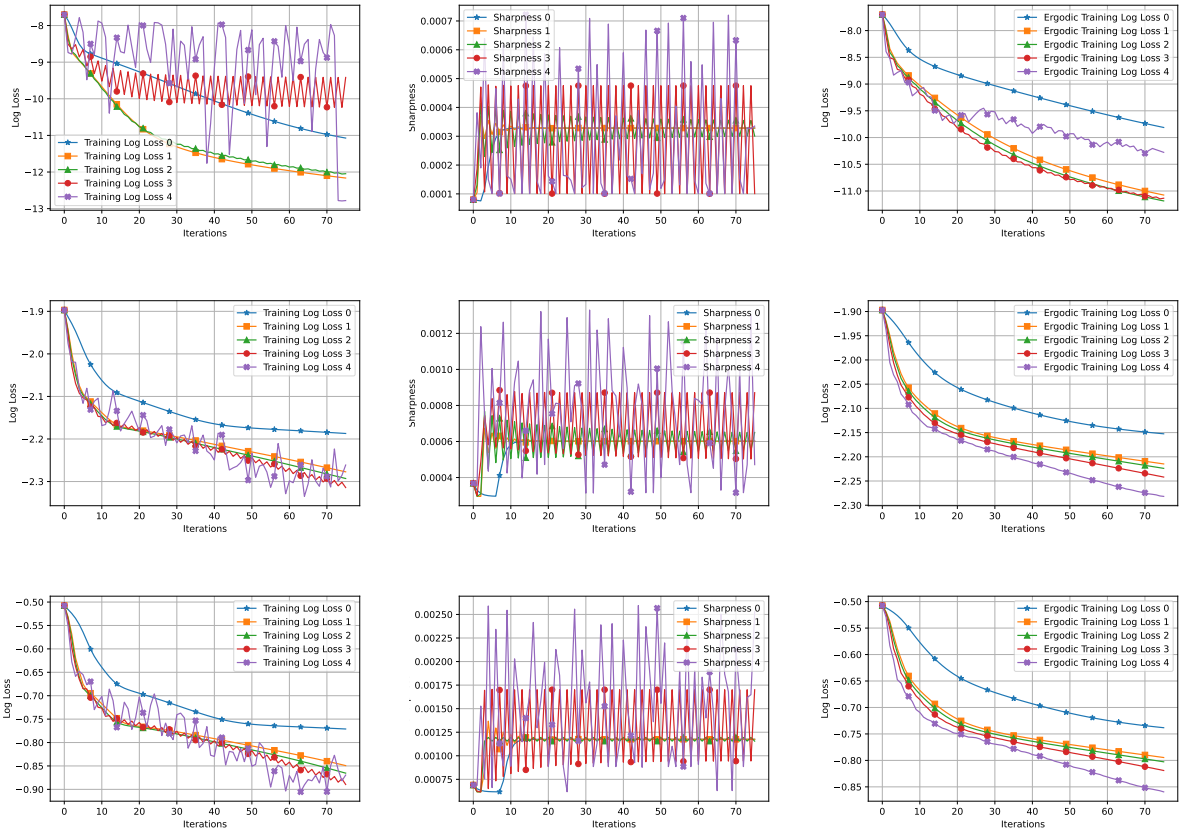


FIGURE A.3. Hidden-layer width =5, with orthogonal data points. Rows from top to bottom represent different levels of noise – mean-zero normal distribution with variance $0, 0.25, 1$. The vertical axes are in log scale for the training loss curves. The second column is about the sharpness of the training loss functions. Numbers $0, 1, 2, 3, 4$ denote different stepsize choices (see Section 4.4.1 for details).

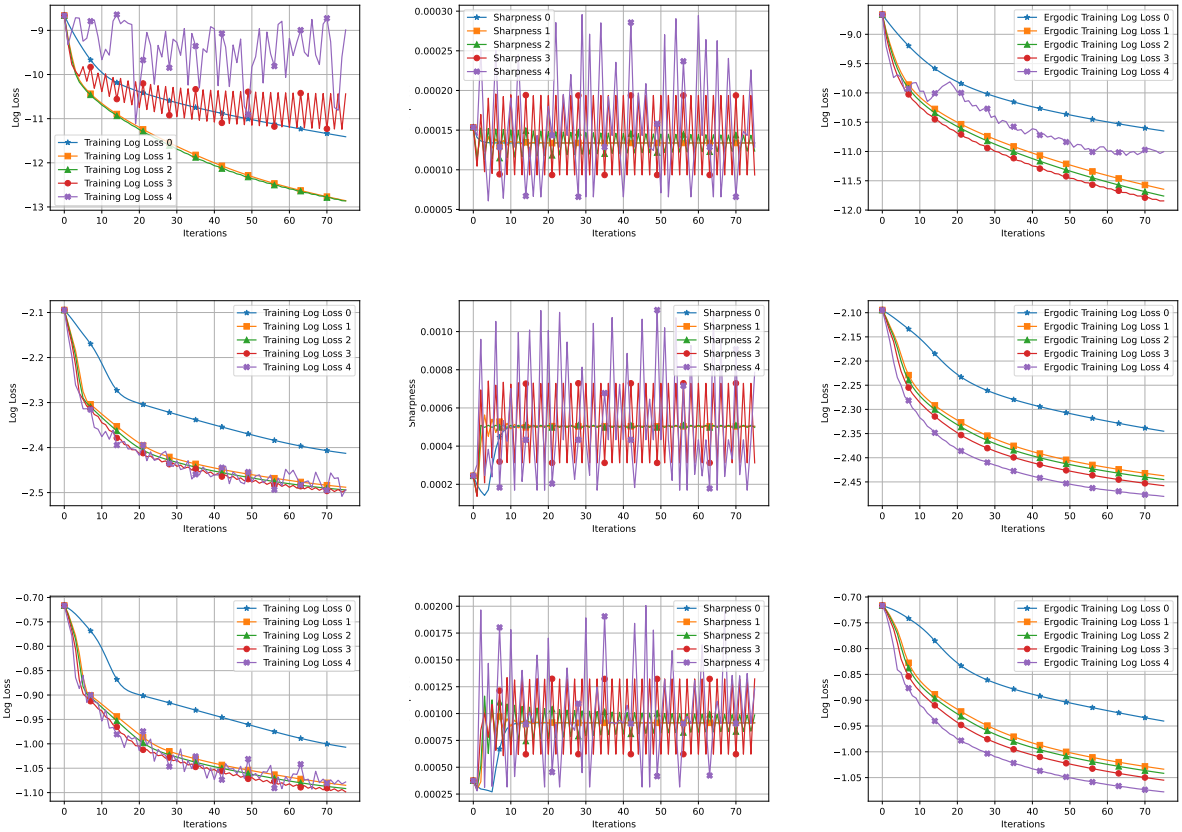


FIGURE A.4. Hidden-layer width =10, with orthogonal data points. Rows from top to bottom represent different levels of noise – mean-zero normal distribution with variance $0, 0.25, 1$. The vertical axes are in log scale for the training loss curves. The second column is about the sharpness of the training loss functions. Numbers $0, 1, 2, 3, 4$ denote different stepsize choices (see Section 4.4.1 for details).

A.6.2. Non-orthogonal data. We next investigate the case when orthogonality condition does not hold. The setup is the same as described in Section 4.4.1 except that $n = 5000$ and each entry of the data matrix $X \in \mathbb{R}^{n\times d}$ is now sampled from a standard normal distribution. We also generate 500 data points from the same distribution for testing. Note that our theory in this work is only applicable for orthogonal data. hence, for these experiments with non-orthogonal data, we first tune the step-size to be as large as possible, say $\eta_{\max}$, so that the training does not diverge and then run the experiments for $\frac{i+1}{5}\eta_{\max}$ with $i = 0, ..., 4$. Hence, the step-sizes for loss and sharpness curves $0, 1, 2, 3, 4$ are chosen to be $10, 20, 30, 40, 50$ for $m = 5, 10$ and $12, 24, 36, 48, 60$ for $m = 25$.

In Figures A.5, A.6 and A.7 we plot the training loss and the testing loss (with and without ergodic trajectory averaging) in log scale. Notably different phases (including the periodic and catapult phases) characterized theoretically for the case of orthogonal data, also appear to be present for the non-orthogonal case. We also make the following intriguing conclusions:

- As a general trend, training roughly in the generalized catapult phase and predicting without doing the ergodic trajectory averaging appears to have the best test error performance.
- In some cases (especially the one with high noise variance), when testing after training in the periodic phase, the test error goes down rapidly in the initial few iterations. Correspondingly, ergodic trajectory averaging after training in the periodic phase, helps to obtain better test error decay compared to ergodic trajectory averaging after training in the catapult phase. However, as mentioned in the previous point, training roughly in the generalized catapult phase and predicting without doing the ergodic trajectory averaging performs the best.
- As discussed in Lim et al. [2022], in various cases, artificially infusing control chaos help to obtain better test accuracy. Given our empirical observations and the results in Lim et al. [2022], it is interesting to design controlled chaos infusion in gradient descent and perform ergodic training averaging to obtain stable and improved test error performance.

Obtaining theoretical results corroborating the above-mentioned observations is challenging future work.

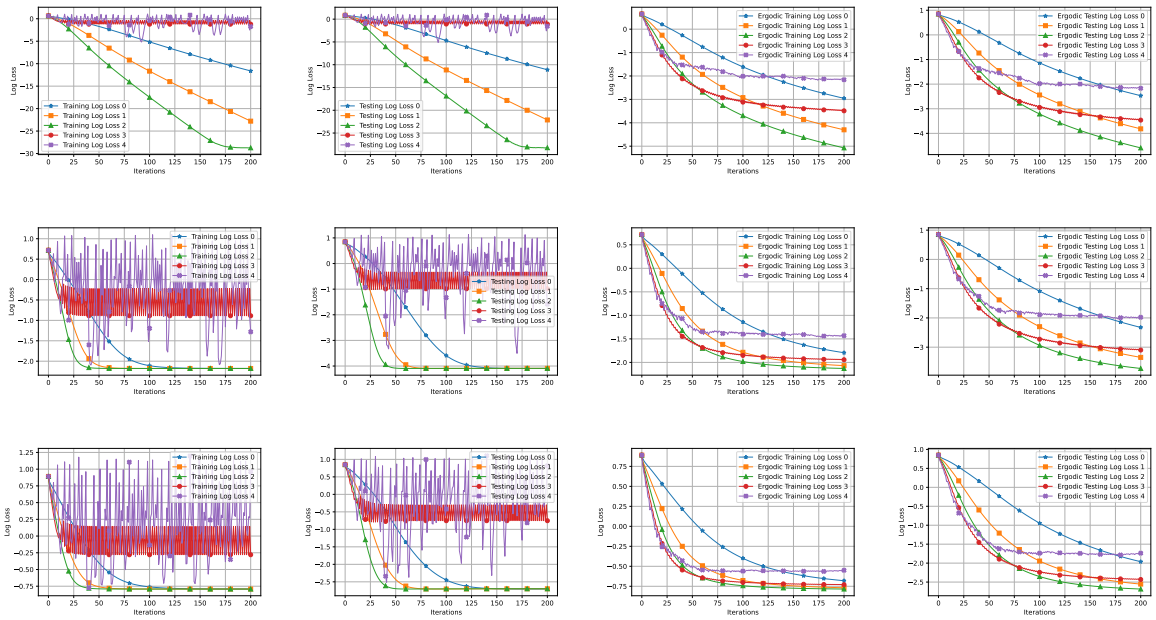


FIGURE A.5. Hidden-layer width=5, with non-orthogonal data points. Rows from top to bottom represent different levels of noise – mean-zero normal distribution with variance $0, 0.25, 1$. The vertical axes are in log scale for loss curves. Numbers $0, 1, 2, 3, 4$ denote different stepsize choices (see Section A.6.2 for details).

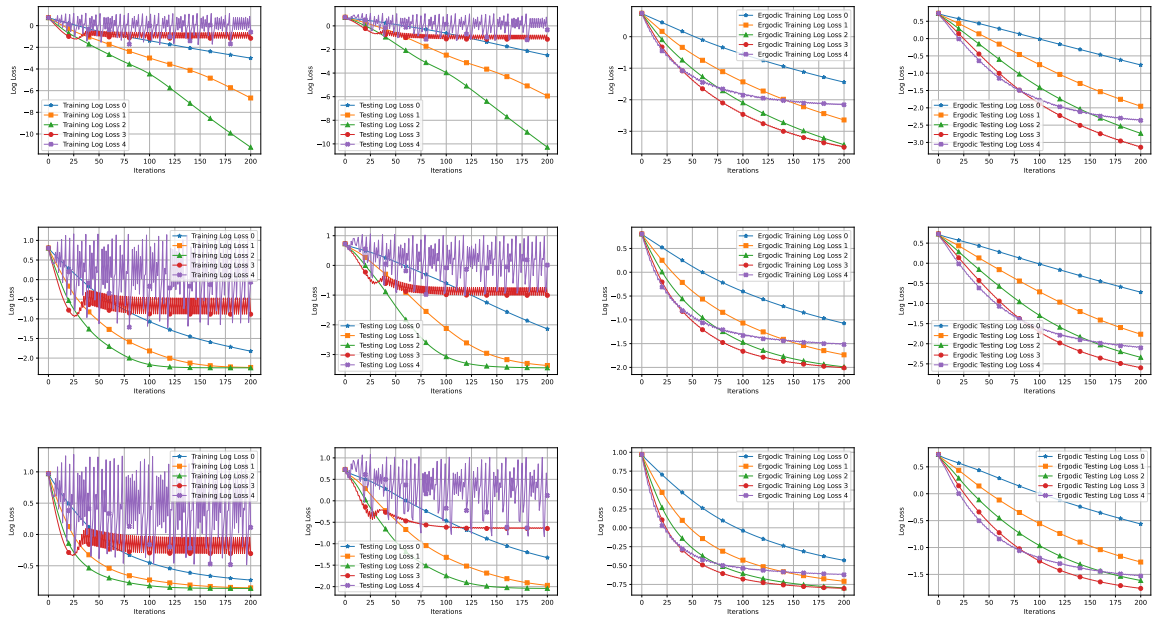


FIGURE A.6. Hidden-layer width=10, with non-orthogonal data points. Rows from top to bottom represent different levels of noise – mean-zero normal distribution with variance $0, 0.25, 1$. The vertical axes are in log scale for loss curves. Numbers $0, 1, 2, 3, 4$ denote different stepsize choices (see Section A.6.2 for details).

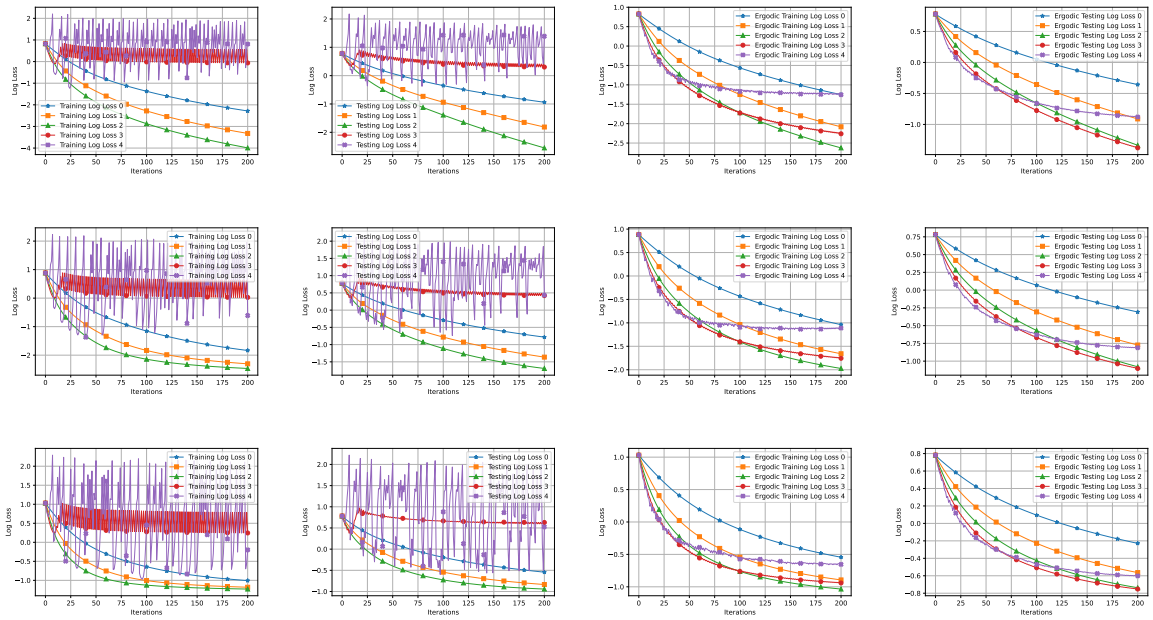


FIGURE A.7. Hidden-layer width=25, with non-orthogonal data points. Rows from top to bottom represent different levels of noise – mean-zero normal distribution with variance $0, 0.25, 1$. The vertical axes are in log scale for loss curves. Numbers $0, 1, 2, 3, 4$ denote different stepsize choices (see Section A.6.2 for details).

A.6.3. Two-layer Neural Network with ReLU. While our main focus in this work is for quadratic activation functions, it is also instructive to examine the dynamics with other activation function, in particular the ReLU activation. Hence, we follow the experimental setup from Section A.6.2, except that the activation function is now ReLU and repeat our experiments. For this case, the step-sizes manually chosen to be $60, 120, 180, 240, 300$ for loss/sharpness curves $0, 1, 2, 3, 4$, respectively.

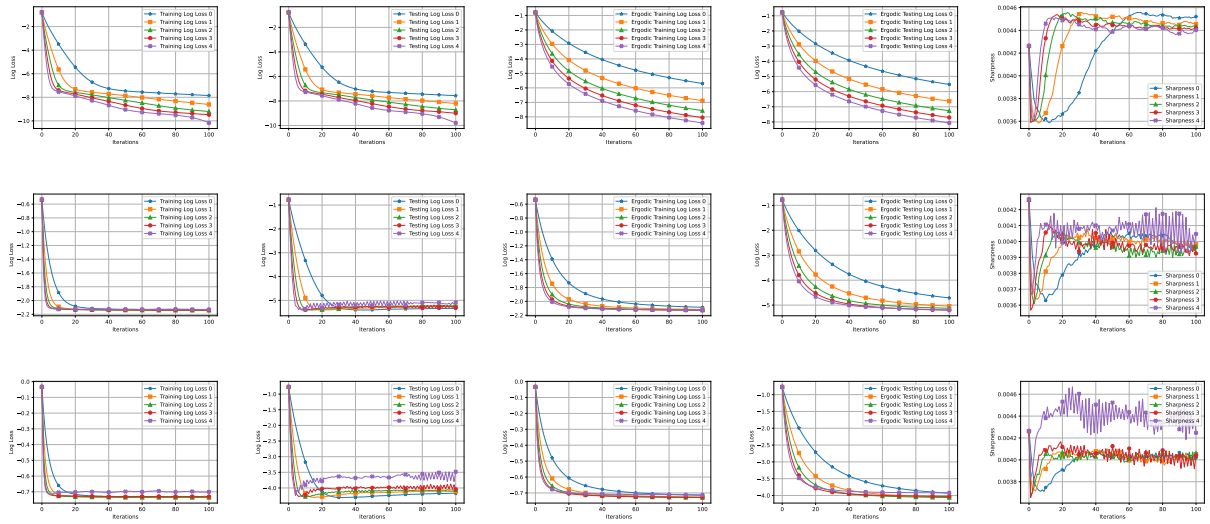


FIGURE A.8. Hidden-layer width=5 with ReLU activation. Rows from top to bottom represent different levels of noise – mean-zero normal distribution with variance 0, 0.25, 1. The vertical axes are in log scale for loss curves. The last column is about the sharpness of the training loss functions. Numbers $0, 1, 2, 3, 4$ denote different stepsize choices (see Section A.6.3 for details).

From Figures A.8 and A.9, (in particular from the sharpness plots), we observe various non-monotonic patterns, roughly including periodic and chaotic patterns. Obtaining a precise theoretical characterization of the training dynamics for this setting is extremely interesting as future work.

A.7. Proofs of Theorems in Chapter 4

A.7.1. Proofs of results in Section 4.2. We first present several technical results required to prove our main results.

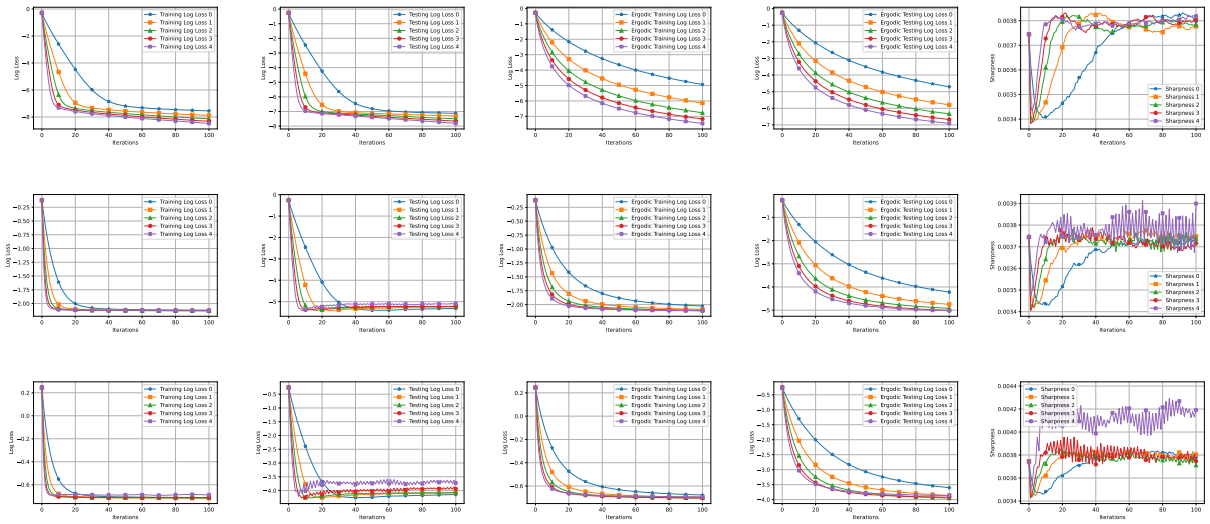


FIGURE A.9. Hidden-layer width=10 with ReLU activation. Rows from top to bottom represent different levels of noise – mean-zero normal distribution with variance 0, 0.25, 1. The vertical axes are in log scale for loss curves. The last column is about the sharpness of the training loss functions. Numbers $0, 1, 2, 3, 4$ denote different stepsize choices (see Section A.6.3 for details).

LEMMA A.7.0.1. *Let $f(x)$ be a polynomial. If all the roots of $f'(x)$ are real and distinct, then we have*

$$\mathsf{S}f(x) = \frac{f'''(x)}{f'(x)} - \frac{3}{2}\left(\frac{f''(x)}{f'(x)}\right)^2 < 0 \quad \textit{for all } x \in I \textit{ with } f'(x) \neq 0.$$

PROOF. See, e.g., the proof of Proposition 11.2 in Devaney [1989]. □

LEMMA A.7.0.2. *Suppose we are given a real-valued continuous function $f(x) : \mathbb{R} \to \mathbb{R}$ and a bounded closed interval $I \subseteq \mathbb{R}$ with $x_0 \in I$. Define $x_k := f^{(k)}(x_0)$. If the sequence $\{x_k\}_{k=0}^{\infty}$ is monotonic, then one of the following holds.*

- *(i) $\{x_k\}_{k=0}^{\infty} \subsetneq I$, i.e., there exists $x_t \notin I$ for some t.*
- *(ii) $\{x_k\}_{k=0}^{\infty} \subseteq I$, and $\lim_{t\to\infty} f^{(t)}(x_0)$ exists and is a fixed point of $f(x)$ in I.*

PROOF. If (i) holds, then the conclusion is true. When (i) does not hold, then $\{x_k\}_{k=0}^{\infty} \subseteq I$. Since this sequence is monotonic and included in a bounded closed interval, we know its limit exists and is in I. Moreover, we have

$$\lim_{t\to\infty} x_t = \lim_{t\to\infty} x_{t+1} = \lim_{t\to\infty} f(x_t) = f(\lim_{t\to\infty} x_t),$$

where the last equality holds since f is continuous. Clearly $\lim_{t\to\infty} x_t$ is a fixed point of f. □

The following lemma characterizes the basic properties of the cubic function f_a defined in (4.2).

LEMMA A.7.0.3. *Suppose $a > 0$. Then $f_a(z)$ has the following properties.*

- *(i) The local minimum and maximum of $f_a(z)$ are at $z = 1$ and $z = \frac{1-2a}{3}$ respectively, and*

$$f_a(1) = -a,\ f_a\left(\frac{1-2a}{3}\right) = \frac{(2a-1)(2a^2+7a-4)}{27} = \frac{4a^3+12a^2-15a+4}{27}.$$

- *(ii) $f_a(z)$ is monotonically increasing on $[-a, \frac{1-2a}{3}]$, monotonically decreasing on $[\frac{1-2a}{3}, 1]$, and monotonically increasing on $[1, 2]$.*
- *(iii) For any $-a \le z \le 2$, we have $-a \le f_a(z) \le \max\left\{f_a\left(\frac{1-2a}{3}\right), 2\right\}$. Moreover, $f_a\left(\frac{1-2a}{3}\right) \le 2$ if and only if $a \le 2$.*

PROOF. Note that we have

$$f_a'(z) = 3z^2 + 2(a-2)z + (1-2a) = (z-1)(3z+2a-1). \tag{A.87}$$

which implies 1 and $\frac{1-2a}{3}$ are critical points of $f_a(z)$. Moreover, by $f_a''(z) = 6z + 2a - 4$ we know $f_a''(1) > 0$ and $f_a''(\frac{1-2a}{3}) < 0$. Hence, they are local minimum and maximum respectively. The rest of (i) is true by calculation. (ii) is true by noticing the expression of $f_a'(z)$ in (A.87). (iii) is a direct conclusion of (i) and (ii) since for $-a \le z \le 2$ we have

$$-a = \min\{f_a(1), f_a(-a)\} \le f_a(z) \le \max\left\{f_a\left(\frac{1-2a}{3}\right), f_a(2)\right\}.$$

By (i) and some calculation we know

$$f_a\left(\frac{1-2a}{3}\right) - 2 = \frac{4a^3+12a^2-15a-50}{27} = \frac{(2a+5)^2(a-2)}{27}.$$

This proves the rest of (iii). □

LEMMA A.7.0.4. *Suppose $2\sqrt{2} - 2 < a \le 1$. Define five subintervals of $[-a, 2]$ as follows.*

$$I_1 = \left[-a, \frac{2-a-\sqrt{a^2+4a}}{2}\right],\ I_2 = \left[\frac{2-a-\sqrt{a^2+4a}}{2}, 0\right],$$
$$I_3 = [0, 0.25],\ I_4 = \left[0.25, \frac{2-a+\sqrt{a^2+4a}}{2}\right],\ I_5 = \left[\frac{2-a+\sqrt{a^2+4a}}{2}, 2\right].$$

Then we have

- *(i)* $f_a(I_1) \subseteq I_1 = I_2,\ f_a(I_4) = I_1 \cup I_2,\ f_a(I_5) = I_3 \cup I_4 \cup I_5$.
- *(ii)* $f_a(I_2) \subseteq I_3,\ f_a(I_3) \subseteq I_2$.

PROOF. We first prove (i). By Lemma A.7.0.3 we know $f_a(z)$ is increasing on I_1, achieving its local minimum at $z = 1$ on I_4, increasing on I_5, then we know

$$f_a(I_1) = \left[f_a(-a), f_a\left(\frac{2 - a - \sqrt{a^2 + 4a}}{2} \right) \right] = [-a, 0] = I_1 \cup I_2.$$

$$f_a(I_4) = \left[f_a(1), \max\left\{ f_a(0.25), f_a\left(\frac{2 - a + \sqrt{a^2 + 4a}}{2} \right) \right\} \right] = [-a, 0] = I_1 \cup I_2.$$

$$f_a(I_5) = \left[f_a\left(\frac{2 - a + \sqrt{a^2 + 4a}}{2} \right), f_a(2) \right] = [0, 2] = I_3 \cup I_4 \cup I_5.$$

This completes the proof of (i).

To prove (ii), observe that when $a \in (2\sqrt{2} - 2, 1]$ we have $\frac{2-a-\sqrt{a^2+4a}}{2} < \frac{1-2a}{3} < 0$. By Lemma A.7.0.3 we know the local maximum of f_a over $I_2 = \left[\frac{2-a-\sqrt{a^2+4a}}{2}, 0 \right]$ is achieved at $\frac{1-2a}{3}$, this together with the fact that $f_a(0) = f_a\left(\frac{2-a-\sqrt{a^2+4a}}{2} \right) = 0$ implies

$$f_a\left(I_2\right) = \left[f_a(0), f_a\left(\frac{1 - 2a}{3} \right) \right] = \left[0, \frac{4a^3 + 12a^2 - 15a + 4}{27} \right] \subseteq [0, 0.25],$$

where the last subset inclusion is true since

$$(4a^3 + 12a^2 - 15a + 4)' = 12a^2 + 24a - 15 > 0,\ \forall a \in (2\sqrt{2} - 2, 1].$$

This implies when $a \in (2\sqrt{2} - 2, 1]$,

$$\frac{4a^3 + 12a^2 - 15a + 4}{27} \leq \frac{(4a^3 + 12a^2 - 15a + 4)|_{a=1}}{27} = \frac{5}{27} < 0.25.$$

On the other hand, we know from Lemma A.7.0.3 that on $I_3 = [0, 0.25] (\subseteq \left[\frac{1-2a}{3}, 1 \right])$ f_a is decreasing. Hence,

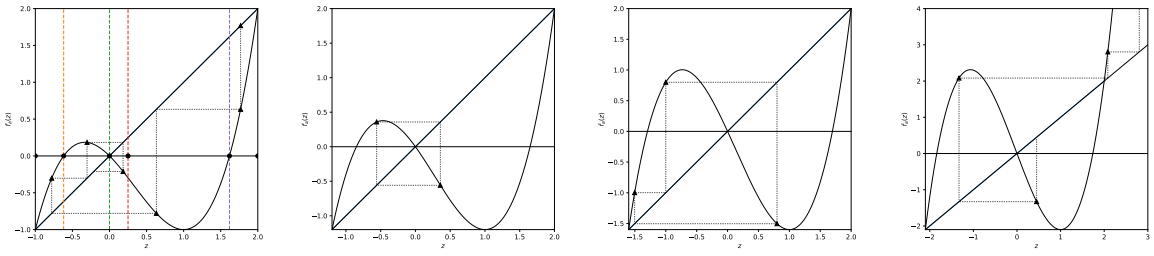
$$f_a(I_3) = [f_a(0.25), f_a(0)] = \left[-\frac{7}{16}a + \frac{9}{16}, 0 \right] \subseteq \left[\frac{2 - a - \sqrt{a^2 + 4a}}{2}, 0 \right] = I_2.$$

where the last subset inclusion is true since

$$f_a(0.25) = -\frac{7}{16}a + \frac{9}{16} > \frac{2 - a - \sqrt{a^2 + 4a}}{2}, \ \forall a \in (2\sqrt{2} - 2, 1].$$

This completes the proof of (ii). □

See Figure A.10(a) for a visualization of the subintervals $I_1, ..., I_5$ for $a = 1$ and an example of the orbit on it.



(a) (b) (c) (d)

FIGURE A.10. From left to right: cubic function $f_1(z)$ with different regions diveded by subintervals and a trajectory of $\{z_i\}_{i=0}^5$, cubic function $f_{1.2}(z)$ with two period-2 point, cubic function $f_{1.6}(z)$ with a period-3 point, and cubic function $f_{2.1}(z)$ with a diverging orbit. We have the cubic curve and the identical mapping line as the solid curves. We use four colored dashed lines in Figure A.10(a) to represent the boundaries that are orthogonal to the endpoints of I_2 and I_4 defined in Lemma A.7.0.4 respectively. The triangle markers represent some terms of a certain orbit, in which horizontal and vertical dotted lines visualize the transitioning trajectory between consecutive terms in an orbit.

LEMMA A.7.0.5. *Suppose* $0 < a \leq 1$ *and* $-a \leq z_0 \leq 2$. *Then we have*

- *(i)* $-a \leq z_t \leq 2$ *for any* t, *and* f_a *does not have a period-2 point on* $[-a, 2]$.
- *(ii) If* z_0 *is chosen from* $[-a, 2]$ *uniformly at random, then* $\lim_{t\to\infty} z_t = 0$ *almost surely. Moreover, if* $0 < a \leq 2\sqrt{2} - 2$, *then almost surely* $|z_{t+1}| \leq |z_t|$ *for all* t. *If* $2\sqrt{2} - 2 < a \leq 2$, *then almost surely* $\{|z_t|\}_{t=0}^{\infty}$ *has catapults.*

PROOF. The boundedness of each iterate (i.e., $z_t \in [-a, 2]$) can be proved by using simple induction and Lemma A.7.0.3, $0 < a \leq 1$, and $-a \leq z_0 \leq 2$. To prove the rest of (i), by (4.2) we know a period-2 point is a solution of

$$f_a^{(2)}(z) = z, \ f_a(z) \neq z$$

which are equivalent to

$$g_a(z)g_a(zg_a(z)) = 1, z \notin \{-a, 0, 2\}. \tag{A.88}$$

Hence it suffices to prove (A.88) do not have a solution. Define

$$h_a(z) = g_a(z) - 1 = (z + a)(z - 2) < 0, \ \forall z \in (-a, 2).$$

We have

$$\begin{aligned}
&g_a(z)g_a(zg_a(z)) - 1 \\
=&h_a(z) + h_a(z)h_a(zg_a(z)) + h_a(zg_a(z)) \\
=&h_a(z)(1 + h_a(zg_a(z))) + (z + a + zh_a(z))(z - 2 + zh_a(z)) \\
=&h_a(z)(1 + h_a(zg_a(z))) + h_a(z) + (z(z - 2) + z(z + a))h_a(z) + z^2h_a^2(z) \\
=&h_a(z)(h_a(zg_a(z)) + z^2h_a(z) + 2z^2 + (a - 2)z + 2). \qquad \text{(A.89)}
\end{aligned}$$

We have

$$\begin{aligned}
&h_a(zg_a(z)) + z^2h_a(z) + 2z^2 + (a - 2)z + 2 \\
=&(zg_a(z) + a)(zg_a(z) - 2) + z^2(z + a)(z - 2) + 2z^2 + (a - 2)z + 2 \\
=&z^2(z^2 + (a - 2)z + 1 - 2a)^2 + (a - 2)z(z^2 + (a - 2)z + 1 - 2a) - 2a \\
&+ z^2(z + a)(z - 2) + 2z^2 + (a - 2)z + 2 \\
=&z^6 + (2a - 4)z^5 + (a^2 - 8a + 7)z^4 - (4a^2 - 12a + 8)z^3 + (5a^2 - 10a + 7)z^2 \\
&- (2a^2 - 6a + 4)z + 2 - 2a \\
=&(z^2 + (a - 1)z + 1 - a)(z^4 + (a - 3)z^3 + (3 - 3a)z^2 + (2a - 2)z + 2). \qquad \text{(A.90)}
\end{aligned}$$

Observe that

$$z^2 + (a - 1)z + (1 - a) \geq (1 - a) - \frac{(a - 1)^2}{4} = \frac{(3 + a)(1 - a)}{4} \geq 0, \ \forall a \in (0, 1]. \tag{A.91}$$

The equalities hold if and only if $z = 0,\ a = 1$. We also have

$$z^4 + (a-3)z^3 + (3-3a)z^2 + (2a-2)z + 2 > 0,\ \forall z \in \{0,1,2\}$$

$$\begin{aligned} &z^4 + (a-3)z^3 + (3-3a)z^2 + (2a-2)z + 2 \\ =&z(z-1)(z-2)\left(a + z + \frac{1}{z} + \frac{1}{z^2-3z+2}\right), \forall z \notin \{0,1,2\}. \end{aligned}$$

For different z we can verify the following inequalities via basic algebra or Young's inequality:

$$z(z-1)(z-2) < 0,\ \left(a + z + \frac{1}{z} + \frac{1}{z^2-3z+2}\right) < 1 + 2 + \frac{1}{2} + \frac{1}{-0.25} < 0, \qquad \forall z \in (1,2).$$

$$z(z-1)(z-2) > 0,\ \left(a + z + \frac{1}{z} + \frac{1}{z^2-3z+2}\right) > 0 + 1 + 1 + 0 > 0, \qquad \forall z \in (0,1).$$

$$z(z-1)(z-2) < 0,\ \left(a + z + \frac{1}{z} + \frac{1}{z^2-3z+2}\right) < 1 - 1 - 1 + \frac{1}{2} < 0, \qquad \forall z \in (-a,0).$$

Thus we may conclude that

$$z^4 + (a-3)z^3 + (3-3a)z^2 + (2a-2)z + 2 > 0,\ \forall z \in (-a,2). \tag{A.92}$$

By (A.89), (A.90), (A.91), (A.92), we know $g_a(z)g_a(zg_a(z)) - 1 \neq 0$ if $z \notin \{-a,0,2\}$. Hence f_a does not have a period-2 point on $[-a,2]$.

To prove the first part in (ii) (the limit converges to 0 almost surely), we will prove

$$(1)\ \lim_{t\to\infty} z_t \in \{-a,0,2\},\ (2)\ \text{The set } S \text{ such that the orbit with } z_0 \in S \text{ has measure 0.} \tag{A.93}$$

We now consider two cases – $a \in (0, 2\sqrt{2}-2]$ and $a \in (2\sqrt{2}-2, 1]$.

Case 1: $a \in (0, 2\sqrt{2}-2]$**.** Note that we have

$$|g_a(z_t)| = |z_t^2 + (a-2)z_t + 1 - 2a| \leq \max\left(|g_a(-a)|, |g_a(2)|, |g_a\left(1 - \frac{a}{2}\right)|\right) = 1,$$

where the last equality holds since $g_a(-a) = g_a(2) = 1$ and $|g_a\left(1-\frac{a}{2}\right)| = \frac{a^2+4a}{4} \leq 1$ for any $a \in (0, 2\sqrt{2}-2]$. Hence, we know

$$|z_{t+1}| = |f_a(z_t)| = |z_t g_a(z_t)| \leq |z_t|,\ \forall z_t \in [-a,2] \tag{A.94}$$

Hence $\lim_{t\to\infty}|z_t|$ exists.

$$\lim_{t\to\infty}|z_t| = \lim_{t\to\infty}|z_{t+1}| = \lim_{t\to\infty}|z_t||g_a(z_t)|$$

Hence, we know

$$\lim_{t\to\infty}|z_t| = 0, \text{ or } \lim_{t\to\infty}|z_t| \neq 0, \lim_{t\to\infty}|g_a(z_t)| = 1.$$

If $\lim_{t\to\infty}|z_t| \neq 0$, then we have two subcases

- Sub-case 1: $\lim_{t\to\infty} z_t$ exists. We can verify that

$$\lim_{t\to\infty} z_t = \lim_{t\to\infty} z_{t+1} = f_a(\lim_{t\to\infty} z_t)$$

  and thus $\lim_{t\to\infty} z_t$ is one of the fixed points of $f_a(z) \in \{-a, 0, 2\}$.
- Sub-case 2: $\lim_{t\to\infty} z_t$ does not exist. Since $\lim_{t\to\infty}|z_t|$ exists, we know there exists an infinite subsequence (denoted as A_1) of $\{z_t\}_{t=0}^{\infty}$ with some limit c and the complement of the sequence, as another infinite subsequence (denoted as A_2), has limit $-c$ for some constant $c > 0$. Hence, we can pick a sequence of the subscripts $k_1 < k_2 < ... < k_n < ...$ such that $z_{k_1}, ..., z_{k_n}, ...$ belong to A_1 and $z_{k_1+1}, ..., z_{k_n+1}, ...$ belong to A_2. Moreover, we have

$$c = \lim_{i\to\infty} z_{k_i} = -\lim_{i\to\infty} z_{k_i+1} = -\lim_{i\to\infty} z_{k_i} g_a(z_{k_i}) = -c g_a(c) \tag{A.95}$$

  This implies that $g_a(c) = -1$, i.e.,

$$c^2 + (a-2)c + 2 - 2a = 0.$$

  From its discriminant $(a-2)^2 - 4(2-2a) = a^2 + 4a - 4 \leq 0$ for $a \in (0, 2\sqrt{2}-2]$ where equality holds only at $2\sqrt{2}-2$, we know $a = 2\sqrt{2}-2$ and thus $c = 2-\sqrt{2}$. However, we can apply the similar trick and pick another sequence $\tilde{k}_1 < \tilde{k}_2 < ... < \tilde{k}_n < ...$ such that $z_{\tilde{k}_1}, ..., z_{\tilde{k}_n}, ...$ belong to A_2 and $z_{\tilde{k}_1+1}, ..., z_{\tilde{k}_n+1}, ...$ belong to A_1. This implies

$$-c = \lim_{i\to\infty} z_{\tilde{k}_i} = -\lim_{i\to\infty} z_{\tilde{k}_i+1} = -\lim_{i\to\infty} z_{k_i} g_a(z_{k_i}) = -(-c) g_a(-c)$$

which gives

$$c^2 - (a-2)c + 2 - 2a = 0.$$

This contradicts with $a = 2\sqrt{2} - 2$ and $c = 2 - \sqrt{2}$. This means case 2 does not exist.

Hence, we know $|z_t|$ is decreasing (not necessarily strictly) and $\lim_{t\to\infty} z_t \in \{-a, 0, 2\}$.

Case 2: $a \in (2\sqrt{2} - 2, 1]$**.** We divide the interval $[-a, 2]$ into the following five parts:

$$I_1 = \left[-a, \frac{2 - a - \sqrt{a^2 + 4a}}{2}\right], \; I_2 = \left[\frac{2 - a - \sqrt{a^2 + 4a}}{2}, 0\right],$$
$$I_3 = [0, 0.25], \; I_4 = \left[0.25, \frac{2 - a + \sqrt{a^2 + 4a}}{2}\right], \; I_5 = \left[\frac{2 - a + \sqrt{a^2 + 4a}}{2}, 2\right].$$

Recall that by Lemma A.7.0.4 we have:

$$f_a(I_1) = I_1 \cup I_2, \; f_a(I_2) \subseteq I_3, \; f_a(I_3) \subseteq I_2, \; f_a(I_4) = I_1 \cup I_2, \; f_a(I_5) = I_3 \cup I_4 \cup I_5. \tag{A.96}$$

We have the following conclusion. Observe that f_a is continuous, and

$$z_{t+1} - z_t = f_a(z_t) - z_t = z_t(z_t + a)(z_t - 2) \geq 0, \; \forall z_t \in I_1 = \left[-a, \frac{2 - a - \sqrt{a^2 + 4a}}{2}\right],$$
$$z_{t+1} - z_t = f_a(z_t) - z_t = z_t(z_t + a)(z_t - 2) \leq 0, \; \forall z_t \in I_5 = \left[\frac{2 - a + \sqrt{a^2 + 4a}}{2}, 2\right].$$

We know if the sequence $\{z_t\}_{t=0}^{\infty}$ visits I_5, by Lemma A.7.0.2 we know either $\lim_{t\to\infty} z_t = 2$ or there exists $M > 0$ such that $z_t \notin I_5$ for any $t \geq M$. Then if the sequence visits I_1 then by Lemma A.7.0.2 either $\lim_{t\to\infty} z_t = -a$ or there exists $\tilde{M} > M > 0$ such that $z_t \in I_2 \cup I_3$ for any $t \geq \tilde{M}$, since $f_a(I_1) \subseteq I_1 \cup I_2$ and $f_a(I_2 \cup I_3) \subseteq I_2 \cup I_3$. Hence, the proof is reduced to the case when $z_0 \in I_2 \cup I_3$. For the case when $z_0 \in I_2 \cup I_3 = \left[\frac{2-a-\sqrt{a^2+4a}}{2}, 0.25\right]$. The key observation is to show that in this interval

$$|z_{t+2}| \leq |z_t|. \tag{A.97}$$

Recall that by Lemma A.7.0.4 (ii) we have

$$f_a(I_2) \subseteq I_3, \; f_a(I_3) \subseteq I_2. \tag{A.98}$$

To prove (A.97), we know it holds when $z_t = 0$. When $z_t \neq 0$, by (A.98) we know $f_a^{(2)}(z_t)$ and z_t have the same sign provided $z_t \in I_2 \cup I_3 = \left[\frac{2-a-\sqrt{a^2+4a}}{2}, 0.25\right]$. This together with

$$f_a^{(2)}(z) = f_a(z)g_a(f_a(z)) = zg_a(z)g_a(zg_a(z))$$

implies that $g_a(z)g_a(zg_a(z)) \geq 0$ when $z \in \left[\frac{2-a-\sqrt{a^2+4a}}{2}, 0\right) \cup (0, 0.25]$. Thus we know

$$|z_{t+2}| = |z_t g_a(z_t)g_a(z_t g_a(z_t))| = |z_t| g_a(z_t)g_a(z_t g_a(z_t)).$$

Thus to prove (A.97) it suffices to show $g_a(z)g_a(zg_a(z)) - 1 \leq 0$, which is true by combining (A.89), (A.90), (A.91), and (A.92). This completes the proof of (1) in (A.93). To prove (2) in (A.93), we first notice that $f_a(z) - z = z(z+a)(z-2) > 0$ for any $z \in (-a, 0)$, and thus $z_{t+1} > z_t$ for any z_t near $-a$. Hence, $\lim_{t\to\infty} z_t = -a$ if and only if there exists t such that $z_t = -a$. This implies that $f_a^{(t)}(z_0) = -a$ for some t. Similarly, $f_a(z) - z < 0$ for any $z \in (0, 2)$, which implies $z_{t+1} < z_t$ for any z_t near 2. Hence, $\lim_{t\to\infty} z_t = 2$ if and only if $z_0 = 2$. Define

$$S = \bigcup_{n=0}^{\infty} f_a^{(-n)}(-a) \cup \{2\}$$

where $f_a^{(-n)}(-a)$ denotes the preimage of $-a$ under $f_a^{(n)}$. Clearly, each preimage is a finite set, and thus S is countable. Hence, we know as long as $z_0 \in [-a, 2]\backslash S$, we have $\lim_{t\to\infty} z_t = 0$. Since S is a countable set and z_0 is chosen uniformly at random, we know $\lim_{t\to\infty} z_t = 0$ almost surely.

For the rest of (ii), we have already proved in (A.94) that $\{|z_t|\}_{t=0}^{\infty}$ is decreasing when $0 < a \leq 2\sqrt{2} - 2$. To see $\{|z_t|\}_{t=0}^{\infty}$ has catapults when $2\sqrt{2} - 2 < a \leq 1$, we consider the following intervals

$$J_1 = [-a, 0] = I_1 \cup I_2, \ J_2 = \left[0, \min\left\{\frac{2 - a + \sqrt{a^2 + 4a - 4}}{2}, 0.25\right\}\right] \subseteq I_3,$$

where we have $a^2 + 4a - 4 > 0$ for $a > 2\sqrt{2} - 2$ so J_2 is well-defined. Notice that

$$0 < z < \frac{2 - a + \sqrt{a^2 + 4a - 4}}{2} \Leftrightarrow g_a(z) < -1, \ z > 0.$$

Hence we know for any $z_t \in J_2$, we will have

$$|z_{t+1}| = |z_t g_a(z_t)| > |z_t|. \tag{A.99}$$

On the other hand, notice that 0 is in the orbit if and only if $z_0 \notin S_0$, where S_0 is defined as

$$S_0 = \bigcup_{n=0}^{\infty} f_a^{(-n)}(0)$$

where $f_a^{-n}(z)$ denotes the set of preimage of z under $f_a^{(n)}$. Note that each preimage is finite and thus S_0 is countable. Hence, we know almost surely the orbit will not contain 0, and recall that by Lemma (A.7.0.4) (ii) and $\lim_{t\to\infty} z_t = 0$, we know there are infinitely many t such that $t \in J_2$, and thus (A.99) holds for infinitely many t almost surely. By definition 4.2.2, we know $\{|z_t|\}$ has catapults almost surely. □

The following theorem indicates that, f_a is chaotic provided that $a > a_*$ where $a_* \in (1, 2)$

LEMMA A.7.0.6. *Suppose* $1 < a \leq 2$ *and* $-a \leq z_0 \leq 2$. *Then we have*

- *(i)* $-a \leq z_t \leq 2$ *for any* t, *and* $f_a(z)$ *has a period-2 point on* $[0, 1]$.
- *(ii) There exists* $a_* \in (1, 2)$ *such that for any* $a \in (a_*, 2)$, f_a *is Li-Yorke chaotic, and for any* $a \in (1, a_*)$, f_a *is not Li-Yorke chaotic.*
- *(iii) If there exists an asymptotically stable orbit and* z_0 *is chosen from* $[-a, 2]$ *uniformly at random, then the orbit of* z_0 *is asymptotically periodic almost surely.*

PROOF. The boundedness of z_t is a direct result of Lemma A.7.0.3 (iii). To prove the rest of (i), we notice that for $a \in (1, 2]$,

$$g_a(0)g_a(0g_a(0)) = (1-2a)^2 > 1, \; g_a(1)g_a(1g_a(1)) = -a < -1.$$

By continuity of $g_a(zg_a(z))$ we know there exists a point $z_0 \in (0, 1)$ such that $g_a(z_0 g_a(z_0)) = 1$. This indicates that $f^{(2)}(z_0) = z_0 g_a(z_0 g_a(z_0)) = z_0$ but clearly $f_a(z_0) \neq z_0$ since $(0, 1)$ does not contain any fixed point of f_a.

To prove (ii), notice that

$$f_1\left(\frac{1-2\times 1}{3}\right) = \frac{5}{27} < 1 < 2 = f_2\left(\frac{1-2\times 2}{3}\right).$$

By continuity of $f_a\left(\frac{1-2a}{3}\right)$ (with respect to a) there exists $c \in (1, 2)$ such that

$$f_c\left(\frac{1-2c}{3}\right) = \frac{(2c-1)(2c^2+7c-4)}{27} = 1. \tag{A.100}$$

Moreover we have

$$f_c(-c) = -c < \frac{1-2c}{3},\ f_c\left(\frac{1-2c}{3}\right) = 1 > \frac{1-2c}{3}.$$

Hence by continuity of $f_c(z)$, we can pick $z_0 \in \left(-c, \frac{1-2c}{3}\right)$ such that $f_c(z_0) = \frac{1-2c}{3}$. We have

$$-c < z_0 < \frac{1-2c}{3} = f_c(z_0). \tag{A.101}$$

By (A.100), (A.101), and Lemma A.7.0.3 (i), we have

$$f_c^{(3)}(z_0) = f_c^{(2)}\left(\frac{1-2c}{3}\right) = f_c(1) = -c \le z_0, \tag{A.102}$$

$$f_c(z_0) = \frac{1-2c}{3} < 1 = f_c(1) = f_c^{(2)}(z_0)\,. \tag{A.103}$$

Combining (A.101), (A.102), (A.103) we can easily verify that

$$f_c^{(3)}(z_0) \le z_0 < f_c(z_0) < f_c^{(2)}(z_0).$$

By Theorem A.8.1 (i.e., Theorem 1 in Li and Yorke [1975]), we know f_c is Li-Yorke chaotic. Moreover, for any $a \in (c, 2]$, we know

$$f_a\left(\frac{1-2a}{3}\right) = \frac{(2a-1)(2a^2+7a-4)}{27} > \frac{(2c-1)(2c^2+7c-4)}{27} = f_c\left(\frac{1-2c}{3}\right) = 1,$$

which together with $f_a(0) = 0 < 1$ implies we can pick y_0 such that

$$\frac{1-2a}{3} < y_0 < 0,\ f_a(y_0) = 1. \tag{A.104}$$

Similarly, we have

$$f_a(-a) = -a < \frac{1-2a}{3} < y_0,\ f_a\left(\frac{1-2a}{3}\right) > 1 > y_0$$

which implies we can pick x_0 such that

$$-a < x_0 < \frac{1-2a}{3},\ f_a(x_0) = y_0. \tag{A.105}$$

Now we know

$$f_a^{(3)}(x_0) < x_0 < f_a(x_0) < f_a^{(2)}(x_0).$$

By Theorem A.8.1 (i.e., Theorem 1 in Li and Yorke [1975]), we know f_a is Li-Yorke chaotic. Hence, we know c defined in (A.100) satisfies that for any $a \in (c, 2]$, f_a is Li-Yorke chaotic. Hence, we know

$$a_* = \inf_{a \in (1,2)} \{a : f_b \text{ is Li-Yorke chaotic for any } b \in [a, 2].\}$$

where the set is not empty, since we have proven c belongs to the above set. This completes the proof of (ii).

To prove (iii), we notice that if $f_a(z)$ has an asymptotically stable periodic orbit, by Theorem A.8.2 (i.e., Theorem 2.7 in Singer [1978]) and the fact that $f_a(x)$ has negative Schwarzian derivative at non-critical points (Lemma A.7.0.1) and we know there exists a critical point c of $f_a(z)$ such that the orbit of c converges to this asymptotically stable orbit. Notice that by Lemma A.7.0.3 we know $c = 1$ or $\frac{1-2a}{3}$. $c = 1$ can be excluded since $f_a(1) = -a$, and $-a$ is an unstable period-1 point. Hence, we know $c = \frac{1-2a}{3}$ is asymptotically periodic. By Theorems A.8.3 and A.8.4 (i.e., Theorem B and Corollary in Nusse [1987]), we know almost surely z_0 will not converge to any periodic orbit if z_0 is chosen from $[-a, 2]$ uniformly at random. This completes the proof.

□

Remarks:

- See Figure A.10(b) for a pair of period-2 points when $a = 1.2$, and Figure A.10(c) for a period-3 orbit when $a = 1.6$. The triangle markers denote the periodic points.
- By Theorem A.8.2 (i.e., Theorem 2.7 in Singer [1978]) and the fact that $-a$ is an unstable period-1 point we know $f_a(z)$ has at most one asymptotically stable periodic orbit.

LEMMA A.7.0.7. *Suppose $a > 2$. z_0 is chosen from $[-a, 2]$ uniformly at random. Then $\lim_{t\to\infty} |z_t| = +\infty$ almost surely.*

PROOF. Notice that by Lemma A.7.0.3 we know

$$f_a\left(\frac{1-2a}{3}\right) = \frac{4a^3 + 12a^2 - 15a + 4}{27} > \frac{(4a^3 + 12a^2 - 15a + 4)|_{a=2}}{27} = 2, \ \forall a > 2,$$

where the inequality holds since $4a^3 + 12a^2 - 15a + 4$ is increasing on $(2, \infty)$. Moreover, we have

$$f_a(z) - z = z(z + a)(z - 2) > 0, \ \forall z \in (2, \infty).$$

Hence we know for the initialization at the critical point $z_0 = \frac{1-2a}{3}$, we have $z_1 > 2$, and the whole sequence is increasing. On the other hand, all fixed points of $f_a(z)$ are no greater than 2, we know z_t will diverge to $+\infty$. For another critical point $z_0 = 1$ we know its orbit converges to the periodic orbit of $z_0 = -a$, which is an unstable period-1 point. Hence, we know from Theorem A.8.2 (i.e., Theorem 2.7 in Singer [1978]) that there does not exist an asymptotically stable periodic orbit, otherwise the orbit of one critical point must converge to it. Hence, by Theorems A.8.3 and A.8.4 (i.e., Theorem B and Corollary in Nusse [1987]) we know $\lim_{t\to\infty} |z_t| = +\infty$ almost surely provided z_0 uniformly chosen from $(-a, 2)$, i.e., almost all points in $[-a, 2]$ converge to the absorbing boundary point $+\infty$. □

A.7.2. Proofs of results in Section 4.3.

Proof of Theorem 4.3.1. Define

$$\alpha^{(t)} := c + \gamma X^\top w^{(t)},\ \beta := y + \frac{c^2}{2\gamma},\ \kappa := \eta\gamma \|X\|^2 .$$

To prove (i), we observe that

$$\nabla_w g(w; X) = (c + \gamma(X^\top w))X$$

Let weights at time t be $w^{(t)}$. Thus, the gradient descent takes the form

$$w^{(t+1)} = w^{(t)} - \eta(g(w^{(t)}; X) - y)(c + \gamma X^\top w^{(t)})X = w^{(t)} - \eta e^{(t)}\alpha^{(t)}X.$$

Simple calculation gives

$$e^{(t)} = \frac{(\alpha^{(t)})^2}{2\gamma} - \beta \tag{A.106}$$

and

$$\alpha^{(t+1)} = (1 - \eta\gamma \|X\|^2 e^{(t)})\alpha^{(t)} = (1 - \kappa e^{(t)})\alpha^{(t)}. \tag{A.107}$$

Hence

$$e^{(t+1)} - e^{(t)} = \frac{1}{2\gamma}\left((\alpha^{(t+1)})^2 - (\alpha^{(t)})^2\right) = \left((1 - \kappa e^{(t)})^2 - 1\right)\frac{(\alpha^{(t)})^2}{2\gamma}$$

which together with (A.106) implies

$$\kappa e^{(t+1)} = \kappa e^{(t)}(\kappa e^{(t)} + \beta\kappa)\left(\kappa e^{(t)} - 2\right) + \kappa e^{(t)}. \tag{A.108}$$

By definition of a and z_t in (4.6) we know $a = \beta\kappa$ and $z_t = \kappa e^{(t)}$. We know (i) holds.

To compute the largest eigenvalue of the Hessain matrix (i.e., the sharpness defined in EoS literature) of the loss in (ii), we notice that the gradient of the loss function takes the form

$$\nabla\ell(w) = (g(w; X) - y)\nabla_w g(w; X).$$

Hence

$$\nabla^2\ell(w) = \nabla_w g(w; X)\nabla_w g(w; X)^\top + (g(w; X) - y)\nabla_w^2 g(w; X) = (\alpha^2 + \gamma e)XX^\top,$$

where we overload the notation and define

$$\alpha = c + \gamma X^\top w,\ e = g(w; X) - y.$$

The sharpness is given by

$$\lambda_{\max}(\nabla^2\ell(w^{(t)})) = ((\alpha^{(t)})^2 + \gamma e^{(t)})\left\|X\right\|^2 = (3\gamma e^{(t)} + 2\gamma y + c^2)\left\|X\right\|^2 = \frac{3z_t + 2a}{\eta}.$$

□

Proof of Theorem 4.3.2. The gradient descent takes the form

$$w^{(t+1)} = w^{(t)} - \frac{\eta}{2n}\sum_{i=1}^{n}\nabla\ell_i(w^{(t)}) = w^{(t)} - \frac{\eta}{n}\sum_{i=1}^{n} e^{(t)}(X_i)\alpha^{(t)}(X_i)X_i.$$

Similarly to (A.106), for each error term $e^{(t)}(X_i)$ we have

$$e^{(t)}(X_i) = \frac{(\alpha^{(t)}(X_i))^2}{2\gamma} - \beta(X_i), \tag{A.109}$$

and

$$\begin{aligned}
&\alpha^{(t+1)}(X_i) = \gamma X_i^\top w^{(t+1)} + c(X_i) \\
=&\gamma\left(X_i^\top w^{(t)} - \frac{\eta}{n}\sum_{j=1}^{n} e^{(t)}(X_j)\alpha^{(t)}(X_j)X_i^\top X_j\right) + c(X_i)
\end{aligned}$$

$$
\begin{aligned}
=&\alpha^{(t)}(X_i) - \frac{\gamma\eta}{n}\sum_{j=1}^{n} e^{(t)}(X_j)\alpha^{(t)}(X_j)X_i^\top X_j \\
=&\alpha^{(t)}(X_i) - \frac{\gamma\eta}{n}\sum_{j=1}^{n}\left(\frac{\alpha^{(t)}(X_j)^3}{2\gamma} - \beta(X_j)\alpha^{(t)}(X_j)\right)X_i^\top X_j
\end{aligned}
$$

We overload the notation and set

$$
\mathbf{X} = (X_1, ..., X_n)^\top,\ \#(\mathbf{X}) = (\#(X_1), ..., \#(X_n))^\top,\ \forall \# \in \{\alpha^{(t)}, e^{(t)}, a, \beta\}.
$$

We can obtain

$$
\alpha^{(t+1)}(\mathbf{X}) = \alpha^{(t)}(\mathbf{X}) - \frac{\eta}{n}\mathbf{X}\mathbf{X}^\top\left(\frac{\alpha^{(t)}(X)^3}{2} - \gamma\beta(X)\odot\alpha^{(t)}(X)\right), \tag{A.110}
$$

where $\odot$ denotes the Hadamard product.

As $\mathbf{X}\mathbf{X}^\top = \mathrm{diag}(\|X_1\|^2, ..., \|X_n\|^2)$, we can rewrite (A.110) as the following non-interacting version for each data point:

$$
\begin{aligned}
\alpha^{(t+1)}(X_i) =&\alpha^{(t)}(X_i) - \frac{\eta\|X_i\|^2}{2n}\left(\alpha^{(t)}(X_i)^3 - 2\gamma\beta(X_i)\alpha^{(t)}(X_i)\right) \\
=&\left(1 - \frac{\gamma\eta\|X_i\|^2}{n}e^{(t)}(X_i)\right)\alpha^{(t)}(X_i).
\end{aligned}
$$

This together with (A.109) implies

$$
\begin{aligned}
&e^{(t+1)}(X_i) - e^{(t)}(X_i) = \frac{1}{2\gamma}\left((\alpha^{(t+1)}(X_i))^2 - (\alpha^{(t)}(X_i))^2\right) \\
=&\left(-\frac{2\gamma\eta\|X_i\|^2}{n}e^{(t)}(X_i) + \frac{\gamma^2\eta^2\|X_i\|^4}{n^2}(e^{(t)}(X_i))^2\right)\left(e^{(t)}(X_i) + \beta(X_i)\right) \\
=&\kappa_n(X_i)e^{(t)}(X_i)\left(\kappa_n(X_i)e^{(t)}(X_i) - 2\right)\left(e^{(t)}(X_i) + \beta(X_i)\right)
\end{aligned}
$$

By definition of $z_i^{(t)}$ and a_i we know

$$
z_i^{(t+1)} = z_i^{(t)}(z_i^{(t)} + a_i)(z_i^{(t)} - 2) + z_i^{(t)} = f_{a_i}(z_i^{(t)}).
$$

The sharpness is given by

$$
\nabla^2\ell(w^{(t)}) = \frac{1}{n}\sum_{i=1}^{n}\left(\nabla_w g(w^{(t)}; X_i)\nabla_w g(w^{(t)}; X_i)^\top + (g(w^{(t)}; X_i) - y_i)\nabla_w^2 g(w^{(t)}; X_i)\right)
$$

$$
\begin{aligned}
=&\frac{1}{n}\sum_{i=1}^{n}\left((\alpha^{(t)}(X_i))^2+\gamma e^{(t)}(X_i)\right)X_iX_i^\top\\
=&\frac{1}{n}\sum_{i=1}^{n}(3\gamma e^{(t)}(X_i)+2\gamma y_i+c^2(X_i))X_iX_i^\top.
\end{aligned}
$$

Therefore we know

$$
\nabla^2\ell(w^{(t)})X_i=\frac{1}{n}(3\gamma e^{(t)}(X_i)+2\gamma y_i+c^2(X_i))\left\|X_i\right\|^2X_i=\frac{3z_i^{(t)}+2a_i}{\eta}X_i,\quad\text{for all } 1\le i\le n.
$$

which means we find n eigenvalues and eigenvectors pairs $\left(\frac{3z_1^{(t)}+2a_1}{\eta},X_1\right),...,\left(\frac{3z_n^{(t)}+2a_n}{\eta},X_n\right)$. Note that $\nabla^2\ell(w^{(t)})$ is a sum of n rank-1 matrices, and we have found n orthogonal eigenvalues. Hence we know $\lambda_{\max}(\nabla^2\ell(w^{(t)}))=\max_{1\le i\le n}\frac{3z_i^{(t)}+2a_i}{\eta}$. This completes the proof. □

Proof of Theorem 4.3.3. Define

$$
\mathbf{A}^{(t)}=\frac{2\eta}{\sqrt{m}dn}\sum_{j=1}^{n}e_j^{(t)}X_jX_j^\top.
$$

Note that we have

$$
\nabla\ell_j^{(t)}(\mathbf{U}^{(t)})=\left(\frac{1}{\sqrt{m}d}\sum_{i=1}^{m}(X_j^\top u_i^{(t)})^2-y_j\right)\left(\frac{2}{\sqrt{m}d}X_jX_j^\top\mathbf{U}^{(t)}\right)=\frac{2}{\sqrt{m}d}e_j^{(t)}X_jX_j^\top\mathbf{U}^{(t)}. \tag{A.111}
$$

This implies that the gradient descent update takes the form

$$
\mathbf{U}^{(t+1)}=\mathbf{U}^{(t)}-\frac{\eta}{n}\sum_{j=1}^{n}\nabla\ell_j^{(t)}(\mathbf{U}^{(t)})=\mathbf{U}^{(t)}-\frac{2\eta}{\sqrt{m}dn}\sum_{j=1}^{n}e_j^{(t)}X_jX_j^\top\mathbf{U}^{(t)}=\left(I-\mathbf{A}^{(t)}\right)\mathbf{U}^{(t)}.
$$

Also we have

$$
\begin{aligned}
e_j^{(t+1)}-e_j^{(t)}=&\frac{1}{\sqrt{m}d}\sum_{i=1}^{m}\left((X_j^\top u_i^{(t+1)})^2-(X_j^\top u_i^{(t)})^2\right)\\
=&\frac{1}{\sqrt{m}d}X_j^\top\left(\mathbf{U}^{(t+1)}(\mathbf{U}^{(t+1)})^\top-\mathbf{U}^{(t)}(\mathbf{U}^{(t)})^\top\right)X_j
\end{aligned}
$$

and

$$
\mathbf{U}^{(t+1)}(\mathbf{U}^{(t+1)})^\top=\left(I-\frac{2\eta}{\sqrt{m}dn}\sum_{j=1}^{n}e_j^{(t)}X_jX_j^\top\right)\mathbf{U}^{(t)}(\mathbf{U}^{(t)})^\top\left(I-\frac{2\eta}{\sqrt{m}dn}\sum_{j=1}^{n}e_j^{(t)}X_jX_j^\top\right).
$$

Hence we know

$$\begin{aligned}
&e_j^{(t+1)} - e_j^{(t)} \\
=&\frac{1}{\sqrt{m}d}\left(X_j^\top \mathbf{A}^{(t)}\mathbf{A}^{(t)}(\mathbf{A}^{(t)})^\top \mathbf{A}^{(t)} X_j - 2X_j^\top \mathbf{A}^{(t)}\mathbf{U}^{(t)}(\mathbf{U}^{(t)})^\top X_j\right) \\
=&\frac{1}{\sqrt{m}d}\left(\frac{4\eta^2}{md^2n^2}(e_j^{(t)})^2 \left\|X_j\right\|^4 X_j^\top \mathbf{U}^{(t)}(\mathbf{U}^{(t)})^\top X_j - \frac{4\eta}{\sqrt{m}dn} e_j^{(t)} \left\|X_j\right\|^2 X_j^\top \mathbf{U}^{(t)}(\mathbf{U}^{(t)})^\top X_j\right) \\
=&\left(\frac{4\eta^2 \left\|X_j\right\|^4}{md^2n^2}(e_j^{(t)})^2 - \frac{4\eta \left\|X_j\right\|^2}{\sqrt{m}dn} e_j^{(t)}\right)\left(e_j^{(t)} + y_j\right),
\end{aligned}$$

where the second equality uses $\mathbf{X}\mathbf{X}^\top = \mathrm{diag}(\left\|X_1\right\|^2, ..., \left\|X_n\right\|^2)$. By definition of $z_i^{(t)}$ and a_i we know

$$z_i^{(t+1)} = f_{a_i}(z_i^{(t)}).$$

Hence we know the training dynamics of this model can be captured by the cubic map as well. □

A.8. Auxiliary Results in Chapter 4

THEOREM A.8.1 (Theorem 1 in Li and Yorke [1975]). *Let I be a compact interval and let $f : I \to I$ be continuous. Assume there is a point $a \in I$ for which the points $b = f(a)$, $c = f^{(2)}(a)$ and $d = f^{(3)}(a)$ satisfy*

$$d \le a < b < c \ (or\ d \ge a > b > c).$$

Then f is Li-Yorke chaotic.

THEOREM A.8.2 (Theorem 2.7 in Singer [1978]). *Let I be a compact interval and let $f : I \to I$ be a three times continuously differentiable function. If the Schwarzian derivative of f satisfies*

$$\mathsf{S}f(x) = \frac{f'''(x)}{f'(x)} - \frac{3}{2}\left(\frac{f''(x)}{f'(x)}\right)^2 < 0 \quad for\ all\ x \in I\ with\ f'(x) \neq 0.$$

Then the stable set of every asymptotically stable orbit of f contains a critical point of f.

THEOREM A.8.3 (Theorem B in Nusse [1987]). *Let I be an interval and let $f : I \to I$ be a three times continuously differentiable function having at least one aperiodic point on I and satisfying:*

- *(i) f has a nonpositive Schwarzian derivative, i.e.,*

$$\mathsf{S}f(x) = \frac{f'''(x)}{f'(x)} - \frac{3}{2}\left(\frac{f''(x)}{f'(x)}\right)^2 \leq 0 \quad \text{for all } x \in I \text{ with } f'(x) \neq 0.$$

- *(ii) The set of points, whose orbits do not converge to an (or the) absorbing boundary point(s) of I for f is a nonempty compact set.*
- *(iii) The orbit of each critical point for f converges to an asymptotically stable periodic orbit of f or to an (or the) absorbing boundary point(s) of I for f.*
- *(iv) The fixed points of $f^{(2)}$ are isolated.*

Then we have

- *(1) The set of points whose orbits do not converge to an asymptotically stable periodic orbit of f or to an (or the) absorbing boundary point(s) of I for f has Lebesgue measure 0;*
- *(2) There exists a positive integer p such that almost every point x in I is asymptotically periodic with $f^{(p)}(x) = x$, provided that $f(I)$ is bounded.*

THEOREM A.8.4 (Corollary in Nusse [1987]). *Assume that $f : \mathbb{R} \to \mathbb{R}$ is a polynomial function having at least one aperiodic point and satisfying the following conditions:*

- *(i) The orbit of each critical point of f converges to an asymptotically stable periodic orbit of f or to an (or the) absorbing boundary point(s) for f;*
- *(ii) Each critical point of f is real.*

Then f satisfies the assumptions (i)-(iv) of Theorem A.8.3.

Bibliography

R. L. Adler, A. G. Konheim, and M. H. McAndrew. Topological entropy. *Transactions of the American Mathematical Society*, 114(2):309–319, 1965.

N. Agarwal, S. Goel, and C. Zhang. Acceleration via fractal learning rate schedules. In *Proceedings of the* 38^{th} *International Conference on Machine Learning*, pages 87–99, 2021.

A. Agarwala and Y. Dauphin. SAM operates far from home: Eigenvalue regularization as a dynamical phenomenon. In *Proceedings of the* 40^{th} *International Conference on Machine Learning*, pages 152–168. PMLR, 2023.

A. Agarwala, F. Pedregosa, and J. Pennington. Second-order regression models exhibit progressive sharpening to the edge of stability. In *Proceedings of the* 40^{th} *International Conference on Machine Learning*, 2023.

K. Ahn, J. Zhang, and S. Sra. Understanding the unstable convergence of gradient descent. In *Proceedings of the* 39^{th} *International Conference on Machine Learning*, pages 247–257. PMLR, 2022.

K. Ahn, S. Bubeck, S. Chewi, Y. T. Lee, F. Suarez, and Y. Zhang. Learning threshold neurons via edge of stability. *Advances in Neural Information Processing Systems*, 36, 2024.

Z. Akhtar, A. S. Bedi, S. T. Thomdapu, and K. Rajawat. Projection-free stochastic bi-level optimization. *IEEE Transactions on Signal Processing*, 70:6332–6347, 2022.

K. T. Alligood, T. D. Sauer, and J. A. Yorke. Chaos: An introduction to dynamical systems., 1997.

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.

J. M. Altschuler and P. A. Parrilo. Acceleration by Stepsize Hedging I: Multi-Step Descent and the Silver Stepsize Schedule. *preprint arXiv:2309.07879*, 2023.

M. Andriushchenko, A. V. Varre, L. Pillaud-Vivien, and N. Flammarion. SGD with large step sizes learns sparse features. In *Proceedings of the* 40^{th} *International Conference on Machine Learning*, pages 903–925. PMLR, 2023.

M. Arbel and J. Mairal. Amortized implicit differentiation for stochastic bilevel optimization. *arXiv preprint arXiv:2111.14580*, 2021.

M. Arbel and J. Mairal. Amortized implicit differentiation for stochastic bilevel optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=3PN4iyXBeF.

Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, A. Sekhari, and K. Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299. PMLR, 2020.

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

S. Arora, Z. Li, and A. Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the* 39^{th} *International Conference on Machine Learning*, pages 948–1024. PMLR, 2022.

B. Aulbach and B. Kieninger. On three definitions of chaos. *Nonlinear Dyn. Syst. Theory*, 1(1): 23–37, 2001.

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.

Y. Bengio. Gradient-based optimization of hyperparameters. *Neural computation*, 12(8):1889–1900, 2000.

L. Bertinetto, J. F. Henriques, P. Torr, and A. Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxnZh0ct7.

Q. Bertrand, Q. Klopfenstein, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization. In *International Conference on Machine Learning*, pages 810–821. PMLR, 2020.

T. Birdal, A. Lou, L. Guibas, and U. Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2021.

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

J. Bracken and J. T. McGill. Mathematical programs with optimization problems in the constraints. *Operations research*, 21(1):37–44, 1973.

B. Branner and J. H. Hubbard. The iteration of cubic polynomials. part I: The global topology of parameter space. *Acta mathematica*, 160(3-4):143–206, 1988.

A. Camuto, G. Deligiannidis, M. A. Erdogdu, M. Gurbuzbalaban, U. Simsekli, and L. Zhu. Fractal structure and generalization properties of stochastic optimization algorithms. In *Advances in Neural Information Processing Systems*, 2021.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International conference on machine learning*, pages 654–663. PMLR, 2017.

S. Chakraborty, A. Bedi, A. Koppel, H. Wang, D. Manocha, M. Wang, and F. Huang. Parl: A unified framework for policy alignment in reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

N. Chandramoorthy, A. Loukas, K. Gatmiry, and S. Jegelka. On the generalization of learning algorithms that do not converge. *Advances in Neural Information Processing Systems*, 35:34241–34257, 2022.

L. Chen and J. Bruna. Beyond the edge of stability via two-step gradient updates. In *Proceedings of the* 40^{th} *International Conference on Machine Learning*, 2023.

L. Chen, J. Xu, and J. Zhang. On bilevel optimization without lower-level strong convexity. *arXiv preprint arXiv:2301.00712*, 2023a.

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, 2021a.

T. Chen, Y. Sun, and W. Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021b.

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.

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

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

X. Chen, A. Roy, Y. Hu, and K. Balasubramanian. Stochastic optimization algorithms for instrumental variable regression with streaming data. *arXiv preprint arXiv:2405.19463*, 2024.

I. Chevyrev, P. K. Friz, A. Korepanov, and I. Melbourne. Superdiffusive limits for deterministic fast–slow dynamical systems. *Probability theory and related fields*, 178(3-4):735–770, 2020.

J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *The* 9^{th} *International Conference on Learning Representations*, 2021.

E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

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

M. Dagréou, P. Ablin, S. Vaiter, and T. Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=wlEOsQ917F.

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

M. Dagréou, P. Ablin, S. Vaiter, and T. Moreau. How to compute hessian-vector products? In *ICLR Blogposts 2024*, 2024. URL https://iclr-blogposts.github.io/2024/blog/bench-hvp/. https://iclr-blogposts.github.io/2024/blog/bench-hvp/.

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

A. Damian, E. Nichani, and J. D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The* 11^{th} *International Conference on Learning Representations*, 2023.

W. De Melo and S. Van Strien. *One-dimensional dynamics*, volume 25. Springer Science & Business Media, 2012.

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

R. Devaney. An introduction to chaotic dynamical systems. *An Introduction to Chaotic Dynamical Systems*, 1989.

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.

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

B. Dupuis, G. Deligiannidis, and U. Simsekli. Generalization bounds using data-dependent fractal dimensions. In *Proceedings of the* 40^{th} *International Conference on Machine Learning*, pages 8922–8968, 2023.

A. Fannjiang and T. Strohmer. The numerics of phase retrieval. *Acta Numerica*, 29:125–228, 2020.

L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pages 1165–1173. PMLR, 2017.

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.

N. Franzová and J. Smítal. Positive sequence topological entropy characterizes chaotic maps. *Proceedings of the American Mathematical Society*, pages 1083–1086, 1991.

S. Frei, G. Vardi, P. L. Bartlett, N. Srebro, and W. Hu. Implicit bias in leaky relu networks trained on high-dimensional data. *arXiv preprint arXiv:2210.07082*, 2022.

H. Gao. On the convergence of momentum-based algorithms for federated stochastic bilevel optimization problems. *arXiv preprint arXiv:2204.13299*, 2022.

H. Gao, X. Wang, L. Luo, and X. Shi. On the convergence of stochastic compositional gradient descent ascent method. In *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

H. Gao, B. Gu, and M. T. Thai. Stochastic bilevel distributed optimization over a network. *arXiv preprint arXiv:2206.15025*, 2022.

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

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

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.

J. Gilmer, B. Ghorbani, A. Garg, S. Kudugunta, B. Neyshabur, D. Cardoze, G. E. Dahl, Z. Nado, and O. Firat. A loss curvature perspective on training instabilities of deep learning models. In *The* 12^{th} *International Conference on Learning Representations*, 2022.

T. Giovannelli, G. Kent, and L. N. Vicente. Inexact bilevel stochastic gradient methods for constrained and unconstrained lower-level problems. *arXiv preprint arXiv:2110.00604*, 2021.

T. Giovannelli, G. Kent, and L. N. Vicente. Bilevel optimization with a multi-objective lower-level problem: Risk-neutral and risk-averse formulations. *arXiv preprint arXiv:2302.05540*, 2023.

B. Goujaud, D. Scieur, A. Dieuleveut, A. B. Taylor, and F. Pedregosa. Super-acceleration with cyclical step-sizes. In *International Conference on Artificial Intelligence and Statistics*, pages 3028–3065. PMLR, 2022.

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.

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.

R. Grazzi, M. Pontil, and S. Salzo. Bilevel optimization with a lower-level contraction: Optimal sample complexity without warm-start. *Journal of Machine Learning Research*, 24(167):1–37, 2023. URL http://jmlr.org/papers/v24/22-1043.html.

B. Grimmer. Provably faster gradient descent via long steps. *preprint arXiv:2307.06324*, 2023.

B. Grimmer, K. Shu, and A. L. Wang. Accelerated Gradient Descent via Long Steps . *preprint arXiv:2309.09961*, 2023.

A. Gu, S. Lu, P. Ram, and T.-W. Weng. Min-max multi-objective bilevel optimization with applications in robust machine learning. In *The Eleventh International Conference on Learning*

Representations, 2023. URL https://openreview.net/forum?id=PvDY71zKsvP.

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*, 2021a.

Z. Guo, Y. Xu, W. Yin, R. Jin, and T. Yang. A novel convergence analysis for algorithms of the ADAM family and beyond. *arXiv preprint arXiv:2104.14840*, 2021b.

J. K. Hale and H. Koçak. *Dynamics and bifurcations*, volume 3. Springer Science & Business Media, 2012.

L. Herrmann, M. Granz, and T. Landgraf. Chaotic dynamics are intrinsic to neural network training with SGD. In *Advances in Neural Information Processing Systems*, 2022.

L. Hodgkinson, U. Simsekli, R. Khanna, and M. Mahoney. Generalization bounds using lower tail exponents in stochastic optimizers. In *Proceedings of the* 39^{th} *International Conference on Machine Learning*, pages 8774–8795, 2022.

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.

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

Q. Hu, Y. Zhong, and T. Yang. Multi-block min-max bilevel optimization with applications in multi-task deep auc maximization. *Advances in Neural Information Processing Systems*, 35: 29552–29565, 2022.

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

M. Huang, X. Chen, K. Ji, S. Ma, and L. Lai. Efficiently escaping saddle points in bilevel optimization. *arXiv preprint arXiv:2202.03684v3*, 2023.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

K. Jaganathan, Y. C. Eldar, and B. Hassibi. Phase retrieval: An overview of recent developments. *Optical Compressive Imaging*, pages 279–312, 2016.

S. Jastrzebski, M. Szymczak, S. Fort, D. Arpit, J. Tabor, K. Cho, and K. Geras. The break-even point on optimization trajectories of deep neural networks. In *The* 8^{th} *International Conference*

on Learning Representations, 2020.

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.

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.

K. Ji, M. Liu, Y. Liang, and L. Ying. Will bilevel optimizers benefit from loops. *arXiv preprint arXiv:2205.14224*, 2022.

R. Jiang, N. Abolfazli, A. Mokhtari, and E. Y. Hamedani. A conditional gradient-based method for simple bilevel optimization with convex lower-level problem. In *International Conference on Artificial Intelligence and Statistics*, pages 10305–10323. PMLR, 2023.

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.

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.

D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

M. Kodryan, E. Lobacheva, M. Nakhodnov, and D. P. Vetrov. Training scale-invariant neural networks on the sphere can happen in three regimes. *Advances in Neural Information Processing Systems*, 35:14058–14070, 2022.

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.

S. Kolyada. Li-Yorke sensitivity and other concepts of chaos. *Ukrainian Mathematical Journal*, 56 (8), 2004.

L. Kong and M. Tao. Stochasticity of deterministic gradient descent: Large learning rate for multiscale objective function. *Advances in Neural Information Processing Systems*, 33:2625–2638, 2020.

Y. Kou, Z. Chen, and Q. Gu. Implicit bias of gradient descent for two-layer relu and leaky relu networks on nearly-orthogonal data. *arXiv preprint arXiv:2310.18935*, 2023.

I. Kreisler, M. S. Nacson, D. Soudry, and Y. Carmon. Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond. In *Proceedings of the* 40^{th} *International Conference on Machine Learning*, 2023.

J. Kwon, D. Kwon, S. Wright, and R. D. Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages 18083–18113. PMLR, 2023.

G. Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

A. Lasota and M. C. Mackey. *Chaos, fractals, and noise: stochastic aspects of dynamics*, volume 97. Springer Science & Business Media, 1998.

V. Lebedev and S. Finogenov. Ordering of the iterative parameters in the cyclical Chebyshev iterative method. *USSR Computational Mathematics and Mathematical Physics*, 11(2):155–170, 1971.

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.

A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, and G. Gur-Ari. The large learning rate phase of deep learning: The catapult mechanism. *preprint arXiv:2003.02218*, 2020.

J. Li, B. Gu, and H. Huang. A fully single loop algorithm for bilevel optimization without Hessian inverse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7426–7434, 2022a.

J. Li, F. Huang, and H. Huang. Local stochastic bilevel optimization with momentum-based variance reduction. *arXiv preprint arXiv: 2205.01608*, 2022b.

J. Li, X. Chen, S. Ma, and M. Hong. Problem-parameter-free decentralized nonconvex stochastic optimization. *arXiv preprint arXiv:2402.08821*, 2024.

T.-Y. Li and J. A. Yorke. Period three implies chaos. *The American Mathematical Monthly*, 82(10): 985–992, 1975.

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.

S. H. Lim, Y. Wan, and U. Simsekli. Chaotic regularization and heavy-tailed limits for deterministic gradient descent. *Advances in Neural Information Processing Systems*, 35:26590–26602, 2022.

T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020a.

T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020b.

B. Liu, M. Ye, S. Wright, P. Stone, and Q. Liu. BOME! Bilevel Optimization Made Easy: A Simple First-Order Approach. *Advances in Neural Information Processing Systems*, 35:17248–17262, 2022.

H. Liu, K. Simonyan, and Y. Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations*, 2019.

R. Liu, Y. Liu, S. Zeng, and J. Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34:8662–8675, 2021.

R. Liu, Y. Liu, W. Yao, S. Zeng, and J. Zhang. Averaged method of multipliers for bi-level optimization without lower-level strong convexity. *arXiv preprint arXiv:2302.03407*, 2023.

E. Lobacheva, M. Kodryan, N. Chirkova, A. Malinin, and D. P. Vetrov. On the periodic behavior of neural network training with batch normalization and weight decay. In *Advances in Neural Information Processing Systems*, 2021.

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.

K. Lyu, Z. Li, and S. Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35:34689–34708, 2022.

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.

J. Milnor. Remarks on iterated cubic maps. *Experimental Mathematics*, 1(1):5–24, 1992.

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

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.

Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.

H. E. Nusse. Asymptotically periodic behaviour in the dynamics of chaotic mappings. *SIAM Journal on Applied Mathematics*, 47(3):498–515, 1987.

E. Ott. *Chaos in dynamical systems*. Cambridge university press, 2002.

S. Oymak. Provable super-convergence with a large cyclical learning rate. *IEEE Signal Processing Letters*, 28:1645–1649, 2021.

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.

B. A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural computation*, 6(1):147–160, 1994.

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

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

Q. Qian, S. Zhu, J. Tang, R. Jin, B. Sun, and H. Li. Robust optimization over multiple domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4739–4746, 2019.

S. Qiu, Z. Yang, X. Wei, J. Ye, and Z. Wang. Single-timescale stochastic nonconvex-concave optimization for smooth nonlinear TD-learning. *arXiv preprint arXiv:2008.10103*, 2020.

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

A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

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.

S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR, 2016.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

T. D. Rogers and D. C. Whitley. Chaos in the cubic mapping. *Mathematical Modelling*, 4(1):9–25, 1983.

C. Rommel, T. Moreau, J. Paillard, and A. Gramfort. Cadda: Class-wise automatic differentiable data augmentation for eeg signals. In *ICLR 2022-International Conference on Learning Representations*, 2022.

H. Shen and T. Chen. On penalty-based bilevel gradient descent method. *arXiv preprint arXiv:2302.05185*, 2023.

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.

D. Singer. Stable orbits and bifurcation of maps of the interval. *SIAM Journal on Applied Mathematics*, 35(2):260–267, 1978.

H. Skjolding, B. Branner-Jørgensen, P. L. Christiansen, and H. E. Jensen. Bifurcations in discrete dynamical systems with cubic maps. *SIAM Journal on Applied Mathematics*, 43(3):520–534, 1983.

L. N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

J. Sohl-Dickstein. The boundary of neural network trainability is fractal. *arXiv preprint arXiv:2402.06184*, 2024.

M. Song and C. Yun. Trajectory alignment: Understanding the edge of stability phenomenon via bifurcation theory. *Advances in Neural Information Processing Systems*, 36, 2024.

D. Sow, K. Ji, Z. Guan, and Y. Liang. A constrained optimization approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022a.

D. Sow, K. Ji, and Y. Liang. On the convergence theory for hessian-free bilevel algorithms. *Advances in Neural Information Processing Systems*, 35:4136–4149, 2022b.

D. Sow, K. Ji, and Y. Liang. On the convergence theory for hessian-free bilevel algorithms. In *Advances in Neural Information Processing Systems*, 2022c.

G. W. Stewart. *Matrix algorithms: volume 1: basic decompositions*. SIAM, 1998.

S. H. Strogatz. *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

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.

D. A. Tarzanagh, M. Li, C. Thrampoulidis, and S. Oymak. Fednest: Federated bilevel, minimax, and compositional optimization. *arXiv preprint arXiv:2205.02215*, 2022.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

I. Tsaknakis, P. Khanduri, and M. Hong. An implicit gradient-type method for linearly constrained bilevel problems. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5438–5442. IEEE, 2022.

K. Van Den Doel and U. Ascher. The chaotic nature of faster gradient descent methods. *Journal of Scientific Computing*, 51:560–581, 2012.

J. Wang, T. Zhang, S. Liu, P.-Y. Chen, J. Xu, M. Fardad, and B. Li. Adversarial attack generation empowered by min-max optimization. *Advances in Neural Information Processing Systems*, 34: 16020–16033, 2021.

M. Wang, E. X. Fang, and H. Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161:419–449, 2017.

Y. Wang, M. Chen, T. Zhao, and M. Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *The* 10^{th} *International Conference on Learning Representations*, 2022.

J. Wu, V. Braverman, and J. D. Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *preprint arXiv:2305.11788*, 2023.

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.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Q. Xiao, H. Shen, W. Yin, and T. Chen. Alternating projected sgd for equality-constrained bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 987–1023. PMLR, 2023.

T. Xiao, K. Balasubramanian, and S. Ghadimi. A projection-free algorithm for constrained stochastic multi-level composition optimization. In *Advances in Neural Information Processing Systems*, volume 35, pages 19984–19996, 2022.

C. Xu, X. Wang, Z. Zheng, and Z. Cai. Stability and bifurcation of collective dynamics in phase oscillator populations with general coupling. *Physical Review E*, 103(3):032307, 2021.

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.

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.

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

S. Yang, M. Wang, and E. X. Fang. Multilevel stochastic gradient methods for nested composition optimization. *SIAM Journal on Optimization*, 29(1):616–659, 2019.

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

B. Yuan, Y. He, J. Davis, T. Zhang, T. Dao, B. Chen, P. S. Liang, C. Re, and C. Zhang. Decentralized training of foundation models in heterogeneous environments. *Advances in Neural Information Processing Systems*, 35:25464–25477, 2022.

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

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

J. Zhang, H. Li, S. Sra, and A. Jadbabaie. Neural network weights do not converge to stationary points: An invariant measure perspective. In *Proceedings of the* 39^{th} *International Conference on Machine Learning*, pages 26330–26346, 2022a.

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.

Y. Zhang, Y. Yao, P. Ram, P. Zhao, T. Chen, M. Hong, Y. Wang, and S. Liu. Advancing model pruning via bi-level optimization. In *Advances in Neural Information Processing Systems*, 2022b.

S. Zhao and Y. Liu. Numerical methods for distributed stochastic compositional optimization problems with aggregative structure. *arXiv preprint arXiv:2211.04532*, 2022.

L. Zhu, C. Liu, A. Radhakrishnan, and M. Belkin. Catapults in SGD: Spikes in the training loss and their impact on generalization through feature learning. *preprint arXiv:2306.04815*, 2023a.

L. Zhu, C. Liu, A. Radhakrishnan, and M. Belkin. Quadratic models for understanding neural network dynamics. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PvJnX3dwsD.

X. Zhu, Z. Wang, X. Wang, M. Zhou, and R. Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *The* 11^{th} *International Conference on Learning Representations*, 2023b.

L. Ziyin, B. Li, J. B. Simon, and M. Ueda. SGD with a Constant Large Learning Rate Can Converge to Local Maxima. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9XhPLAjjRB.