UC Davis
UC Davis Electronic Theses and Dissertations

Title
Towards Computational and Sample Efficiency in Stochastic Optimization

Permalink
https://escholarship.org/uc/item/59w1k62c

Author
Xiao, Tesi

Publication Date
2023

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Towards Computational and Sample Efficiency in Stochastic Optimization

By

TESI XIAO
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Krishnakumar Balasubramanian, Chair

Miles Lopes

Xiaodong Li

Committee in Charge

2023

© Tesi Xiao, 2023. All rights reserved.

To my beloved family

Contents

Abstract v

Acknowledgments vi

Chapter 1. Overview of The Dissetation 1

1.1. Preliminaries 1

1.2. Organization 7

Chapter 2. Proximal Averaged Stochastic Approximation 9

2.1. Introduction 9

2.2. Methodology 11

2.3. Convergence Analysis 12

2.4. Discussion and Conclusion 25

Chapter 3. Decentralized Proximal Averaged Stochastic Approximation 27

3.1. Introduction 27

3.2. Methodology 31

3.3. Convergence Analysis 34

3.4. Experiments 39

3.5. Discussion and Conclusion 42

Chapter 4. Conditional Gradient-Based Nested Averaged Stochastic Approximation 43

4.1. Introduction 43

4.2. Methodology 50

4.3. Convergence Analysis 50

4.4. Numerical Experiments for $T = 1$ 57

4.5. Discussion and Conclusion 58

Chapter 5. Stochastic Conditional Gradient Methods under Interpolation-like Conditions 60

5.1. Introduction 60
5.2. Preliminaries and Assumptions 64
5.3. Improved Complexities for Stochastic Conditional Gradient Methods 68
5.4. Experiments 72
5.5. Discussion and Conclusion 73

Appendix A. Appendix of Chapter 3 74
A.1. Experimental Details 74
A.2. Proof of Theorem 3.1 74
A.3. Discussion on Different Types of Consensus Errors 90

Appendix B. Appendix of Chapter 4 92
B.1. Technical Lemmas 92
B.2. Proof of Theorem 4.1 93
B.3. Proofs for Section 4.3.1 103
B.4. Proof of Theorem 4.3 110

Appendix C. Appendix of Chapter 5 114
C.1. Proof for Theorem 5.1 114
C.2. Proof of Theorem 5.2 117
C.3. Zeroth-order SGD under Growth Conditions 121

Bibliography 124

Towards Computational and Sample Efficiency in Stochastic Optimization

Abstract

Stochastic optimization is a crucial tool in machine learning, statistics, and operations research, and developing efficient algorithms for stochastic optimization is of great importance. This dissertation focuses on stochastic composite optimization, where the objective function is composed of a smooth expected value function and a deterministic non-smooth component. We propose a class of algorithms called proximal averaged stochastic approximation (Prox-ASA), which estimates the gradient using a moving average approach. We prove the theoretical convergence of Prox-ASA to a first-order stationary point in different settings, including expectation, high probability, and almost surely asymptotically. In addition, we show that Prox-ASA can be applied to address decentralized problems and stochastic compositional optimization problems. For the non-convex constrained setting with expensive projection, we propose a novel class of conditional gradient based algorithms for solving stochastic multi-level compositional optimization problems that obtain the same sample complexity of the single-level setting under standard assumptions. Lastly, we demonstrate that by leveraging interpolation-like conditions satisfied by overparameterized models, we can improve the oracle complexities of stochastic conditional gradient methods.

Acknowledgments

I am immensely grateful to my advisor, Prof. Krishnakumar Balasubramanian, without whom this dissertation would not have been possible. Krishna has been an outstanding Ph.D. advisor, providing me with invaluable guidance and support throughout my journey. His availability to discuss my research, constructive feedback, and constant push to help me achieve my full potential have been exemplary. I have been inspired to delve deeper into my research and develop a strong foundation in my field by his expertise, knowledge, and passion for the subject. Krishna's support has extended beyond just research as he has provided me with opportunities to attend conferences, workshops, and other academic events that have allowed me to expand my knowledge and network with other researchers in my field. Moreover, Krishna has always been patient, kind, and understanding during difficult times in our general life outside of research. Overall, I consider myself extremely fortunate to have had the opportunity to work with Krishna over the past five years.

I would like to extend my thanks to the esteemed members of my dissertation committee, Prof. Miles Lopes, Prof. Xiaodong Li, Prof. Bala Rajaratnam, and Prof. Thomas Strohmer, for their invaluable contributions to my research. Their time, effort, and expertise in evaluating my work have been immensely beneficial in helping me refine and improve my dissertation. The feedback, constructive criticism, and thought-provoking questions they posed during the defense have been instrumental in shaping my research and enhancing its quality. I am grateful for their critical insights and guidance throughout the dissertation process. Thank you for helping me achieve this significant milestone in my academic journey.

I would also like to express my gratitude to the outstanding faculty members and friendly staff in the Department of Statistics at UC Davis. In particular, I would like to thank Prof. Hans Mueller, who is my initial advisor in the first year and the instructor of *Generalized Linear Models*, for his invaluable guidance and expertise. I want to thank Prof. Miles Lopes again for his inspiring lectures in *High-dimensional Statistics*. Furthermore, I would like to extend my sincere gratitude to several faculty members and staff, including Prof. Debashis Paul, Prof. Wolfgang Polonik, Prof. Hao Chen, Prof. Jiming Jiang, Prof. Thomas Lee, Prof. Prabir Burman, Prof. Jie Peng, Prof. Ethan Anderes, Prof. Alexander Aue, Prof. Shizhe Chen, Pete Scully, Sarah Driver, Andi Carr, and many others. Their unwavering support have been crucial to my academic growth and success.

Furthermore, I would like to thank many great research collaborators. I would like to acknowledge the significant contributions of Prof. Saeed Ghadimi to my research projects in the field of optimization. His expertise and guidance have been invaluable in shaping the direction of my research and in achieving successful outcomes. In addition, I would like to thank Dr. Xuanqing Liu and Prof. Cho-Jui Hsieh for offering me stimulating research opportunities during my undergraduate studies, which encouraged me to pursue a doctoral degree. I would also like to express my gratitude to Dr. Xia Xiao, Ming Chen, and Youlong Chen for the valuable research internship at ByteDance. Their mentorship and guidance have been instrumental in helping me gain practical experience in recommendation systems and neural architecture search. Besides, I want to thank my great mentors during my internship at Amazon, Dr. Branislav Kveton, Dr. Sumeet Katariya, Dr. Tanmay Gangwani, and Dr. Anshuka Rangi. Their expertise and knowledge have expanded my understanding of reinforcement learning and have helped me to develop new skills in the field of information retrieval.

Meeting amazing peers at Davis has been a source of immense gratitude for me. Jiaxiang Li, in particular, has been a part of my life since the very first day of my undergraduate studies, and we have been roommates at Davis for three years. Jiaxiang is not only an exceptional research peer but also one of my best friends who has always been kind and supportive. Additionally, I would like to express my appreciation to all my friends here, such as Ye He, Xuxing Chen, Shouwei Hui, Abhshek Roy, and many others. It has been fortunate for me to have these friends through the Ph.D. journey.

Finally, I wish to convey my profound gratitude to my parents for their constant support and understanding throughout my journey. Moreover, I want to express my heartfelt appreciation to my wife, Zhuo Ren, for her unwavering support, encouragement, and love. Zhuo has always been there to celebrate my accomplishments, comfort me during challenging times, and inspire me to pursue my aspirations. I am forever thankful for her presence in my life.

CHAPTER 1

Overview of The Dissetation

Stochastic optimization is a branch of mathematical optimization that deals with optimizing a function that involves random variables, which is widely used in machine learning, statistics, and operations research. In this dissertation, we consider the following stochastic optimization problem:

$$\min_{x\in\mathbb{R}^d} \quad \{\Phi(x) = F(x) + \Psi(x)\}, \qquad F(x) = \mathbb{E}_{\xi\sim\mathcal{D}}[G(x,\xi)] \tag{1.1}$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable function and $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a simple but possibly non-smooth function. Moreover, the function $F(x)$ is an expected-valued function in the form of $F(x) = \mathbb{E}_{\xi\sim\mathcal{D}}[G(x,\xi)]$, where the expectation is taken over the random vector $\xi \in \Xi$ with an underlying distribution denoted by $\mathcal{D}$. We aim to propose iterative algorithms that utilize the information of $G(x,\xi)$ to solve (1.1), assuming access to a sampling oracle of the random vector ξ. In designing such algorithms, it is crucial to consider both the computational cost and the sample complexity. To this end, this dissertation presents various theoretical contributions to the computational and sample efficiency in the field of stochastic optimization.

1.1. Preliminaries

In this section, we introduce several preliminaries to establish the foundation for this dissertation.

1.1.1. Notations. All vectors considered in this dissertation are in Euclidean space. $\|\cdot\|$ denotes the ℓ_2-norm for vectors and Frobenius norm for matrices. $\|\cdot\|_2$ denotes the spectral norm for matrices. $\mathbf{1}$ represents the all-one vector, and $\mathbf{I}$ is the identity matrix as a standard practice. For an extended valued function $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, its effective domain is written as $\mathrm{dom}(\Psi) = \{x \mid \Psi(x) < +\infty\}$. A function Ψ is said to be proper if $\mathrm{dom}(\Psi)$ is non-empty. For any proper closed convex function Ψ, $x \in \mathbb{R}^d$, the proximal operator is defined as

$$\mathbf{prox}_{\Psi}(x) = \arg\min_{y\in\mathbb{R}^d} \left\{\frac{1}{2}\|y-x\|^2 + \Psi(y)\right\}. \tag{1.2}$$

For $x, z \in \mathbb{R}^d$ and $\gamma > 0$, the proximal gradient mapping of z at x is defined as

$$\mathcal{G}(x, z, \gamma) = \frac{1}{\gamma}\left(x - \mathbf{prox}_{\Psi}^{\gamma}(x - \gamma z)\right). \tag{1.3}$$

For any convex and compact set $\mathcal{X} \subset \mathbb{R}^d$, we define the indicator function as follows

$$\mathbb{1}_{\{x \in \mathcal{X}\}} = \begin{cases} 0 & \text{if } x \in \mathcal{X}, \\ +\infty & \text{if } x \notin \mathcal{X}. \end{cases} \tag{1.4}$$

Its corresponding proximal operator is the orthogonal projection onto the set $\mathcal{X}$, which is denoted as

$$\mathbf{proj}_{\mathcal{X}}(x) = \underset{y \in \mathbb{R}^d}{\arg\min}\, \|y - x\|^2. \tag{1.5}$$

All random objects are properly defined in a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and write $x \in \mathcal{H}$ if x is $\mathcal{H}$-measurable given a sub-σ-algebra $\mathcal{H} \subseteq \mathscr{F}$ and a random vector x. We use $\sigma(\cdot)$ to denote the σ-algebra generated by all the argument random vectors.

1.1.2. Function Class.

We present the formal definition of various function classes that will be discussed in the dissertation.

Definition 1.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous differentiable function. Then, we say*

(i) $f(x)$ is convex if and only if

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y \in \mathbb{R}^d.$$

(ii) $f(x)$ is μ-strongly convex ($\mu > 0$) if and only if

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

(iii) $f(x)$ is L_f-Lipschitz continuous ($L_f > 0$) if and only if

$$|f(x) - f(y)| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

(iv) $f(x)$ is $L_{\nabla f}$-smooth if and only if $\nabla f(x)$ is $L_{\nabla f}$-Lipschitz continuous, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq L_{\nabla f}\|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

Equivalently, $f(x)$ is $L_{\nabla f}$-smooth if and only if

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L_{\nabla f}}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

1.1.3. Algorithm. Next, we present two fundamental classes of algorithms that this dissertation builds upon and that have been extensively studied in the literature.

Gradient Descent-Type Methods. One of the most widely used optimization algorithms to solve (1.1) when $\Psi(x) = 0$ is Stochastic Gradient Descent (SGD), which is based on the idea of iteratively moving in the direction of an estimator of the negative gradient of the objective function to approach the minimum. Specifically, at each iteration $k \in \mathbb{N}$, (mini-batch) SGD finds the next iterate x^{k+1} based on the current iterate x^k and gradient estimator z^k:

$$z^k = \frac{1}{|\mathcal{B}_k|} \sum_{\xi \in \mathcal{B}_k} \nabla G(x, \xi), \tag{1.6}$$

$$x^{k+1} = x^k - \gamma_k z^k, \tag{1.7}$$

where $\gamma_k > 0$ is the step size and $\mathcal{B}_k = \{\xi_1^k, \ldots, \xi_{|\mathcal{B}_k|}^k\}$ is a batch of samples used to evaluate z^k. For any general proximable $\Psi(x)$, a natural extension of SGD is called stochastic proximal gradient descent (Prox-SGD), in which the update rule (1.7) at the k-th iteration is replaced by

$$x^{k+1} = \mathbf{prox}_{\gamma_k \Psi}(x^k - \gamma_k z^k). \tag{1.8}$$

For the constrained case where $\Psi(x) = \mathbb{1}_{\{x \in \mathcal{X}\}}$ for a compact and convex set $\mathcal{X} \subset \mathbb{R}^d$, the update rule in (1.8) yields

$$x^{k+1} = \mathbf{proj}_{\mathcal{X}}(x^k - \gamma_k z^k), \tag{1.9}$$

which corresponds to the projected SGD for solving the constrained problem.

Conditional Gradient-Type Methods. The Conditional Gradient method, also known as Frank-Wolfe (FW) method, was proposed first by [**FW56**] to solve the constrained optimization problem. It has obtained renewed interest in the machine learning and optimization communities due to their projection-free nature [**Jag13**]. Its stochastic variants were also proposed and analyzed subsequently [**HL16**]. Unlike the update rule in (1.9) that uses a projection step to satisfy the constraints, the conditional gradient method finds the next iterate in the constraint set by

$$d^k = \operatorname*{arg\,min}_{d \in \mathcal{X}} \left\langle d^k - x^k, z^k \right\rangle, \tag{1.10}$$

$$x^{k+1} = x^k + \gamma_k (d^k - x^k). \tag{1.11}$$

The step size $\gamma_k \in (0, 1)$ guarantees that $x^{k+1} \in \mathcal{X}$ if $x^k \in \mathcal{X}$. The conditional gradient-type algorithms are more favorable than projection-based algorithms when computing d^k in (1.10) is much more efficient than solving the projection step; see Table 1 in [**Jag13**].

1.1.4. Complexity. To analyze the complexity of iterative algorithms designed for solving stochastic optimization problems, as described in (1.1), it is important to take into account not only the iteration complexity (the number of iterations required to obtain a solution) and per-iteration complexity (the computational complexity for each iteration), but also the sample complexity (the number of samples needed to obtain a solution). In particular, we consider the following types of oracles in this dissertation.

- Proximal Oracle (PO): Given $x \in \mathbb{R}^d$ and a proper convex and closed function $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, we say a procedure is a *Proximal Oracle* if it computes the proximal mapping of x:

$$\min_{y \in \mathbb{R}^d} \left\{ \frac{1}{2} \|y - x\|^2 + \Psi(y) \right\}, \tag{1.12}$$

  When $\Psi = \mathbb{1}_{\{x \in \mathcal{X}\}}$, then PO computes the orthogonal projection of x onto $\mathcal{X}$.
- Linear Minimization Oracle (LMO): Given $z \in \mathbb{R}^d$ and a convex and compact set $\mathcal{X} \subset \mathbb{R}^d$, we say a procedure is a *Linear Minimization Oracle* if it computes the solution of the following problem:

$$\min_{d \in \mathcal{X}} \langle d, z \rangle. \tag{1.13}$$

The Proximal Oracle is computationally efficient for some special cases. For example, when $\Psi(x) = \|x\|_1$, the corresponding proximal mapping has the following analytical form:

$$[\mathbf{prox}_{\|x\|_1}(x)]_i = \begin{cases} x_i - 1, & \text{if } x > 1, \\ x_i + 1, & \text{if } x < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1.14}$$

This operator is also known as the soft thresholding operator [**BL08**]. When $\Psi = \mathbb{1}_{\{x \in \mathcal{X}\}}$ for a convex and compact set $\mathcal{X} \subset \mathbb{R}^d$, then PO computes the orthogonal projection of x onto $\mathcal{X}$. In some specific scenarios, such as when $\mathcal{X}$ represents a trace norm ball of matrices, the projection operator may not be as computationally efficient as the Linear Minimization Oracle. While the projection onto the trace norm ball requires the full singular value decomposition, LMO only calculates the top eigenvalue (or singular value) using the standard Lanczos' algorithm [**Jag13**]. We also introduce the Stochastic First-Order Oracle and Stochastic Zeroth-Order Oracle, which algorithms in the dissertation build upon.

- Stochastic First-Order Oracle (SFO): Given a function $G(x, \xi)$, $x \in \mathbb{R}^d$, and $\xi \in \Xi$, we say a procedure is a *Stochastic First-Order Oracle* if it computes the gradient of G w.r.t. x, i.e.,

$$\nabla G(x, \xi).$$

- Stochastic Zeroth-Order Oracle (SZO): Given a function $G(x, \xi)$, $x \in \mathbb{R}^d$, and $\xi \in \Xi$, we say a procedure is a *Stochastic First-Order Oracle* if it computes the function value:

$$G(x, \xi).$$

In training artificial neural networks, the forward-backward pass provides an excellent example of calling SZO and SFO methods for the loss function; see Figure 1.1. Designing algorithms with fewer SFO and SZO becomes crucial as the time and space complexities for feedforward passes and backward propagation increase significantly with deeper and wider neural networks.

1.1.5. Concentration Inequality. To ensure completeness, we offer a brief overview of sub-Gaussian and sub-exponential random variables, which serve as the fundamentals for deriving high-probability outcomes.

Definition 1.2. (Sub-gaussian and Sub-exponential)

(a) A random variable X is K-sub-gaussian if there exists $K > 0$ such that $\mathbb{E}[\exp(X^2/K^2)] \leq 2$. The sub-gaussian norm of X, denoted $\|X\|_{\psi_2}$, is defined to be the smallest K. That is to say,

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}[\exp(X^2/t^2)] \leq 2\right\}.$$

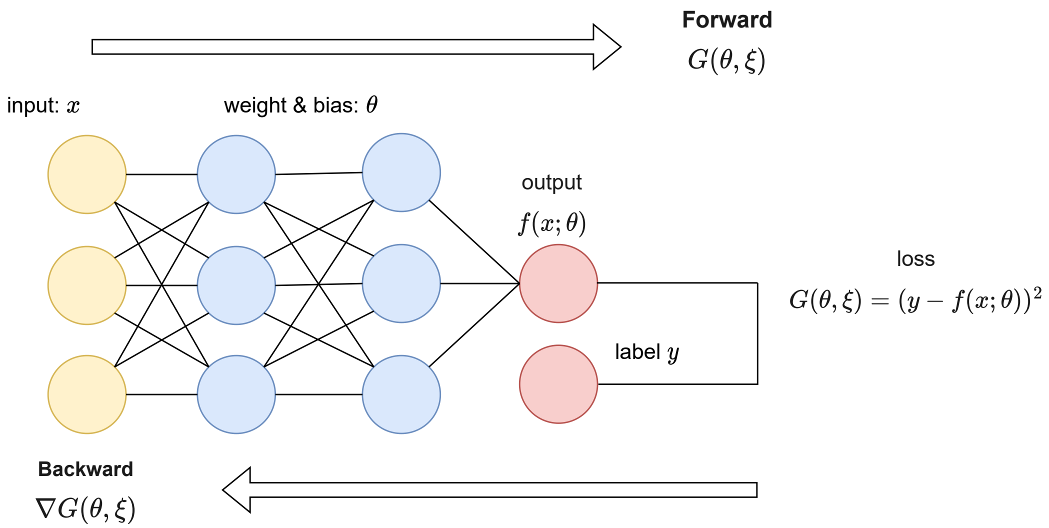


FIGURE 1.1. The feedforward pass and backward propagation in a fully Connected feed-forward network. With a slight abuse of notation, we denote the optimization variable as θ and the data sample as $\xi = (x, y)$.

(b) A random variable X is K-sub-exponential if there exists $K > 0$ such that $\mathbb{E}[\exp(|X|/K)] \leq 2$. The sub-exponential norm of X, denoted $\|X\|_{\psi_1}$, is defined to be the smallest K. That is to say,

$$\|X\|_{\psi_1} = \inf\left\{t > 0 : \mathbb{E}[\exp(|X|/t)] \leq 2\right\}.$$

The above characterization is based on the so-called orlicz norm of a random variable. There are equivalent definitions of sub-gaussian and sub-exponential random variables. We refer readers to Proposition 2.5.2 and Proposition 2.7.1 in [**Ver18**]. In particular, we will also use another definition of sub-gaussian random variables based on the moment-generating function given below.

LEMMA 1.1. (Sub-gaussian M.G.F. [**Ver18**]) *If a random variable X is K-sub-gaussian with $\mathbb{E}[X] = 0$, then $\mathbb{E}[\exp(\lambda X)] \leq \exp(c\lambda^2 K^2)$ $\forall \lambda \in \mathbb{R}$, where cx is a absolute constant.*

The following two lemmas are essential in our proof.

LEMMA 1.2. (Sub-exponential is sub-gaussian squared [**Ver18**]) *A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover, $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$.*

LEMMA 1.3. (Generalized Freedman-type Inequality [**HLPR19**]) *Let $(\Omega, \mathscr{F}, (\mathscr{F}_i), P)$ be a filtered probability space, (X_i) and (K_i) be adapted to $(\mathscr{F}_i)$, and $n \in \mathbb{N}$. Suppose for all $i \in [n]$, $K_{i-1} \geq 0$,*

$\mathbb{E}[X_i|\mathscr{F}_{i-1}] = 0$, *and* $\mathbb{E}\left[\exp(\lambda X_i)|\mathscr{F}_{i-1}\right] \le \exp(\lambda^2 K_i^2)$. *Then for any* $t, b \ge 0, a > 0$,

$$\Pr\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k} X_i \ge t \text{ and } 2\sum_{i=1}^{k} K_{i-1}^2 \le a\sum_{i=1}^{k} X_i + b\right\}\right) \le \exp\left(-\frac{t}{4a + 8b/t}\right). \tag{1.15}$$

1.2. Organization

We start with Chapter 2 by investigating a class of algorithms referred to as proximal averaged stochastic approximation (Prox-ASA), which uses a moving average approach to estimate the gradient in another sequence. We prove the theoretical convergence of Prox-ASA to a first-order stationary point in expectation, with high probability, and almost surely asymptotically under different conditions. Furthermore, it is worth noting that this algorithmic framework has the potential to address stochastic compositional optimization problems and decentralized problems. However, the specifics of these applications will be discussed in Chapter 3 and Chapter 4, respectively.

In Chapter 3, we extend Prox-ASA to solve decentralized optimization algorithms, where n agents work together to optimize the objective function. We propose a class of single-time scale algorithms that achieves the optimal sample complexity using constant batch sizes. Unlike prior work, our algorithms have comparable complexity without requiring large batch sizes, more complex per-iteration operations (such as double loops), or stronger assumptions. Our theoretical findings are supported by extensive numerical experiments, which demonstrate the superiority of our algorithms over previous approaches.

In Chapter 4, we extend Prox-ASA for solving non-convex constrained stochastic multi-level compositional optimization problems, where the objective function is a nested composition of T functions with only noisy evaluations of the functions and their gradients being available. Leveraging the technique of conditional gradient sliding, we propose the first class of projection-free algorithms that obtain the same sample complexity of the single-level setting under standard assumptions. Notably, the dependence of these complexity bounds on ϵ and T are separate in the sense that changing one does not impact the dependence of the bounds on the other. Moreover, our algorithm is parameter-free and does not require any (increasing) order of mini-batches to converge, unlike the common practice in the analysis of stochastic conditional gradient-type algorithms.

The last chapter of this dissertation is separate from the preceding topics. In Chapter 5, we study the convergence of stochastic conditional gradient methods for overparametrized models. We show that one could leverage the interpolation-like conditions satisfied by such models to obtain improved oracle complexities. Specifically, when the objective function is convex, we show that the conditional gradient method requires $\mathcal{O}(\epsilon^{-2})$ calls to the stochastic gradient oracle to find an ϵ-optimal solution. Furthermore, by including a gradient sliding step, we show that the number of calls reduces to $\mathcal{O}(\epsilon^{-1.5})$. We also establish similar improved results in the zeroth-order setting, where only noisy function evaluations are available. Notably, the above results are achieved without any variance reduction techniques, thereby demonstrating the improved performance of vanilla versions of conditional gradient methods for over-parametrized machine learning problems.

CHAPTER 2

Proximal Averaged Stochastic Approximation

2.1. Introduction

In this chapter, we investigate a class of proximal algorithms for solving the general non-convex regularized stochastic optimization problem:

$$\min_{x \in \mathbb{R}^d} \quad \{\Phi(x) = F(x) + \Psi(x)\}, \tag{2.1}$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is a continously differentiable function and $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a convex but possibly non-smooth function. In addition, the function $F(x)$ is an expected-valued function in the form of $F(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[G(x, \xi)]$, where the expectation is taken over the random vector ξ with an underlying distribution denoted by $\mathcal{D}$.

Since the analytical expression for $\nabla F(x)$ is often unknown, conventional gradient-based algorithms for solving deterministic problems are not applicable. To tackle stochastic optimization problems where $\Psi(x) = 0$, stochastic gradient descent (SGD) [**RM51**] serves as the foundation. SGD employs the stochastic gradient by invoking SFO over a single random sample ξ^{k+1}:

$$x^{k+1} = x^k - \gamma_k z^k, \quad z^k = \nabla G(x^k, \xi^{k+1}). \tag{2.2}$$

The presence of a regularizer $\Psi(x)$ generalizes the smooth optimization problem where $\Psi(x) = 0$, leading to numerous practical applications. For instance, the framework can be utilized for training *sparse* models by incorporating a non-smooth L_1 regularizer on the weights. This is done to compress models for deployment on memory-constrained devices, as outlined in [**LWK17**, **WWW$^+$16**]. To solve Problem (2.1), non-smoothness in $\Psi(x)$ can make the problem unsuitable for SGD, thereby necessitating the use of subgradient approaches. However, such methods can adversely impact convergence performance [**BM08**]. Fortunately, when Ψ has a special structure, the problem can be solved more efficiently. To be specific, if the proximal oracle for Ψ is computationally efficient, the

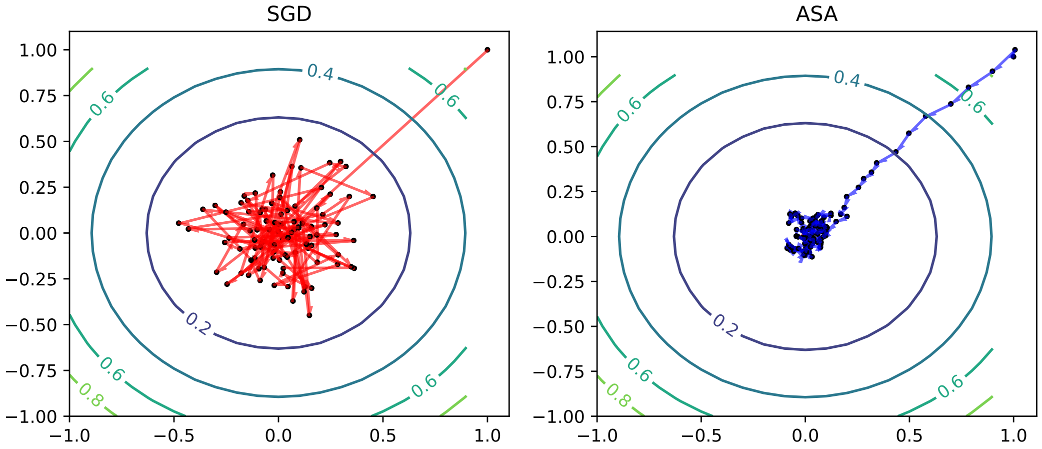


FIGURE 2.1. Level set plot of a quadratic function with the iterates of SGD (left) with constant stepsize and ASA (right). The central point is the solution x_*. Gaussian noises are added to obtain the stochastic gradients.

stochastic proximal gradient algorithm (Prox-SGD) can be utilized to solve the problem:

$$x^{k+1} = \mathbf{prox}_{\gamma_k \Psi}(x^k - \gamma_k z^k). \tag{2.3}$$

The literature contains extensive research on the convergence of proximal gradient algorithms for deterministic (non)-convex problems, as demonstrated by studies such as [**SRB11**, **CLK⁺12**, **HZSL13**, **KNS16**, **SYVS21**]. Furthermore, several recent contributions have focused on the stochastic setting. In particular, [**AFM17**, **RVV20**] provide proof of the convergence of Prox-SGG in the convex setting, while [**XJY19**, **GS21**] establish (non)-asymptotic analyses of Prox-SGD for minimizing general non-convex functions. Additionally, numerous studies have analyzed stochastic proximal gradient-type methods in the finite-sum problem [**SSZ12**, **SSZ14**, **Nit14**]; however, we do not delve into these studies since they are distinct from those in the stochastic setting.

The primary challenge associated with (Prox)-SGD is the presence of random noise. Merely having an unbiased estimator of the gradient $\nabla F(x)$ does not suffice to ensure convergence of the iterates. Figure 2.1 (left) illustrates this point by plotting the iterates of SGD with a constant step size used to minimize a quadratic function. As seen, the iterates of SGD do not converge to the solution and instead form a cluster of points around the solution. SGD fails to converge in this example because the stochastic gradients z^k do not converge to zero. This is in contrast to GD, where the algorithm terminates naturally as $\nabla F(x^k) \to 0$ when $x^k \to x_*$.

There are two classic techniques for addressing the variance in stochastic gradients: *diminishing stepsizes* [**RM51**] and *mini-batching*. However, tuning the sequence of decreasing stepsizes is challenging since the method may terminate early before reaching the solution or continue for excessive time. Moreover, the per-iteration cost increases with the batch size. A large batch size may prolong the duration of each iterate without updating optimization parameters frequently enough. Moreover, several modern variance-reduced methods are proposed for finite-sum problems to ensure that $\mathbb{E}[\|z^k - \nabla F(x^k)\|^2] \to 0$ as $k \to +\infty$, including [**SLRB17**, **DBLJ14**, **SSZ13**, **JZ13**]. In the stochastic setting, a commonly-used technique is aggregating past stochastic gradients [**Rus08**, **Xia09**, **GRW20**]. Specifically, [**GRW20**] propose the Averaged Stochastic Approximation (ASA) and prove its convergence to the first-order stationary point for non-convex objective functions. As shown in Figure 2.1, ASA exhibits stable convergence in contrast to SGD. In the subsequent sections, we extend ASA to solve general regularized non-convex stochastic optimization problems described in (2.1) and establish theoretical convergence to a first-order stationary point in expectation, with high probability, and almost surely.

2.2. Methodology

In this section, we present Algorithm 1 - **Prox**imal **A**veraged **S**tochastic **A**ppro-ximation (Prox-ASA), which leverages a common averaging technique in stochastic optimation [**Rus08**, **MHK18a**, **GRW20**] to reduce the variance of gradient estimation. In particular, the algorithm generates two sequences of variables, namely, the approximate solutions $\{x^k\}$ and approximate gradients $\{z^k\}$. We let $\mathscr{F}_0 = \emptyset$ and $\mathscr{F}_k$ be the σ-algebra generated by $\{x^1, z^1, \ldots, x^k, z^k\}$ for $k \geq 1$. The update rule for approximate gradients is given by

$$z^{k+1} = (1-\tau_k) z^k + \tau_k v^{k+1}, \quad \tau_k \in (0, 1], \tag{2.4}$$

where $\mathbb{E}[v^{k+1} | \mathscr{F}_k] = \mathbb{E}[\nabla G(x^k, \xi^{k+1})] = \nabla F(x^k)$. It is easy to observe that z^k is a biased estimator of the gradient that aggregates k stochastic gradients computed over the previous samples when $z^0 = 0$, i.e.,

$$z^k = \sum_{i=1}^{k} \alpha_i v^i, \qquad \text{where } \alpha_i = \tau_{i-1} \prod_{j=i}^{k-1} (1-\tau_j), \quad \sum_{i=1}^{k} \alpha_i = 1. \tag{2.5}$$

Algorithm 1 Proximal Averaged Stochastic Approximation (Prox-ASA)

```
**Input:** $z_0 = \mathbf{0}, \gamma, \{\tau_k\}_{\geq 0}, N$
**for** $k = 0, 1, \ldots, N-1$ **do**
    $y^k = \mathbf{prox}_{\gamma\Psi}\left(x^k - \gamma z^k\right)$
    $x^{k+1} = (1-\tau_k)x^k + \tau_k y^k$
    Obtain $\xi^{k+1}$ and compute the stochastic gradient $v^{k+1} = \nabla G(x^k, \xi^{k+1})$
    $z^{k+1} = (1-\tau_k)z^k + \tau_k v^{k+1}$
**end for**
```

Given the approximate gradient z^k for each k, the approximate solution x^k is updated as follows:

$$y^k = \mathbf{prox}_{\gamma\Psi}\left(x^k - \gamma z^k\right), \tag{2.6}$$

$$x^{k+1} = (1-\tau_k)x^k + \tau_k y^k. \tag{2.7}$$

The update rule above comprises two components: (i) a proximal gradient descent step in Eq. (2.6) that employs a biased gradient estimator and a constant stepsize, and (ii) a moving average step in Eq. (2.7) that is sometimes referred to as a relaxation step or a damped update. It is worth noting that Prox-ASA is a single time-scale algorithm that employs the same τ_k for updating both x^k and z^k. It is also possible to extend it to utilize two sets of weights $\{a\tau_k\}$ and $\{b\tau_k\}$ with a constant scaling factor $a, b > 0$; see [**GRW20**]. For simplicity, we employ the same weights throughout the sequel to establish the convergence results.

2.3. Convergence Analysis

2.3.1. Convergence Criteria. We first discuss the convergence criteria in the following analysis. Nonconvex optimization problems are NP-hard because finding a global minimum involves exploring a large search space with many local minima, local maxima, and saddle points. This makes it computationally infeasible to find an optimal solution in a reasonable amount of time, especially for problems with high-dimensional input spaces. This chapter's primary focus is to analyze an algorithm's effectiveness in discovering a first-order stationary point of (2.1).

Definition 2.1 (First-Order Stationary Point). *A point x_* is stationary point of $F(x) + \Psi(x)$ if*

$$x_* - \mathbf{prox}_{\Psi}(x_* - \nabla F(x_*)) = \mathbf{0},$$

i.e., $\mathbf{0} \in \nabla F(x_*) + \partial\Psi(x_*)$.

Empirical evidence suggests that an (approximate) first-order stationary point can be highly effective in practice. For instance, deep neural networks often have a loss surface with numerous local minima. These minima are believed to have varying degrees of flatness that may play a role in generalization. To establish the non-asymptotic convergence results, we introduce the ϵ-first-order stationary point in which $\epsilon > 0$ measures the non-stationarity.

Definition 2.2 (ϵ-First-Order Stationary Point). *A point $\bar{x}$ is ϵ-stationary point of $F(x) + \Psi(x)$ if $\|\bar{x} - \mathbf{prox}_\Psi(\bar{x} - \nabla F(\bar{x}))\|^2 \leq \epsilon$*

It is worth noting that the stepsize does not play a role in the above definitions. This is due to the following fact that characterizes the relations between proximal gradient mappings defined under different stepsizes.

Lemma 2.1. *Let $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper convex and closed function. For any $x, z \in \mathbb{R}^d$ and $\gamma > 0$,*

$$\min(1, \gamma) \leq \frac{\left\|x - \mathbf{prox}_{\gamma\Psi}(x - \gamma z)\right\|}{\|x - \mathbf{prox}_\Psi(x - z)\|} \leq \max(1, \gamma), \tag{2.8}$$

Proof. Denote the subdifferential of $\Psi(x)$ as $\partial\Psi(x)$ and $y_+(\beta) := \mathbf{prox}_{\beta\Psi}(x - \gamma z)$ for simplicity. By the optimality condition, we have $\mathbf{0}$ is a subgradient of $H(y) = \langle z, y - x\rangle + \frac{1}{2\beta}\|y - x\|^2 + \Psi(y)$ at $y_+(\beta)$, i.e.,

$$\mathbf{0} \in z + \beta^{-1}(y_+(\beta) - x) + \partial\Psi(y_+(\beta)).$$

Hence, there exists a subgradient of $\Psi(y)$ at $y_+(\beta)$, denoted by $\tilde{\nabla}\Psi(y_+(\beta))$, such that

$$\tilde{\nabla}\Psi(y_+(\beta)) = -z - \beta^{-1}(y_+(\beta) - x).$$

By the convexity of Ψ, we have for any $y \in \mathbb{R}^d$,

$$\Psi(y) - \Psi(y_+(\beta)) \geq \left\langle \tilde{\nabla}\Psi(y_+(\beta)), y - y_+(\beta) \right\rangle = \left\langle -z - \beta^{-1}(y_+(\beta) - x), y - y_+(\beta) \right\rangle,$$

Then, setting $y = y_+(\gamma), \beta = 1$ and $y = y_+(1), \beta = \gamma$ in the above inequality respectively, we obtain

$$\Psi(y_+(\gamma)) - \Psi(y_+(1)) \geq \langle -z - (y_+(1) - x), y_+(\gamma) - y_+(1)\rangle,$$

$$\Psi(y_+(1)) - \Psi(y_+(\gamma)) \geq \left\langle -z - \gamma^{-1}(y_+(\gamma) - x), y_+(1) - y_+(\gamma) \right\rangle.$$

Adding the above equalities together, we obtain

$$\big\langle (y_+(1)-x) - \gamma^{-1}(y_+(\gamma)-x), (y_+(1)-x) - (y_+(\gamma)-x) \big\rangle \le 0.$$

This implies that

$$(1+\gamma^{-1})\langle y_+(\gamma)-x, y_+(1)-x\rangle \ge \|y_+(1)-x\|^2 + \gamma^{-1}\left\|y_+(\gamma)-x\right\|^2,$$

Using the Cauchy-Schwartz inequality and rearranging the terms, we get

$$\left(\frac{\|y_+(\gamma)-x\|}{\|y_+(1)-x\|}\right)^2 - (1+\gamma)\frac{\|y_+(\gamma)-x\|}{\|y_+(\gamma)-x\|} + \gamma \le 0,$$

which is equivalent to

$$\left(\frac{\|y_+(\gamma)-x\|}{\|y_+(1)-x\|} - \gamma\right)\left(\frac{\|y_+(\gamma)-x\|}{\|y_+(1)-x\|} - 1\right) \le 0.$$

That is to say, $\frac{\|y_+(\gamma)-x\|}{\|y_+(1)-x\|}$ is between 1 and γ. □

2.3.2. Assumption. Next, we list and discuss the assumptions made in this work.

Assumption 2.1. *The functions $F(x)$ and $\Psi(x)$ satisfy:*

(1) $\Phi(x) = F(x) + \Psi(x) \ge \Phi_* > -\infty$ for all $x \in \mathbb{R}^d$.

(2) $F(x)$ is $L_{\nabla F}$-smooth.

(3) $\Psi(x)$ is proper, convex, and closed.

For stochastic oracles, we assume that the stochastic gradient $\nabla G(\cdot, \xi^{k+1})$ is unbiased conditioned on the filtration $\mathscr{F}_k$.

Assumption 2.2 (Unbiasness). *For any $k \ge 0, x \in \mathscr{F}_k$, and $1 \le i \le n$,*

$$\mathbb{E}\left[\nabla G(x, \xi^{k+1})\middle|\mathscr{F}_k\right] = \nabla F(x).$$

In addition, we consider three standard assumptions on the variance.

Assumption 2.3 (Bounded variance). *For any $k \ge 0, x \in \mathscr{F}_k$,*

$$\mathbb{E}\left[\left\|\nabla G(x, \xi^{k+1}) - \nabla F(x)\right\|^2\middle|\mathscr{F}_k\right] \le \sigma^2.$$

Assumption 2.4 (Bounded second-moment). *For any $k \geq 0, x \in \mathscr{F}_k$,*

$$\mathbb{E}\left[\left\|\nabla G(x, \xi^{k+1})\right\|^2 \middle| \mathscr{F}_k\right] \leq \sigma^2.$$

Assumption 2.5 (Sub-gaussian noise). *For any $k \geq 0, x \in \mathscr{F}_k$, $\left\|\nabla G(x, \xi^{k+1}) - \nabla F(x)\right\| |\mathscr{F}_k$ is K-sub-Gaussian.*

The unbiasedness and bounded variance assumptions (Assumption 2.2 and 2.3) are standard in the literature and also typically satisfied in several practical stochastic optimization problems [**Lan20**]. The assumption of the bounded second moment (Assumption 2.4), which implies the Lipschitz continuity of $F(x)$, is considerably stronger than Assumption 2.3. It is also straightforward to see that Assumption 2.3 together with the Lipschitz continuity of $F(x)$ imply Assumption 2.4. The Assumption 2.5 is commonly used in the literature to derive high-probability bounds; see [**HK14**, **HLPR19**, **LO20**, **ZCC$^+$18**]. It is worth highlighting that we assume the noise vector to be norm-sub-Gaussian rather than sub-Gaussian vectors to eliminate the dependence on dimensions in our bounds [**JNG$^+$19**]. It is also feasible to relax these assumptions for stochastic optimization in scenarios with heavy tails [**HM21**]; however, our present theory is based on the standard assumptions listed above.

2.3.3. Merit Function. Our proof relies on the merit function below:

$$W_\lambda(x^k, z^k) = \underbrace{\Phi(x^k) - \Phi_*}_{\text{function value gap}} + \underbrace{\Psi(x^k) - \eta(x^k, z^k)}_{\text{primal convergence}} + \lambda \underbrace{\left\|\nabla F(x^k) - z^k\right\|^2}_{\text{dual convergence}}, \tag{2.9}$$

where

$$\eta(x, z) = \min_{y \in \mathbb{R}^d} \left\{\langle z, y - x\rangle + \frac{1}{2\gamma}\|y - x\|^2 + \Psi(y)\right\} \tag{2.10}$$

Given that $y^k = \mathbf{prox}_{\gamma\Psi}\left(x^k - \gamma z^k\right)$ is the minimizer of a $1/\gamma$-strongly convex function, we have

$$\left\langle z^k, y^k - x^k\right\rangle + \frac{1}{2\gamma}\|y^k - x^k\|^2 + \Psi(y^k) \leq \Psi(x^k) - \frac{1}{2\gamma}\|y^k - x^k\|^2,$$

which implies $\Psi(x^k) - \eta(x^k, z^k) \geq \frac{1}{2\gamma}\left\|x^k - y^k\right\|^2 = \frac{1}{2\gamma}\left\|x^k - \mathbf{prox}_{\gamma_\Psi}\left(x^k - \gamma z^k\right)\right\|^2$. Consequently, the current merit function comprises three terms that constrain the gap in function value, the convergence of iterates in the primal space x^k, and the convergence of dual variables z^k to $\nabla F(x^k)$, respectively. The following lemma characterizes the smoothness of $\eta(\cdot,\cdot)$, which plays an important role in the subsequent analysis.

Lemma 2.2. *Let $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a closed proper convex function. Let $\eta(x, z)$ be the function defined in* (2.10). *Then, $\nabla\eta$ is C_γ-Lipschitz continuous where*

$$C_\gamma = 2\sqrt{\left(1+\frac{1}{\gamma}\right)^2 + \left(1+\frac{\gamma}{2}\right)^2}. \tag{2.11}$$

Proof. Recall that the Moreau envelope of a convex and closed function Ψ multiplied by a scalar γ is defined by

$$\mathrm{env}_{\gamma\Psi}(x) = \min_{y\in\mathbb{R}^d}\left\{\frac{1}{2\gamma}\|y-x\|^2 + \Psi(y)\right\},$$

and its gradient is given by $\nabla\mathrm{env}_{\gamma\Psi}(x) = \frac{1}{\gamma}(x - \mathbf{prox}_{\gamma\Psi}(x))$ where $\mathbf{prox}_{\gamma\Psi}(x) = \arg\min_{y\in\mathbb{R}^d}\left\{\frac{1}{2\gamma}\|y-x\|^2 + \Psi(y)\right\}$. Note that $\eta(x,z) = \mathrm{env}_{\gamma\Psi}(x-\gamma z) - \frac{\gamma}{2}\|z\|^2$. Therefore, the partial gradients of η are given by

$$\nabla_x\eta(x,z) = -z - \gamma^{-1}\left(\mathbf{prox}_{\gamma\Psi}(x-\gamma z) - x\right), \quad \nabla_z\eta(x,z) = \mathbf{prox}_{\gamma\Psi}(x-\gamma z) - x.$$

Hence, for any (x, z) and (x', z'),

$$\begin{aligned}\|\nabla\eta(x,z) - \nabla\eta(x',z')\| &\le \|\nabla_x\eta(x,z) - \nabla_x\eta(x',z')\| + \|\nabla_z\eta(x,z) - \nabla_z\eta(x',z')\| \\ &\le 2(1+1/\gamma)\|x-x'\| + (2+\gamma)\|z-z'\| \le C_\gamma\|(x,z)-(x',z')\|.\end{aligned}$$

□

2.3.4. Non-asymptotic Convergence. We state the non-asymptotic convergence results in this subsection. The first results show that the randomly selected iterate from the sequences generated by Prox-ASA is an ϵ-first-order stationary point in expectation.

Theorem 2.1 (Non-asymptotic convergence in expectation). *Suppose Assumption 2.1, 2.2, 2.3 hold. Let $\gamma > 0$, $\tau_k = \frac{1}{\sqrt{N}}$, R be a random integer uniformly from $\{1, \dots, N\}$. Then, for any sufficiently large $N \ge N_0(\gamma, L_{\nabla F})$, the sequences generated by Algorithm 1 satisfy*

$$\textit{(Primal)}\quad \mathbb{E}\left[\left\|y^R - x^R\right\|^2\right] \lesssim \frac{1}{\sqrt{N}} \qquad \textit{(Dual)}\quad \mathbb{E}\left[\left\|z^R - \nabla F(x^R)\right\|^2\right] \lesssim \frac{1}{\sqrt{N}},$$

which together imply that

$$\mathbb{E}\left[\left\|x^R - \mathbf{prox}_\Psi(x^R - \nabla F(x^R))\right\|^2\right] \lesssim \frac{1}{\sqrt{N}}$$

i.e., x^R and y^R is an ϵ-first-order stationary point in expectation if $N \gtrsim \epsilon^{-2}$. Furthermore, if Assumption 2.4 holds and $\Psi(x)$ is Lipschtiz continuous on its effective domain, then $N_0(\gamma, L_{\nabla F}) = 1$.

We also establish the high probability convergence results in the next theorem.

THEOREM 2.2 (Non-asymptotic convergence with high probabiltiy). *Suppose Assumption 2.1, 2.2, 2.5 hold. Let $\gamma > 0$, $\tau_k = \frac{1}{\sqrt{N}}$, R be a random integer uniformly from $\{1, \ldots, N\}$. Then, for any sufficiently large $N \geq N_0(\gamma, L_{\nabla F}, K)$, with probability $1 - \delta$, the sequences generated by Algorithm 1 satisfy*

$$\text{(Primal)} \quad \frac{1}{N}\sum_{k=1}^{N}\left\|y^k - x^k\right\|^2 \lesssim \frac{K^2\log(1/\delta)}{\sqrt{N}} \qquad \text{(Dual)} \quad \frac{1}{N}\sum_{k=1}^{N}\left\|z^k - \nabla F(x^k)\right\|^2 \lesssim \frac{K^2\log(1/\delta)}{\sqrt{N}},$$

which together imply that

$$\frac{1}{N}\sum_{k=1}^{N}\left\|x^k - \mathbf{prox}_\Psi(x^k - \nabla F(x^k))\right\|^2 \lesssim \frac{K^2\log(1/\delta)}{\sqrt{N}}.$$

2.3.5. Proof of Non-asymptotic Convergence. In this subsection, we present the proof of Theorem 2.1 and 2.2. First, the following lemma plays an essential role in our analysis that characterizes the decrease of the merit function.

LEMMA 2.3. *Suppose Assumption 2.1 holds. Let $W_\lambda(\cdot,\cdot)$ be the merit function defined in (2.9) for $\lambda > 0$. The sequences $\{x^k, z^k\}_{k\geq 0}$ generated by Algorithm 1 satisfy*

$$\begin{aligned} W_\lambda(x^{k+1}, z^{k+1}) - W_\lambda(x^k, z^k) &\leq \tau_k\left\{-\gamma^{-1} + \frac{(L_{\nabla F} + C_\gamma)\tau_k}{2} + \lambda L_{\nabla F}^2\right\}\left\|x^k - y^k\right\|^2 \\ &+ \tau_k\left\{-\lambda + \frac{C_\gamma\tau_k}{2}\right\}\left\|\nabla F(x^k) - z^k\right\|^2 + \frac{(C_\gamma + 2\lambda)\tau_k^2}{2}\left\|\Delta^{k+1}\right\|^2 + \tau_k r^{k+1}. \end{aligned} \tag{2.12}$$

where C_γ is defined in (2.11), $\Delta^{k+1} = v^{k+1} - \nabla F(x^k)$, and

$$r^{k+1} = \left\langle \Delta^{k+1}, x^k - y^k + [2\lambda(1-\tau_k) + C_\gamma\tau_k]\left(\nabla F(x^k) - z^k\right) + 2\lambda\left(\nabla F(x^{k+1}) - \nabla F(x^k)\right)\right\rangle \tag{2.13}$$

PROOF. By the smoothness of F (Assumption 2.1) and η (Lemma 2.2), we have

$$\begin{aligned} F(x^{k+1}) - F(x^k) &\leq \left\langle \nabla F(x^k), x^{k+1} - x^k \right\rangle + \frac{L_{\nabla F}}{2} \left\| x^{k+1} - x^k \right\|^2 \\ &= -\tau_k \left\langle \nabla F(x^k), x^k - y^k \right\rangle + \frac{L_{\nabla F} \tau_k^2}{2} \left\| x^k - y^k \right\|^2, \end{aligned} \tag{2.14}$$

and

$$\begin{aligned} \eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1}) &\leq \left\langle -z^k - \gamma^{-1}(y^k - x^k), x^k - x^{k+1} \right\rangle + \left\langle y^k - x^k, z^k - z^{k+1} \right\rangle \\ &\quad + \frac{C_\gamma}{2} \left(\left\| x^{k+1} - x^k \right\|^2 + \left\| z^{k+1} - z^k \right\|^2 \right) \\ &= 2\tau_k \left\langle z^k, y^k - x^k \right\rangle + \gamma^{-1}\tau_k \|x^k - y^k\|^2 + \tau_k \left\langle v^{k+1}, x^k - y^k \right\rangle \\ &\quad + \frac{C_\gamma}{2} \left(\tau_k^2 \left\| x^k - y^k \right\|^2 + \left\| z^{k+1} - z^k \right\|^2 \right). \end{aligned} \tag{2.15}$$

Since y^k is the minimizer of a $1/\gamma$-strongly convex function, i.e.,

$$\left\langle z^k, y^k - x^k \right\rangle + \frac{1}{2\gamma} \left\| y^k - x^k \right\|^2 + \Psi(y^k) \leq \Psi(x^k) - \frac{1}{2\gamma} \left\| y^k - x^k \right\|^2,$$

which together with (2.15) gives

$$\begin{aligned} \eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1}) \leq & -\gamma^{-1}\tau_k \left\| x^k - y^k \right\|^2 + \tau_k \left\langle v^{k+1}, x^k - y^k \right\rangle \\ & + 2\tau_k \left(\Psi(x^k) - \Psi(y^k) \right) + \frac{C_\gamma}{2} \left(\tau_k^2 \left\| x^k - y^k \right\|^2 + \left\| z^{k+1} - z^k \right\|^2 \right). \end{aligned} \tag{2.16}$$

By the convexity of Ψ, we have

$$\Psi(x^{k+1}) - \Psi(x^k) \leq (1 - \tau_k)\Psi(x^k) + \tau_k \Psi(y^k) - \Psi(x^k) = \tau_k \left(\Psi(y^k) - \Psi(x^k) \right). \tag{2.17}$$

Combining (2.14), (2.16), and (2.17), we have

$$\begin{aligned} &\left[\Phi(x^{k+1}) + \Psi(x^{k+1}) - \eta(x^{k+1}, z^{k+1}) \right] - \left[\Phi(x^k) + \Psi(x^k) - \eta(x^k, z^k) \right] \leq -\gamma^{-1}\tau_k \left\| x^k - y^k \right\|^2 \\ &\quad + \tau_k \left\langle v^{k+1} - \nabla F(x^k), x^k - y^k \right\rangle + \frac{(L_{\nabla F} + C_\gamma)\tau_k^2}{2} \left\| x^k - y^k \right\|^2 + \frac{C_\gamma}{2} \left\| z^{k+1} - z^k \right\|^2. \end{aligned} \tag{2.18}$$

Noting that $z^{k+1} - z^k = \tau_k(\nabla F(x^k) - z^k) + \tau_k \Delta^{k+1}$ where $\Delta^{k+1} = v^{k+1} - \nabla F(x^k)$, we can get

$$\left\| z^{k+1} - z^k \right\|^2 = \tau_k^2 \left\{ \left\| \nabla F(x^k) - z^k \right\|^2 + \left\| \Delta^{k+1} \right\|^2 + 2 \left\langle \Delta^{k+1}, \nabla F(x^k) - z^k \right\rangle \right\}. \tag{2.19}$$

In addition, by the update rule of z^k, we have

$$\begin{aligned}
\left\|\nabla F(x^{k+1}) - z^{k+1}\right\|^2 &= \left\|(1-\tau_k)\left[\nabla F(x^k) - z^k\right] + \nabla F(x^{k+1}) - \nabla F(x^k) + \tau_k \Delta^{k+1}\right\|^2 \\
&= \left\|(1-\tau_k)\left[\nabla F(x^k) - z^k\right] + \nabla F(x^{k+1}) - \nabla F(x^k)\right\|^2 + \tau_k^2 \left\|\Delta^{k+1}\right\|^2 + \vartheta^{k+1} \\
&\le (1-\tau_k)\left\|\nabla F(x^k) - z^k\right\|^2 + \frac{1}{\tau_k}\left\|\nabla F(x^{k+1}) - \nabla F(x^k)\right\|^2 + \tau_k^2 \left\|\Delta^{k+1}\right\|^2 + \vartheta^{k+1} \\
&\le (1-\tau_k)\left\|\nabla F(x^k) - z^k\right\|^2 + \tau_k L_{\nabla F}^2 \left\|x^k - y^k\right\|^2 + \tau_k^2 \left\|\Delta^{k+1}\right\|^2 + \vartheta^{k+1} \qquad (2.20)
\end{aligned}$$

where $\vartheta^{k+1} := 2\tau_k \left\langle \Delta^{k+1}, (1-\tau_k)\left[\nabla F(x^k) - z^k\right] + \nabla F(x^{k+1}) - \nabla F(x^k)\right\rangle$.

Combining (2.18), (2.19), (2.20), and the definition of $W(x^k, z^k)$ in (2.9), we have

$$\begin{aligned}
W_\lambda(x^{k+1}, z^{k+1}) - W_\lambda(x^k, z^k) &\le \tau_k \left\{-\gamma^{-1} + \frac{(L_{\nabla F} + C_\gamma)\tau_k}{2} + \lambda L_{\nabla F}^2\right\} \left\|x^k - y^k\right\|^2 \\
&+ \tau_k \left\{-\lambda + \frac{C_\gamma \tau_k}{2}\right\} \left\|\nabla F(x^k) - z^k\right\|^2 + \frac{(C_\gamma + 2\lambda)\tau_k^2}{2} \left\|\Delta^{k+1}\right\|^2 + \tau_k r^{k+1}.
\end{aligned}$$

where r^{k+1} is defined in (2.13). □

Next, we shall demonstrate the proof of Theorem 2.1.

Proof of Theorem 2.1. For simplicity, set $\gamma = \frac{c}{L_{\nabla F}}$ and $\lambda = \frac{c}{2L_{\nabla F}}$. Then, we have

$$C_\gamma = 2\sqrt{\left(1 + \frac{1}{\gamma}\right)^2 + \left(1 + \frac{\gamma}{2}\right)^2} \le \gamma + \frac{2}{\gamma} + 4 = \frac{2L_{\nabla F}}{c} + \frac{c}{L_{\nabla F}} + 4 \qquad (2.21)$$

For $k \ge 1$, choosing τ_k such that $\tau_k \le \min\{\frac{1}{4}, \frac{L_{\nabla F}}{4cC_\gamma}, \frac{cL_{\nabla F}^{-1}}{4C_\gamma}\}$, we can re-organize the terms in (2.12) as

$$\begin{aligned}
W_\lambda(x^{k+1}, z^{k+1}) - W_\lambda(x^k, z^k) &\le -\tau_k \left\{\frac{L_{\nabla F}}{4c}\left\|x^k - y^k\right\|^2 + \frac{c}{4L_{\nabla F}}\left\|\nabla F(x^k) - z^k\right\|^2\right\} \\
&+ \tau_k^2 \left\{\frac{C_\gamma + 2\lambda}{2}\left\|\Delta^{k+1}\right\|^2\right\} + \tau_k r^{k+1}. \qquad (2.22)
\end{aligned}$$

When $k = 0$, setting $\tau_0 = 1$, we have

$$\begin{aligned}
W_\lambda(x^1, z^1) - W_\lambda(x^0, z^0) &\le \left\{-\gamma^{-1} + \frac{(L_{\nabla F} + C_\gamma)}{2} + \lambda L_{\nabla F}^2\right\} \left\|x^0 - y^0\right\|^2 \\
&+ \left\{-\lambda + \frac{C_\gamma}{2}\right\} \left\|\nabla F(x^0) - z^0\right\|^2 + \frac{(C_\gamma + 2\lambda)}{2} \left\|\Delta^1\right\|^2 + r^1. \qquad (2.23)
\end{aligned}$$

Telescoping (2.22) from $k = 1$ to $k = N$, together with (2.23), (2.9), and $z^0 = \mathbf{0}$, we have

$$
\begin{aligned}
\sum_{k=1}^{N} \tau_k \left\{ \frac{L_{\nabla F}}{4c} \left\| x^k - y^k \right\|^2 + \frac{c}{4L_{\nabla F}} \left\| \nabla F(x^k) - z^k \right\|^2 \right\} &\le \Phi(x^0) - \Phi_* + \Psi(x^0) - \Psi(y^0) \\
&\quad + \left\{ \frac{c^2 + c - 3}{2c} L_{\nabla F} + \frac{C_\gamma}{2} \right\} \left\| x^0 - y^0 \right\|^2 + \frac{C_\gamma}{2} \left\| \nabla F(x^0) \right\|^2 \\
&\quad + \sum_{k=0}^{N} \tau_k^2 \left\{ \frac{C_\gamma + (c/L_{\nabla F})}{2} \left\| \Delta^{k+1} \right\|^2 \right\} + \sum_{k=0}^{N} \tau_k r^{k+1}.
\end{aligned}
\tag{2.24}
$$

Setting $\tau_k = \frac{1}{\sqrt{N}}$ for all $k \ge 1$, taking the expectation of the above inequality, and noting that under Assumption 2.3

$$
\mathbb{E}\left[\Delta^{k+1} \middle| \mathscr{F}_k \right] = 0, \quad \mathbb{E}\left[\left\| \Delta^{k+1} \right\|^2 \middle| \mathscr{F}_k \right] \le \sigma^2,
$$

we have for $N \ge N_0 = \max\left\{ 16, \frac{16c^2 C_\gamma^2}{L_{\nabla F}^2}, \frac{16 C_\gamma^2 L_{\nabla F}^2}{c^2} \right\}$

$$
\begin{aligned}
\mathbb{E}\left[\left\| y^R - x^R \right\|^2 \right] = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left[\left\| y^k - x^k \right\|^2 \right] &\le \frac{4c \left(C_\gamma L_{\nabla F} + c \right) \sigma^2}{L_{\nabla F} \sqrt{N}} \\
+ \frac{2c \left\{ 2 \left[\Phi(x^0) - \Phi_* + \Psi(x^0) - \Psi(y^0) \right] + \left[(c+1) L_{\nabla F} + C_\gamma \right] \left\| x^0 - y^0 \right\|^2 + C_\gamma \left\| \nabla F(x^0) \right\|^2 \right\}}{L_{\nabla F} \sqrt{N}}
\end{aligned}
\tag{2.25}
$$

$$
\begin{aligned}
\mathbb{E}\left[\left\| \nabla F(x^R) - z^R \right\|^2 \right] = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left[\left\| \nabla F(x^R) - z^R \right\|^2 \right] &\le \frac{4L_{\nabla F} \left(C_\gamma L_{\nabla F} + c \right) \sigma^2}{c \sqrt{N}} \\
+ \frac{2L_{\nabla F} \left\{ 2 \left[\Phi(x^0) - \Phi_* + \Psi(x^0) - \Psi(y^0) \right] + \left[(c+1) L_{\nabla F} + C_\gamma \right] \left\| x^0 - y^0 \right\|^2 + C_\gamma \left\| \nabla F(x^0) \right\|^2 \right\}}{c \sqrt{N}}.
\end{aligned}
\tag{2.26}
$$

By the triangle inequality and the non-expansiveness of the proximal operator, we obtain

$$
\left\| x^k - \mathbf{prox}_{\gamma\Psi}\left(x^k - \gamma \nabla F(x^k) \right) \right\| \le \left\| x^k - y^k \right\| + \gamma \left\| \nabla F(x^k) - z^k \right\|.
\tag{2.27}
$$

Therefore, we can obtain

$$
\begin{aligned}
\mathbb{E}\left[\left\| x^R - \mathbf{prox}_{\gamma\Psi}\left(x^R - \gamma \nabla F(x^R) \right) \right\|^2 \right] &\le 2\mathbb{E}\left[\left\| x^R - y^R \right\|^2 + \gamma^2 \left\| \nabla F(x^R) - z^R \right\|^2 \right] \\
\le \frac{4c \left\{ 2 \left[\Phi(x^0) - \Phi_* + \Psi(x^0) - \Psi(y^0) \right] + \left[(c+1) L_{\nabla F} + C_\gamma \right] \left\| x^0 - y^0 \right\|^2 + C_\gamma \left\| \nabla F(x^0) \right\|^2 \right\}}{L_{\nabla F} \sqrt{N}}
\end{aligned}
$$

$$+ \frac{8c\left(C_\gamma L_{\nabla F} + c\right)\sigma^2}{L_{\nabla F}\sqrt{N}}, \tag{2.28}$$

which completes the proof without assuming the bounded second moment of stochastic gradients. We then prove the last part of Theorem 2.1. Firstly, if $\Psi(x)$ is L_Ψ-Lipschitz continuous, then by the optimality of y^k we can have

$$\frac{1}{\gamma}\left\|y^k - x^k\right\|^2 \leq \Psi(x^k) - \Psi(y^k) + \left\langle z^k, x^k - y^k\right\rangle \leq L_\Psi\left\|x^k - y^k\right\| + \left\|z^k\right\|\left\|x^k - y^k\right\|,$$

which implies that

$$\left\|y^k - x^k\right\| \leq \gamma\left(L_\Psi + \left\|z^k\right\|\right). \tag{2.29}$$

Noting that

$$z^k = \sum_{i=1}^{k}\alpha_{i,k}v^i, \qquad \text{where } \alpha_{i,k} = \tau_{i-1}\prod_{j=i}^{k-1}(1-\tau_j), \quad \sum_{i=1}^{k}\alpha_{i,k} = 1, \tag{2.30}$$

we can further bound $\mathbb{E}\left[\|z^k\|^2\right]$ using Assumption 2.4, i.e.,

$$\mathbb{E}\left[\left\|z^k\right\|^2\right] = \sum_{i=1}^{k}\alpha_i\mathbb{E}\left[\|v^i\|^2\right] \leq \sigma^2. \tag{2.31}$$

Therefore, $\mathbb{E}\left[\|x^k - y^k\|^2\right] \leq 2\gamma^2\left(L_\Psi^2 + \sigma^2\right)$. Furthermore,

$$\begin{aligned}\mathbb{E}\left[\left\|z^{k+1} - z^k\right\|^2\right] &= \tau_k^2\mathbb{E}\left[\left\|v^{k+1} - z^k\right\|^2\right] \leq 2\tau_k^2\left\{\mathbb{E}\left[\left\|v^{k+1}\right\|^2\right] + \mathbb{E}\left[\left\|\sum_{i=0}^{k-1}\alpha_{i,k}v^{i+1}\right\|^2\right]\right\} \\ &\leq 2\tau_k^2\left\{\mathbb{E}\left[\left\|v^{k+1}\right\|^2\right] + \sum_{i=0}^{k-1}\alpha_{i,k}\mathbb{E}\left[\|v^{i+1}\|^2\right]\right\} = 4\tau_k^2\sigma^2.\end{aligned} \tag{2.32}$$

With these results, we can then use (2.18) and (2.20) to obtain another basic inequality:

$$\begin{aligned}W_\lambda(x^{k+1}, z^{k+1}) - W_\lambda(x^k, z^k) &\leq \tau_k\left\{(-\gamma^{-1} + \lambda L_{\nabla F}^2)\left\|x^k - y^k\right\|^2 - \lambda\left\|\nabla F(x^k) - z^k\right\|^2\right\} \\ + \tau_k^2\left\{\lambda\tau_k^2\left\|\Delta^{k+1}\right\|^2 + \frac{(L_{\nabla F} + C_\gamma)}{2}\left\|x^k - y^k\right\|^2\right\} &+ \frac{C_\gamma}{2}\left\|z^{k+1} - z^k\right\|^2 + \tau_k\dot{r}^{k+1}.\end{aligned} \tag{2.33}$$

where $\dot{r}^{k+1} = \big\langle \Delta^{k+1}, x^k - y^k + 2\lambda\left((1-\tau_k)\left[\nabla F(x^k) - z^k\right] + \nabla F(x^{k+1}) - \nabla F(x^k)\right)\big\rangle$. In this scenario, the proof can be completed using analogous arguments when taking the expectation of (2.33) without any restrictions on τ_k, i.e., the convergence results hold for any $N \geq 1$. □

In the high probability results, we aim to handle the tail probability for two terms involving the mean-zero noise with the sub-gaussian norm, $\|\Delta^{k+1}\|^2$ and $\langle \Delta^{k+1}, \Lambda^k\rangle$, where (Δ^k) and (Λ^k) are adapted to $(\mathscr{F}_k)$. Our proof leverages Lemma 1.2 and Lemma 1.3 to control the probability of these two terms being too large.

Proof of Theorem 2.2. We start with presenting the lemma below, which leverages concentration inequalities to show a high-probability upper bound for terms involved in the previous analysis.

Lemma 2.4. *Suppose Assumption 2.2 and 2.5 hold. For any $\delta_1, \delta_2, a > 0$, we have*

(a) with probability at least $1-\delta_1$, $\sum_{k=0}^N \tau_k^2 \left\|\Delta^{k+1}\right\|^2 \lesssim K^2 \log(2/\delta_1) \sum_{k=0}^N \tau_k^2$;

(b) with probability at least $1-\delta_2$,

$$\sum_{k=0}^{N} \tau_k \left\langle \Delta^{k+1}, x^k - y^k + [2\lambda(1-\tau_k) + C_\gamma \tau_k]\left(\nabla F(x^k) - z^k\right) + 2\lambda\left(\nabla F(x^{k+1}) - \nabla F(x^k)\right)\right\rangle$$

$$\lesssim 4a\log(1/\delta_2) + \frac{6cK^2}{a}\sum_{k=0}^{N} \tau_k^2 \left\{(1+4\lambda^2 L_{\nabla F}^2)\left\|x^k - y^k\right\|^2 + (4\lambda^2 + C_\gamma^2)\left\|\nabla F(x^k) - z^k\right\|^2\right\}.$$

We first show (a). Using the law of total expectation, we have

$$\mathbb{E}\left[\exp\left(\frac{\|\tau_k \Delta^{k+1}\|^2}{\tau_k^2 K^2}\right)\right] \leq 2,$$

which implies that $\|\tau_k\Delta^{k+1}\|^2$ is $\tau_k^2 K^2$-sub-exponential. Thus, we have with probability at least $1-\delta_1$,

$$\sum_{k=0}^{N} \tau_k^2 \left\|\Delta^{k+1}\right\|^2 \lesssim K^2 \log(2/\delta_1) \sum_{k=0}^{N} \tau_k^2. \tag{2.34}$$

To prove (b), we apply Lemma 1.1 and Lemma 1.3 with

$$X_i = \tau_k \left\langle \Delta^{k+1}, x^k - y^k + [2\lambda(1-\tau_k) + C_\gamma \tau_k] \left(\nabla F(x^k) - z^k\right) + 2\lambda \left(\nabla F(x^{k+1}) - \nabla F(x^k)\right)\right\rangle,$$
$$K_i = K\tau_k \left\| x^k - y^k + [2\lambda(1-\tau_k) + C_\gamma \tau_k] \left(\nabla F(x^k) - z^k\right) + 2\lambda \left(\nabla F(x^{k+1}) - \nabla F(x^k)\right)\right\|,$$
$$b = 0, t = 4a \log(1/\delta_2).$$

We obtain that for all $a > 0$ with probability at least $1 - \delta_2$, $\sum_{i=0}^N X_i \leq 4a\log(1/\delta_2)$ and

$$\begin{aligned}\sum_{i=0}^N X_i &\lesssim \frac{2K^2}{a} \sum_{k=0}^N \tau_k^2 \left\| x^k - y^k + [2\lambda(1-\tau_k) + C_\gamma \tau_k] \left(\nabla F(x^k) - z^k\right) + 2\lambda \left(\nabla F(x^{k+1}) - \nabla F(x^k)\right)\right\|^2 \\ &\lesssim \frac{6K^2}{a} \sum_{k=0}^N \tau_k^2 \left\{ \left\| x^k - y^k \right\|^2 + (4\lambda^2 + C_\gamma^2) \left\| \nabla F(x^k) - z^k \right\|^2 + 4\lambda^2 L_{\nabla F}^2 \left\| x^k - y^k \right\|^2 \right\},\end{aligned}$$

where the second inequality comes from the Lipschitzness of ∇F.

With the above lemma, we can now complete the proof of Theorem 2.2 by setting $\delta_1 = \delta_2 = \delta/2$ and following the similar arguments as in the proof of Theorem 2.1.

□

2.3.6. Almost Surely Asymptotic Convergence. In this subsection, we also establish the asymptotic convergence of Prox-ASA. We select any time-varying positive $\{\tau_k\}$ that satisfy

$$\sum_{k=0}^\infty \tau_k = \infty \quad \text{and} \quad \sum_{k=0}^\infty \tau_k^2 < \infty. \tag{2.35}$$

These two requirements are standard in the examination of stochastic approximation [**RM51**,**BT00**]. The first condition is necessary to move away from the initial point as much as desired, while the second condition is necessary to maintain control over the variance of the noise. In the following theorem, we derive the almost surely asymptotic convergence under above conditions.

THEOREM 2.3 (Almost Surely Asymptotic Convergence). *Suppose Assumption 2.1, 2.2, 2.3 hold. If $\gamma > 0$ and $\sum_{k=0}^\infty \tau_k = \infty$ and $\sum_{k=0}^\infty \tau_k^2 < \infty$, then with probability 1, the sequences generated by Algorithm 1 satisfy*

$$\lim_{k\to\infty} \|y^k - x^k\| = 0,$$
$$\lim_{k\to\infty} \|z^k - \nabla F(x^k)\| = 0,$$

$$\lim_{k\to\infty} \|x^k - \mathbf{prox}_{\Psi}(x^k - \nabla F(x^k))\| = 0.$$

The proof of the above outcomes can be traced back to [**RM51**]. Nonetheless, we will simplify the lengthy proof for better understanding as follows.

Proof of Theorem 2.3. Noting that by (2.24), with the conditon that $\sum_{k=1}^{\infty} \tau_k^2 \leq \infty$, we have

$$\sum_{k=0}^{\infty} \tau_k \mathbb{E}\left[\left\|x^k - y^k\right\|^2\right] \lesssim \sum_{k=0}^{\infty} \tau_k^2 < \infty, \quad \sum_{k=0}^{\infty} \tau_k \mathbb{E}\left[\left\|\nabla F(x^k) - z^k\right\|^2\right] \lesssim \sum_{k=0}^{\infty} \tau_k^2 < \infty, \tag{2.36}$$

which implies that $\sum_{k=0}^{\infty} \tau_k \left\|x^k - y^k\right\|^2 < \infty$ and $\sum_{k=0}^{\infty} \tau_k \left\|\nabla F(x^k) - z^k\right\|^2 < \infty$ with probability 1. From these inequality and the condition that $\sum_{k=0}^{\infty} \tau_k = \infty$, we can derive the fact that with probability 1

$$\liminf_{k\to\infty} \left\|x^k - y^k\right\|^2 = 0, \qquad \liminf_{k\to\infty} \left\|\nabla F(x^k) - z^k\right\|^2 = 0 \tag{2.37}$$

It remains to show $\limsup_{k\to\infty} \left\|x^k - y^k\right\|^2 = \limsup_{k\to\infty} \left\|\nabla F(x^k) - z^k\right\|^2 = 0$. As it turns out, the lemma below from [**Ora20**] is essentially all that we require. This lemma can be proved by contradiction assuming $\limsup_{k\to\infty} b_k = \lambda \in (0, +\infty)$. For simplicity, we will omit its proof.

Lemma 2.5 ([**Ora20**]). *Let $\{b_k\}_{k\geq 1}, \{\tau_k\}_{k\geq 1}$ be two non-negative sequences and $\{u^k\}_{k\geq 1}$ a sequence of vectors in $\mathbb{R}^d$. Let $p \geq 1$ and assume $\sum_{k=1}^{\infty} \tau_k b_k^p < \infty$ and $\sum_{k=1}^{\infty} \tau_k = \infty$. If there exists $M \geq 0$ such that $|b_{k+t} - b_k| \leq M\left(\sum_{i=k}^{k+t-1} \tau_i b_i + \left\|\sum_{i=k}^{k+t-1} \tau_i u^i\right\|\right)$, where $\{u^k\}$ is such that $\left\|\sum_{k=1}^{\infty} \tau_i u^k\right\| < \infty$, then b_k converges to 0.*

With this lemma, we can now check the asymptotic convergence of $\left\{\left\|x^k - y^k\right\|, \left\|\nabla F(x^k - z^k)\right\|\right\}$. Observe that by the triangle inequality and non-expansiveness of the proximal operator,

$$\begin{aligned}
&\left|\left\|x^{k+t} - y^{k+t}\right\| - \left\|x^k - y^k\right\|\right| \\
&\leq \left\|(x^{k+t} - y^{k+t}) - (x^k - y^k)\right\| \leq \left\|x^{k+t} - x^k\right\| + \left\|y^{k+t} - y^k\right\| \\
&\leq 2\left\|x^{k+t} - x^k\right\| + \gamma\left\|z^{k+t} - z^k\right\| \leq 2\sum_{i=k}^{k+t-1} \left\|x^{i+1} - x^i\right\| + \gamma\left\|\sum_{i=k}^{k+t-1} (z^{i+1} - z^i)\right\|
\end{aligned}$$

$$\leq \max\{2,\gamma\}\left\{\sum_{i=k}^{k+t-1}\tau_i\left\|x^i-y^i\right\|+\left\|\sum_{i=k}^{k+t-1}(z^{i+1}-z^i)\right\|\right\}$$

$$\leq \max\{2,\gamma\}\left\{\sum_{i=k}^{k+t-1}\tau_i\left(\left\|x^i-y^i\right\|+\left\|\nabla F(x^i)-z^i\right\|\right)+\left\|\sum_{i=k}^{k+t-1}\tau_k\left(v^{k+1}-\nabla F(x^k)\right)\right\|\right\} \tag{2.38}$$

Moreover, by the smoothness of $F(x)$

$$\left|\left\|\nabla F(x^{k+t})-z^{k+t}\right\|-\left\|\nabla F(x^k)-z^k\right\|\right|$$

$$\leq \left\|(\nabla F(x^{k+t})-z^{k+t})-(\nabla F(x^k)-z^k)\right\| \leq L_{\nabla F}\left\|x^{k+t}-x^k\right\|+\left\|z^{k+t}-z^k\right\|$$

$$\leq \max\{1,L_{\nabla F}\}\left\{\sum_{i=k}^{k+t-1}\tau_i\left(\left\|x^i-y^i\right\|+\left\|\nabla F(x^i)-z^i\right\|\right)+\left\|\sum_{i=k}^{k+t-1}\tau_k\left(v^{k+1}-\nabla F(x^k)\right)\right\|\right\} \tag{2.39}$$

Thus, combing (2.38) and (2.39) and apply the triangle inequality again, we have

$$|b_{k+t}-b_k| \leq M\left(\sum_{i=k}^{k+t-1}\tau_i b_i+\left\|\sum_{i=k}^{k+t-1}u^i\right\|\right),$$

where

$$b_k=\left\|x^k-y^k\right\|+\left\|\nabla F(x^k)-z^k\right\|, \quad u^k=v^{k+1}-\nabla F(x^k), \quad M=\max\{2,L_{\nabla F},\gamma\}.$$

Note that $\{\sum_{k=0}^N \tau_k u^k\}_{N=0,1,\dots}$ is a martingale whose variance is bounded by $\sigma^2\sum_{k=0}^\infty \tau_k^2<\infty$. Hence, $\{\sum_{k=0}^N \tau_k u^k\}_{N=0,1,\dots}$ is a martingale in L^2, so it converges in L^2 with probability 1. Overall, with probability 1 the assumptions of Lemma 2.5 are verified with $p=2$. Therefore, with probability 1, we have

$$\lim_{k\to\infty}\left\|x^k-y^k\right\|=0, \qquad \lim_{k\to\infty}\left\|\nabla F(x^k)-z^k\right\|=0,$$

which together implies that

$$\lim_{k\to\infty}\left\|x^k-\mathbf{prox}_\Psi(x^k-\nabla F(x^k))\right\|=0.$$

□

2.4. Discussion and Conclusion

In this chapter, we introduce and examine a novel class of proximal gradient methods, Prox-ASA, to solve non-convex stochastic optimization problems. Despite the absence of theoretical improvement in the convergence rate when compared to (Prox-)SGD, Prox-ASA exhibits a decreasing variance of the gradient estimator z^k, ensuring that $\mathbb{E}\left[\left\|z^k - \nabla F(x^k)\right\|^2\right] \to 0$ as $k \to +\infty$. Consequently, a reliable terminating criterion of $\left\|\mathbf{prox}_{\Psi}(x^k - z^k) - x^k\right\|$ can be employed. Additionally, this algorithm class matches the lower bound of the sample complexity established for smooth problems $(\Psi = 0)$ [**ACD$^+$19**]. Finally, this algorithm class can be further extended to tackle numerous challenging problems in which (Prox-)SGD may not yield satisfactory results, which we will illustrate in the next two chapters.

CHAPTER 3

Decentralized Proximal Averaged Stochastic Approximation

3.1. Introduction

Decentralized optimization is a flexible paradigm for solving complex optimization problems in a distributed manner, and has numerous applications in fields such as machine learning, robotics, and control systems. It has attracted increased attention due to the following benefits: (i) *Robustness*: Decentralized optimization is more robust than centralized optimization because each agent can operate independently, making the system more resilient to failures compared to a centralized system where a coordinator failure or overload can halt the entire system. (ii) *Privacy*: Decentralized optimization can provide greater privacy because each agent only has access to a limited subset of observations, which may help to protect sensitive information. (iii) *Scalability*: Decentralized optimization is highly scalable as it can handle a large datasets in a distributed manner, thereby solving complex optimization problems that are difficult or even impossible to solve in a centralized setting.

Specifically, we consider the following decentralized composite optimization problems in which n agents collaborate to solve

$$\min_{x\in\mathbb{R}^d} \Phi(x) := F(x) + \Psi(x),\ F(x) := \frac{1}{n}\sum_{i=1}^{n} F_i(x), \tag{3.1}$$

where each function $F_i(x)$ is a smooth function only known to the agent i; $\Psi(x)$ is non-smooth, convex, and shared across all agents; $\Phi(x)$ is bounded below by $\Phi_* > -\infty$. We consider the stochastic setting where the exact function values and derivatives of F_i's are not available. In particular, we assume that $F_i(x) = \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[G_i(x,\xi_i)]$, where ξ_i is a random vector and $\mathcal{D}_i$ is the distribution used to generate samples for agent i. The agents form a connected and undirected network and can communicate with their neighbors to cooperatively solve (3.1). The communication network can be represented with $\mathbb{G} = (\mathcal{V}, \mathbf{W})$ where $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ denotes all devices and $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{n\times n}$ is the weighted adjacency matrix indicating how two agents are connected.

A majority of the existing decentralized stochastic algorithms for solving (3.1), require large batch sizes to achieve convergence. The few algorithms that operate with constant batch sizes mainly rely on complicated variance reduction techniques and require stronger assumptions to establish convergence results. To the best of our knowledge, the question of whether it's possible to develop decentralized stochastic optimization algorithms to solve (3.1) without the above mentioned limitations, remains unresolved.

To address this, we propose the two decentralized stochastic proximal algorithms, `Prox-DASA` and `Prox-DASA-GT`, for solving (3.1) and make the following **contributions**:

- We show that `Prox-DASA` is capable of achieving convergence in both homogenous and bounded heterogeneous settings while `Prox-DASA-GT` works for general decentralized heterogeneous problems.
- We show that both algorithms find an ϵ-stationary point in $\mathcal{O}(n^{-1}\epsilon^{-2})$ iterations using only $\mathcal{O}(1)$ stochastic gradient samples per agent and m communication rounds at each iteration, where m can be any positive integer. A topology-independent transient time can be achieved by setting $m = \lceil \frac{1}{\sqrt{1-\rho}} \rceil$, where ρ is the second-largest eigenvalue of the communication matrix.
- Through extensive experiments we demonstrate the superiority of our algorithms over prior works.

A summary of our results and comparison to prior work is provided in Table 3.1.

3.1.1. Related Works on Decentralized Composite Optimization.

Motivated by wide applications in constrained optimization [**LN13**, **MFGP17**] and non-smooth problems with a composite structure as (3.1), arising in signal processing [**LT10**, **MBG10**, **PEK14**] and machine learning [**FSS15**, **HHZ17**], several works have studied the decentralized composite optimization problem in (3.1), a natural generalization of smooth optimization. For example, [**SLWY15**, **LSY19**, **AYS19**, **YZLZ20**, **XTSS21**, **LLT$^+$21**, **SSD22**, **WL22**] studied (3.1) in the convex setting. Furthermore, [**FSS15**, **DLS16**, **HHZ17**, **ZY18**, **SS19**] studied (3.1) in the deterministic setting.

Although there has been a lot of research investigating decentralized composite optimization, the stochastic non-convex setting, which is more broadly applicable, still lacks a full understanding. [**WZC$^+$21**] proposes `SPPDM`, which uses a proximal primal-dual approach to achieve $\mathcal{O}(\epsilon^{-2})$ sample

TABLE 3.1. Comparison of decentralized proximal gradient based algorithms to find an ϵ-stationary solution to stochastic composite optimization in the nonconvex setting. The sample complexity is defined as the number of required samples per agent to obtain an ϵ-stationary point (see Definition 3.1). We omit a comparison with `SPPDM` [**WZC$^+$21**] as their definition of stationarity differs from ours; see Appendix A.3 for further discussions.

Algorithm	**Batch Size**	**Sample Complexity**	**Communication Complexity**	**Linear Speedup?**	**Remark**
`ProxGT-SA` [**XDKK21**]	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(n^{-1}\epsilon^{-2})$	$\mathcal{O}(\log(n)\epsilon^{-1})$	✓	
`ProxGT-SR-O` [**XDKK21**]	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(n^{-1}\epsilon^{-1.5})$	$\mathcal{O}(\log(n)\epsilon^{-1})$	✓	(i) double-loop; (ii) mean-squared smoothness

`DEEPSTORM` [**MBMXC22**]	$\mathcal{O}(\epsilon^{-0.5})$ then $\mathcal{O}(1)^*$	$\mathcal{O}(n^{-1}\epsilon^{-1.5})$	$\mathcal{O}(n^{-1}\epsilon^{-1.5})$	✓	(i) two time-scale;
					(ii) mean-squared smoothness;
					(iii) double gradient evaluations per iteration

	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-1.5}\lvert\log\epsilon\rvert^{-1.5})$	$\mathcal{O}(\epsilon^{-1.5}\lvert\log\epsilon\rvert^{-1.5})$	✗	
`Prox-DASA` (Alg. 2)	$\mathcal{O}(1)$	$\mathcal{O}(n^{-1}\epsilon^{-2})$	$\mathcal{O}(n^{-1}\epsilon^{-2})$	✓	bounded heterogeneity
`Prox-DASA-GT` (Alg. 3)	$\mathcal{O}(1)$	$\mathcal{O}(n^{-1}\epsilon^{-2})$	$\mathcal{O}(n^{-1}\epsilon^{-2})$	✓	

* It requires $\mathcal{O}(\epsilon^{-0.5})$ batch size in the first iteration and then $\mathcal{O}(1)$ for the rest (see m_0 in Algorithm 1 in [**MBMXC22**]).

complexity. `ProxGT-SA` and `ProxGT-SR-O` [**XDKK21**] incorporate stochastic gradient tracking and multi-consensus update in proximal gradient methods and obtain $\mathcal{O}(n^{-1}\epsilon^{-2})$ and $\mathcal{O}(n^{-1}\epsilon^{-1.5})$ sample complexity respectively, where the latter further uses a `SARAH` type variance reduction method [**PNPTD20**, **WJZ$^+$19**]. A recent work [**MBMXC22**] proposes `DEEPSTORM`, which leverages a `STORM` type of variance reduction technique [**CO19**] and gradient tracking to obtain $\mathcal{O}(n^{-1}\epsilon^{-1.5})$ and $\tilde{\mathcal{O}}(\epsilon^{-1.5})$ sample complexity under different stepsize choices. Nevertheless, existing works either require stronger assumptions [**MBMXC22**] or increasing batch sizes [**WZC$^+$21**, **XDKK21**].

3.1.2. Notations. $\|\cdot\|$ denotes the ℓ_2-norm for vectors and Frobenius norm for matrices. $\|\cdot\|_2$ denotes the spectral norm for matrices. $\mathbf{1}$ represents the all-one vector, and $\mathbf{I}$ is the identity matrix as a standard practice. We identify vectors at agent i in the subscript and use the superscript for the algorithm step. For example, the optimization variable of agent i at step k is denoted as x_i^k, and z_i^k is the corresponding dual variable. We use uppercase bold letters to represent the matrix that collects all the variables from nodes (corresponding lowercase) as columns. We add an overbar to a letter to denote the average over all nodes. For example, we denote the optimization variables over all nodes at step k as $\mathbf{X}_k = \left[x_1^k, \ldots, x_n^k\right]$. The corresponding average over all nodes can be thereby defined as

$$\bar{x}^k = \frac{1}{n}\sum_{i=1}^{n} x_i^k = \frac{1}{n}\mathbf{X}_k\mathbf{1}, \quad \bar{\mathbf{X}}_k = [\bar{x}^k, \ldots, \bar{x}^k] = \bar{x}^k\mathbf{1}^\top = \frac{1}{n}\mathbf{X}_k\mathbf{1}\mathbf{1}^\top.$$

For an extended valued function $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, its effective domain is written as $\mathrm{dom}(\Psi) = \{x \mid \Psi(x) < +\infty\}$. A function Ψ is said to be proper if $\mathrm{dom}(\Psi)$ is nonempty. For any proper closed convex function Ψ, $x \in \mathbb{R}^d$, and scalar $\gamma > 0$, the proximal operator is defined as

$$\mathbf{prox}_{\gamma\Psi}(x) = \underset{y\in\mathbb{R}^d}{\arg\min}\left\{\frac{1}{2\gamma}\|y - x\|^2 + \Psi(y)\right\}.$$

For $x, z \in \mathbb{R}^d$ and $\gamma > 0$, the proximal gradient mapping of z at x is defined as

$$\mathcal{G}(x, z, \gamma) = \frac{1}{\gamma}\left(x - \mathbf{prox}_{\gamma\Psi}(x - \gamma z)\right).$$

All random objects are properly defined in a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and write $x \in \mathcal{H}$ if x is $\mathcal{H}$-measurable given a sub-σ-algebra $\mathcal{H} \subseteq \mathscr{F}$ and a random vector x. We use $\sigma(\cdot)$ to denote the σ-algebra generated by all the agument random vectors.

3.1.3. Assumptions.

Next, we list and discuss the assumptions made in this work.

Assumption 3.1. *The weighted adjacency matrix* $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{n\times n}$ *is symmetric and doubly stochastic, i.e.,*

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

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

Assumption 3.2. *All functions* $\{F_i\}_{1\leq i\leq n}$ *have Lipschitz continuous gradients with Lipschitz constants* $L_{\nabla F_i}$*, respectively. Therefore,* ∇F *is* $L_{\nabla F}$*-Lipchitz continous with* $L_{\nabla F} = \max_{1\leq i\leq n}\{L_{\nabla F_i}\}$.

Assumption 3.3. *The function* $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ *is a closed proper convex function.*

For stochastic oracles, we assume that each node i at every iteration k is able to obtain a local random data vector ξ_i^k. The induced natural filtration is given by $\mathscr{F}_0 = \{\emptyset, \Omega\}$ and

$$\mathscr{F}_k := \sigma\left(\xi_i^t \mid i = 1, \ldots, n,\ t = 1, \ldots, k\right), \forall k \geq 1.$$

We require that the stochastic gradient $\nabla G_i(\cdot, \xi_i^{k+1})$ is unbiased conditioned on the filteration $\mathscr{F}_k$.

Assumption 3.4 (Unbiasness). *For any $k \geq 0, x \in \mathscr{F}_k$, and $1 \leq i \leq n$,*

$$\mathbb{E}\left[\nabla G_i(x, \xi_i^{k+1}) \middle| \mathscr{F}_k\right] = \nabla F_i(x).$$

Assumption 3.5 (Independence). *For any $k \geq 0, 1 \leq i, j \leq n, i \neq j$, ξ_i^{k+1} is independent of $\mathscr{F}_k$, and ξ_i^{k+1} is independent of ξ_j^{k+1}.*

In addition, we consider two standard assumptions on the variance and heterogeneity of stochastic gradients.

Assumption 3.6 (Bounded variance). *For any $k \geq 0, x \in \mathscr{F}_k$, and $1 \leq i \leq n$,*

$$\mathbb{E}\left[\left\|\nabla G_i(x, \xi_i^{k+1}) - \nabla F_i(x)\right\|^2 \middle| \mathscr{F}_k\right] \leq \sigma_i^2.$$

Let $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$.

Assumption 3.7 (Gradient heterogeneity). *There exists a constant $\nu \geq 0$ such that for all $1 \leq i \leq n, x \in \mathbb{R}^d$, $\|\nabla F_i(x) - \nabla F(x)\| \leq \nu$.*

Remark. *The above assumption of gradient heterogeneity is standard [**LZZ+17**] and less strict than the bounded second moment assumption on stochastic gradients which implies lipschtizness of functions $\{F_i\}$. However, this assumption is only required for the convergence analysis of `Prox-DASA` and can be bypassed by employing a gradient tracking step.*

3.2. Methodology

Several algorithms have been developed to solve Problem (3.1) in the stochastic setting; see Table 3.1. However, the most recent two types of algorithms that achieve (near)-optimal sample complexities have certain drawbacks: (i) **increasing batch sizes**: `ProxGT-SA`, `Prox-SR-O`, and `DEEPSTORM` with constant step sizes (Theorem 1 in [**MBMXC22**]) require batches of stochastic gradients with batch sizes inverse proportional to tolerance ϵ; (ii) **algorithmic complexities**: `ProxGT-SR-O` and `DEEPSTORM` are either double-looped or two-time-scale, and require stochastic gradients evaluated at different parameter values over the same sample, i.e., $\nabla G_i(x, \xi)$ and $\nabla G_i(x', \xi)$. These variance reduction techniques are unfavorable when gradient evaluations are computationally expensive such as forward-backward steps for deep neural networks. (iii) **theoretical weakness**: the convergence analyses of `ProxGT-SR-O` and `DEEPSTORM` are established under the *stronger* assumption of mean-squared lipschtizness of stochastic gradients. In addition, Theorem 2 in [**MBMXC22**] fails to provide linear-speedup results for one-sample variant of `DEEPSTORM` with diminishing stepsizes.

3.2.1. Decentralized Proximal Averaged Stochastic Approximation. To address the above limitations, we propose **D**ecentralized **Prox**imal **A**veraged **S**tochastic **A**ppro-ximation (`Prox-DASA`) which leverages a common averaging technique in stochastic optimation [**Rus08**, **MHK18a**, **GRW20**] to reduce the error of gradient estimation. In particular, the sequences of dual variables $\mathbf{Z}^k = [z_1^k, \dots, z_n^k]$ that aim to approximate gradients are defined in the following recursion:

$$\begin{aligned} \mathbf{Z}^{k+1} &= \left\{(1-\alpha_k)\mathbf{Z}^k + \alpha_k \mathbf{V}^{k+1}\right\} \mathbf{W}^m \\ \mathbf{V}^{k+1} &= [v_1^{k+1}, \dots, v_n^{k+1}], \end{aligned}$$

where each v_i^{k+1} is the local stochastic gradient evaluated at the local variable x_i^k. For complete graphs where each pair of graph vertices is connected by an edge and there is no consensus error for optimization variables, i.e., $\mathbf{W} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $x_i^k = x_j^k, \forall i, j$, the averaged dual variable over nodes $\bar{z}^k$ follows the same averaging rule as in centralized algorithms:

$$\begin{aligned} \bar{z}^{k+1} &= (1-\alpha_k)\bar{z}^k + \alpha_k \bar{v}^{k+1} \\ \mathbb{E}[\bar{v}^{k+1} | \mathscr{F}_k] &= \nabla F(\bar{x}^k). \end{aligned}$$

To further control the consensus errors, we employ a multiple consensus step for both primal and dual iterates $\{x_i^k, z_i^k\}$ which multiply the matrix of variables from all nodes by the weight matrix m times. A pseudo code of `Prox-DASA` is given in Algorithm 2.

3.2.2. Gradient Tracking. The constant ν defined in Assumption 3.7 measures the heterogeneity between local gradients and global gradients, and hence the variance of datasets of different agents. To remove ν in the complexity bound, [**TLY**$^+$**18**] proposed the D^2 algorithm, which modifies the x update in D-PSGD [**LZZ**$^+$**17**]. However, it requires one additional assumption on the eigenvalues of the mixing matrix $\mathbf{W}$. Here we adopt the gradient tracking technique, which was first introduced to deterministic distributed optimization to improve the convergence rate [**XZSX15**, **DLS16**, **NOS17**, **QL17**], and was later proved to be useful in removing the data

Algorithm 2 Prox-DASA

```
Input: $x_i^0 = z_i^0 = \mathbf{0}, \gamma, \{\alpha_k\}_{\geq 0}, m$
for $k = 0, 1, \ldots, K-1$ do
    # Local Update
    for $i = 1, 2, \ldots, n$ (in parallel) do
        $y_i^k = \mathbf{prox}_{\gamma\Psi}\left(x_i^k - \gamma z_i^k\right)$
        $\tilde{x}_i^{k+1} = (1-\alpha_k)x_i^k + \alpha_k y_i^k$
        # Compute stochastic gradient
        $v_i^{k+1} = \nabla G_i(x_i^k, \xi_i^{k+1})$
        $\tilde{z}_i^{k+1} = (1-\alpha_k)z_i^k + \alpha_k v_i^{k+1}$
    end for
    # Communication
    $[x_1^{k+1}, \ldots, x_n^{k+1}] = [\tilde{x}_1^{k+1}, \ldots, \tilde{x}_n^{k+1}]\mathbf{W}^m$
    $[z_1^{k+1}, \ldots, z_n^{k+1}] = [\tilde{z}_1^{k+1}, \ldots, \tilde{z}_n^{k+1}]\mathbf{W}^m$
end for
```

variance (i.e., ν) dependency in the stochastic case [**ZY19**, **LZSH19**, **PN21**, **KLS21**]. In the convergence analysis of Prox-DASA, an essential step is to control the heterogeneity of stochastic gradients, i.e., $\mathbb{E}[\left\|\mathbf{V}^{k+1} - \bar{\mathbf{V}}^{k+1}\right\|^2]$, which requires bounded heterogeneity of local gradients (Assumption 3.7). To pypass this assumption, we employ a gradient tracking step by replacing $\mathbf{V}^{k+1}$ with pseudo stochastic gradients $\mathbf{U}^{k+1} = [u_1^{k+1}, \ldots, u_n^{k+1}]$, which is updated as follows:

$$\mathbf{U}^{k+1} = \left(\mathbf{U}^k + \mathbf{V}^{k+1} - \mathbf{V}^k\right)\mathbf{W}^m.$$

Provided that $\mathbf{U}^0 = \mathbf{V}^0$ and $\mathbf{W}\mathbf{1} = \mathbf{1}$, one can show that $\bar{u}^k = \bar{v}^k$ at each step k. In addition, with the consensus procedure over $\mathbf{U}^k$, the heterogeneity of pseudo stochastic gradients $\mathbb{E}[\left\|\mathbf{U}^{k+1} - \bar{\mathbf{U}}^{k+1}\right\|^2]$ can be bounded above. The proposed algorithm, which we name as Prox-DASA with Gradient Tracking (Prox-DASA-GT), is presented in Algorithm 3.

3.2.3. Consensus Algorithm. In practice, we can leverage accelerated consensus algorithms, e.g., [**LM11**, **Ols17**], to speed up the multiple consensus step $\mathbf{W}^m$ to achieve improved communication complexities when $m > 1$. Specifically, we can replace $\mathbf{W}^m$ by a Chebyshev-type polynomial of $\mathbf{W}$, which can improve the ρ-dependency of the communication complexity from a factor of $\frac{1}{1-\rho}$ to $\frac{1}{\sqrt{1-\rho}}$.

Then, we have the following lemma.

Algorithm 3 `Prox-DASA-GT`

```
Input: $x_i^0 = u_i^0 = z_i^0 = \mathbf{0}, \gamma, \{\alpha_k\}_{\geq 0}, m$
for $k = 0, 1, \ldots, K$ do
  # Local Update
  for $i = 1, 2, \ldots, n$ (in parallel) do
    $y_i^k = \mathbf{prox}_{\gamma\Psi}\left(x_i^k - \gamma z_i^k\right)$
    $\tilde{x}_i^{k+1} = (1-\alpha_k)x_i^k + \alpha_k y_i^k$
    # Compute stochastic gradient
    $v_i^{k+1} = \nabla G_i(x_i^k, \xi_i^{k+1})$
    $\tilde{u}_i^{k+1} = u_i^k + v_i^{k+1} - v_i^k$
    $\tilde{z}_i^{k+1} = (1-\alpha_k)z_i^k + \alpha_k \tilde{u}_i^{k+1}$
  end for
  # Communication
  $[x_1^{k+1}, \ldots, x_n^{k+1}] = [\tilde{x}_1^{k+1}, \ldots, \tilde{x}_n^{k+1}]\mathbf{W}^m$
  $[u_1^{k+1}, \ldots, u_n^{k+1}] = [\tilde{u}_1^{k+1}, \ldots, \tilde{u}_n^{k+1}]\mathbf{W}^m$
  $[z_1^{k+1}, \ldots, z_n^{k+1}] = [\tilde{z}_1^{k+1}, \ldots, \tilde{z}_n^{k+1}]\mathbf{W}^m$
end for
```

Algorithm 4 Chebyshev Mixing Protocol

```
Input: Matrix $\mathbf{X}$, mixing matrix $\mathbf{W}$, rounds $m$ Set $\mathbf{A}_0 = \mathbf{X}, \mathbf{A}_1 = \mathbf{X}\mathbf{W}, \rho = \max\{|\lambda_2(\mathbf{W})|, |\lambda_n(\mathbf{W})|\} < 1, \mu_0 = 1, \mu_1 = \frac{1}{\rho}$
for $t = 1, \ldots, m-1$ do
  $\mu_{t+1} = \frac{2}{\rho}\mu_t - \mu_{t-1}$
  $\mathbf{A}_{t+1} = \frac{2\mu_t}{\rho\mu_{t+1}}\mathbf{A}_t\mathbf{W} - \frac{\mu_{t-1}}{\mu_{t+1}}\mathbf{A}_{t-1}$
end for
Output: $\mathbf{A}_m$
```

Lemma 3.1. *Suppose* $\mathbf{W}$ *satisfies Assumption 3.1. Let* $\mathbf{A}_0, \mathbf{A}_m$ *be the input and output matrix of Algorithm 4 respectively. Then, we have*

$$\left\|\mathbf{A}_m - \bar{\mathbf{A}}_m\right\| \leq 2\left(1 - \sqrt{1-\rho}\right)^m \left\|\mathbf{A}_0 - \bar{\mathbf{A}}_0\right\|.$$

Hence, we obtain a linear convergence rate of $\left(1 - \sqrt{1-\rho}\right)$ instead of ρ. By virtue of that, we can set $m = \lceil \frac{1}{\sqrt{1-\rho}} \rceil$ to obtain a topology-independent iteration complexity.

3.3. Convergence Analysis

3.3.1. Notion of Stationarity. For centralized optimization problems with non-convex objective function $F(x)$, a standard measure of non-stationarity of a point $\bar{x}$ is the squared norm of proximal gradient mapping of $\nabla F(\bar{x})$ at $\bar{x}$, i.e.,

$$\|\mathcal{G}(\bar{x}, \nabla F(\bar{x}), \gamma)\|^2 = \left\| \frac{1}{\gamma} \left(x - \mathbf{prox}_{\gamma\Psi}(\bar{x} - \gamma \nabla F(\bar{x})) \right) \right\|^2 .$$

For the smooth case where $\Psi(x) \equiv 0$, the above measure is reduced to $\|\nabla F(\bar{x})\|^2$.

However, in the decentralized setting with a connected network G, we solve the following equivalent reformulated consensus optimization problem:

$$\begin{aligned} &\min_{x_1,\ldots,x_n \in \mathbb{R}^d} && \frac{1}{n}\sum_{i=1}^{n} \{F_i(x_i) + \Psi(x_i)\} \\ &\text{s.t.} && x_i = x_j, \ \forall (i,j). \end{aligned} \tag{3.2}$$

To measure the non-stationarity in Problem (3.2), one should not only consider the stationarity violation at each node but also the consensus errors over the network. Therefore, **[XDKK21]** and **[MBMXC22]** define an ϵ-stationary point $\mathbf{X} = [x_1, \ldots, x_n]$ of Problem 3.2 as

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left\{ \|\mathcal{G}(x_i, \nabla F(x_i), \gamma)\|^2 + L_{\nabla F}^2 \|x_i - \bar{x}\|^2 \right\} \right] \le \epsilon. \tag{3.3}$$

In this work, we use a general measure as follows.

Definition 3.1. *Let $\mathbf{X} = [x_1, \ldots, x_n]$ be random vectors generated by a decentralized algorithm to solve Problem 3.2 and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. We say that $\mathbf{X}$ is an ϵ-stationary point of Problem 3.2 if*

$$\begin{aligned} \mathbb{E}\left[\|\mathcal{G}(\bar{x}, \nabla F(\bar{x}), \gamma)\|^2 \right] &\le \epsilon, && \textit{(stationarity violation)} \\ \mathbb{E}\left[\frac{L_{\nabla F}^2}{n} \left\|\mathbf{X} - \bar{\mathbf{X}}\right\|^2 \right] &\le \epsilon. && \textit{(consensus error)} \end{aligned}$$

The next inequality characterizes the difference between the gradient mapping at $\bar{x}$ and x_i, which relates our definition to (3.3). Noting that by non-expansiveness of the proximal operator, we have

$$\|\mathcal{G}(x_i, \nabla F(x_i), \gamma) - \mathcal{G}(\bar{x}, \nabla F(\bar{x}), \gamma)\| \le \frac{2 + \gamma L_{\nabla F}}{\gamma} \|x_i - \bar{x}\| ,$$

which implies that

$$\frac{1}{n}\sum_{i=1}^{n} \|\mathcal{G}(x_i, \nabla F(x_i), \gamma)\|^2 \lesssim \|\mathcal{G}(\bar{x}, \nabla F(\bar{x}), \gamma)\|^2 + \frac{1}{\gamma^2 n} \left\|\mathbf{X} - \bar{\mathbf{X}}\right\|^2 .$$

3.3.2. Main Results.

We present the complexity results of our algorithms below.

Theorem 3.1. *Suppose Assumption 3.1, 3.2, 3.3, 3.4, 3.5, 3.6 hold and the total number of iterations $K \geq K_0$, where K_0 is a constant that only depends on constants $(n, L_{\nabla F}, \varrho(m), \gamma)$, where $\varrho(m) = \frac{(1+\rho^{2m})\rho^{2m}}{(1-\rho^{2m})^2}$. Let C_0 be some initialization-dependent constant and R be a random integer uniformly distributed over $\{1, 2, \dots, K\}$. Suppose we set $\alpha_k \asymp \sqrt{\frac{n}{K}}, \gamma \asymp \frac{1}{L_{\nabla F}}$.*

(Prox-DASA) *Suppose Assumption 3.7 also holds. Then, for Algorithm 2 we have*

$$\mathbb{E}\left[\left\|\mathcal{G}(\bar{x}^R, \nabla F(\bar{x}^R), \gamma)\right\|^2\right] \lesssim \frac{\gamma^{-1}C_0 + \sigma^2}{\sqrt{nK}} + \frac{n(\sigma^2 + \gamma^{-2}\nu^2)\varrho(m)}{K},$$

$$\mathbb{E}\left[\frac{L_{\nabla F}^2}{n}\left\|\mathbf{X}_R - \bar{\mathbf{X}}_R\right\|^2\right] \lesssim \frac{n(\sigma^2 + \gamma^{-2}\nu^2)\varrho(m)}{K}.$$

(Prox-DASA-GT) *For Algorithm 3, we have*

$$\mathbb{E}\left[\left\|\mathcal{G}(\bar{x}^R, \nabla F(\bar{x}^R), \gamma)\right\|^2\right] \lesssim \frac{\gamma^{-1}C_0 + \sigma^2}{\sqrt{nK}} + \frac{n\sigma^2\varrho(m)}{K},$$

$$\mathbb{E}\left[\frac{L_{\nabla F}^2}{n}\left\|\mathbf{X}_R - \bar{\mathbf{X}}_R\right\|^2\right] \lesssim \frac{n\sigma^2\varrho(m)}{K}.$$

In Theorem 3.1 for simplicity we assume $\gamma \asymp \frac{1}{L_{\nabla F}}$, which can be relaxed to $\gamma > 0$. We have the following corollary characterizing the complexity of Algorithm 2 and 3 for finding ϵ-stationary points. The proof is immediate.

Corollary 3.1. *Under the same conditions of Theorem 3.1, provided that $K \gtrsim n^3\varrho(m)$, for any $\epsilon > 0$ the sample complexity per agent for finding ϵ-stationary points in Algorithm 2 and 3 are $\mathcal{O}(\max\{n^{-1}\epsilon^{-2}, K_T\})$ where the transient time $K_T \asymp \max\{K_0, n^3\varrho(m)\}$.*

Remark (Sample Complexity). *For a sufficiently small $\epsilon > 0$, Corrollary 3.1 implies that the sample complexity of Algorithm 2 and 3 matches the optimal lower bound $\mathcal{O}(n^{-1}\epsilon^{-2})$ in decentralized* smooth *stochastic non-convex optimization **[LDS21]**.*

Remark (Transient Time and Communication Complexity). *Our algorithms are able to achieve convergence with a single communication round per iteration, i.e., $m = 1$, leading to a topology-independent $\mathcal{O}(n^{-1}\epsilon^{-2})$ communication complexity. In this case, however, the transient time K_T still depends on ρ, as is also the case for smooth optimization problems **[XKK21]**. If we consider multiple consensus steps per iteration with the communication complexity being $\mathcal{O}(mn^{-1}\epsilon^{-2})$, setting $m \asymp \lceil\frac{1}{1-\rho}\rceil$ (or $m \asymp \lceil\frac{1}{\sqrt{1-\rho}}\rceil$ for accelerated consensus algorithms) results in a topology-independent transient time given that $\varrho(m) \asymp 1$.*

3.3.3. Proof Sketch. Here, we present a sketch of our convergence analyses and defer details to Appendix. Our proof relies on the merit function below:

$$W(\bar{x}^k, \bar{z}^k) = \underbrace{\Phi(\bar{x}^k) - \Phi_*}_{\text{function value gap}} + \underbrace{\Psi(\bar{x}^k) - \eta(\bar{x}^k, \bar{z}^k)}_{\text{primal convergence}} + \lambda \underbrace{\left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2}_{\text{dual convergence}},$$

where $\eta(x,z) = \min\limits_{y \in \mathbb{R}^d} \left\{ \langle z, y - x\rangle + \frac{1}{2\gamma}\|y - x\|^2 + \Psi(y) \right\}$. Let $y_+^k := \mathbf{prox}_{\gamma\Psi}\left(\bar{x}^k - \gamma \bar{z}^k\right)$. Then, the proximal gradient mapping of $\bar{z}^k$ at $\bar{x}^k$ is $\mathcal{G}(\bar{x}^k, \bar{z}^k, \gamma) = \frac{1}{\gamma}(\bar{x}^k - y_+^k)$. Since y_+^k is the minimizer of a $1/\gamma$-strongly convex function, we have

$$\left\langle \bar{z}^k, y_+^k - \bar{x}^k \right\rangle + \frac{1}{2\gamma}\|y_+^k - \bar{x}^k\|^2 + \Psi(y_+^k) \leq \Psi(\bar{x}^k) - \frac{1}{2\gamma}\|y_+^k - \bar{x}^k\|^2,$$

which implies $\Psi(\bar{x}^k) - \eta(\bar{x}^k, \bar{z}^k) \geq \frac{\gamma}{2}\left\|\mathcal{G}(\bar{x}^k, \bar{z}^k, \gamma)\right\|^2$.

Following standard practices in optimization, we set $\gamma = \frac{1}{L_{\nabla F}}$ below for simplicity. However, our algorithms do not require any restriction on the choice of γ.

Step 1: Leveraging the merit function with $\lambda \asymp \gamma$, we can first obtain an essential lemma (Lemma 11 in Appendix) in our analyses, which says that for sequences $\{x_i^k, z_i^k\}_{1 \leq i \leq n, k \geq 0}$ generated by `Prox-DASA(-GT)` (Algorithm 2 or 3) with $\alpha_k \lesssim \min\{1, (1+\gamma)^{-2}, \gamma^2(1+\gamma)^{-2}\}$, we have

$$W(\bar{x}^{k+1}, \bar{z}^{k+1}) - W(\bar{x}^k, \bar{z}^k) \leq -\alpha_k \left\{ \Theta^k + \Upsilon^k + \alpha_k \Lambda^k + r^{k+1} \right\},$$

where $\mathbb{E}[r^{k+1} \mid \mathscr{F}_k] = 0$, $\Lambda^k \asymp \gamma \left\|\bar{\Delta}^{k+1}\right\|^2$,

$$\Theta^k \asymp \frac{1}{\gamma}\|\bar{x}^k - \bar{y}^k\|^2 + \gamma \left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2,$$
$$\Upsilon^k \asymp \frac{\gamma}{n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2 + \frac{1}{n\gamma}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2,$$

and $\bar{\Delta}^{k+1} = \bar{v}^{k+1} - \frac{1}{n}\sum_{i=1}^n \nabla F_i(x_i^k) = \bar{u}^{k+1} - \frac{1}{n}\sum_{i=1}^n \nabla F_i(x_i^k)$ (for `Prox-DASA-GT`). Thus, by telescoping and taking expectation with respect to $\mathscr{F}_0$, we have

$$\begin{aligned} &\sum_{k=0}^K \alpha_k \mathbb{E}\left[\left\|\bar{x}^k - \bar{y}^k\right\|^2 + \gamma^2 \left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2\right] \\ &\lesssim \gamma W(\bar{x}^0, \bar{z}^0) + \gamma^2\sigma^2 \boxed{\sum_{k=0}^K \frac{\alpha_k^2}{n}} + \sum_{k=0}^K \frac{\alpha_k \left\{\mathbb{E}\left[\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 + \gamma^2 \left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2\right]\right\}}{n}. \end{aligned} \tag{3.4}$$

Step 2: We then analyze the consensus errors. Without loss of generality, we consider $\mathbf{X}_0 = \bar{\mathbf{X}}_0 = \mathbf{0}$, i.e., all nodes have the same initialization at $\mathbf{0}$. For $m \in \mathbb{N}_+$, define

$$\varrho(m) = \frac{(1+\rho^{2m})\rho^{2m}}{(1-\rho^{2m})^2}.$$

Then, we have the following fact:

- $\varrho(m)$ is monotonically decreasing with the maximum value being $\varrho(1) = \frac{(1+\rho^2)\rho^2}{(1-\rho^2)^2} := \varrho_1$;
- $\varrho(m) \leq 1$ if and only if $\rho^{2m} \leq \frac{1}{3}$.

With the definition of $\varrho(m)$ and assuming $0 < \alpha_{k+1} \leq \alpha_k \leq 1$, we can show the consensus errors have the following upper bounds.

`Prox-DASA`: Let $\alpha_k \lesssim \varrho(m)^{-\frac{1}{2}}$, we have

$$\sum_{k=0}^{K} \frac{\alpha_k}{n} \mathbb{E}\left[\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2\right] \leq \sum_{k=0}^{K} \frac{\gamma^2 \alpha_k}{n} \mathbb{E}\left[\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2\right] \lesssim (\gamma^2\sigma^2 + \nu^2)\varrho(m) \boxed{\sum_{k=0}^{K} \alpha_k^3}. \tag{3.5}$$

`Prox-DASA-GT`: Let $\alpha_k \lesssim \min\{\varrho(m)^{-1}, \varrho(m)^{-\frac{1}{2}}\}$, we have

$$\begin{aligned} \sum_{k=0}^{K} \frac{\alpha_k}{n} \mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2\right] &\leq \sum_{k=0}^{K} \frac{\gamma^2 \alpha_k}{n} \mathbb{E}\left[\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2\right] \\ &\lesssim \varrho(m)^2 \boxed{\sum_{k=0}^{K} \alpha_k^3} \left\{\gamma^2\sigma^2 + \alpha_k^2 \mathbb{E}\left[\|\bar{x}^k - \bar{y}^k\|^2\right]\right\}. \end{aligned} \tag{3.6}$$

We can also see that to obtain a topology-independent iteration complexity, the number of communication rounds can be set as $m = \lceil \frac{\log 3}{2(1-\rho)} \rceil$, which implies $\varrho(m) \leq 1$.

In addition, we have the following fact that relates the consensus error of $\mathbf{Y}$ to the consensus errors of $\mathbf{X}$ and $\mathbf{Z}$:

$$\left\|y_+^k - \bar{y}^k\right\|^2 + \frac{1}{n}\left\|\mathbf{Y}_k - \bar{\mathbf{Y}}_k\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|y_i^k - y_+^k\right\|^2 \leq \frac{2}{n}\left\{\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2 + \gamma^2\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2\right\}. \tag{3.7}$$

Step 3: Let R be a random integer with

$$\Pr(R = k) = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k}, \quad k = 1, 2, \ldots, K,$$

and dividing both sides of (3.5) by $\sum_{k=1}^{K} \alpha_k$, we can obtain that for `Prox-DASA`, the consensus error of $\mathbf{X}_R$ satisfies

$$\mathbb{E}\left[\frac{1}{n}\left\|\mathbf{X}_R - \bar{\mathbf{X}}_R\right\|^2\right] \lesssim (\gamma^2\sigma^2 + \nu^2)\varrho(m)\frac{\sum_{k=0}^{K}\alpha_k^3}{\sum_{k=1}^{K}\alpha_k}.$$

Moreover, noting that

$$\|\mathcal{G}(\bar{x}, \nabla F(\bar{x}), \gamma)\|^2 \lesssim \frac{1}{\gamma^2}\left\{\left\|\bar{x}^k - \bar{y}^k\right\|^2 + \left\|y_+^k - \bar{y}^k\right\|^2\right\} + \left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2,$$

and combining (3.4) with (3.5), we can get

$$\mathbb{E}\left[\left\|\mathcal{G}(\bar{x}^R, \nabla F(\bar{x}^R), \gamma)\right\|^2\right] \lesssim \underbrace{\boxed{\frac{W(\bar{x}^0, \bar{z}^0)}{\gamma\sum_{k=1}^{K}\alpha_k}}}_{\text{initialization-related term}} + \underbrace{\boxed{\sigma^2\frac{\sum_{k=0}^{K}\alpha_k^2}{n\sum_{k=1}^{K}\alpha_k}}}_{\text{variance-related term}} + \underbrace{\boxed{(\sigma^2 + \gamma^{-2}\nu^2)\varrho(m)\frac{\sum_{k=0}^{K}\alpha_k^3}{\sum_{k=1}^{K}\alpha_k}}}_{\text{consensus error}}.$$

Thus, setting $\alpha_k \asymp \sqrt{\frac{n}{K}}$, we obtain the convergence results of `Prox-DASA`:

$$\mathbb{E}\left[\left\|\mathcal{G}(\bar{x}^R, \nabla F(\bar{x}^R), \gamma)\right\|^2\right] \lesssim \frac{\gamma^{-1}W(\bar{x}^0, \bar{z}^0) + \sigma^2}{\sqrt{nK}} + \frac{n(\sigma^2 + \gamma^{-2}\nu^2)\varrho(m)}{K},$$

$$\mathbb{E}\left[\frac{1}{\gamma^2 n}\left\|\mathbf{X}_R - \bar{\mathbf{X}}_R\right\|^2\right] \lesssim \frac{n(\sigma^2 + \gamma^{-2}\nu^2)\varrho(m)}{K}.$$

For `Prox-DASA-GT`, we can complete the proof with similar arguments by combining (3.6) with (3.4) and noting that $\varrho(m)^2\alpha_k^4 \lesssim 1$.

3.4. Experiments

3.4.1. Synthetic Data. To demonstrate the effectiveness of our algorithms, we first evaluate our algorithms using synthetic data for solving sparse single index models [**AB13**] in the decentralized setting. We consider the homogeneous setting where the data sample at each node $\xi = (X, Y)$ is generated from the same single index model $Y = g(X^\top\theta_*) + \varepsilon$, where $X, \theta \in \mathbb{R}^d$ and $\mathbb{E}[\varepsilon|X] = 0$. In this case, we solve the following L_1-regularized least square problems:

$$\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[(Y - g(X^\top\theta))^2\right] + \lambda\|\theta\|_1$$

In particular, we set $\theta_* \in \mathbb{R}^{100}$ to be a sparse vector and $g(\cdot) = (\cdot)^2$ which corresponds to the sparse phase retrieval problem [**JEH16**]. We simulate streaming data samples with batch size = 1 for training and 10,000 data samples per node for evaluations, where X and ϵ are sampled independently

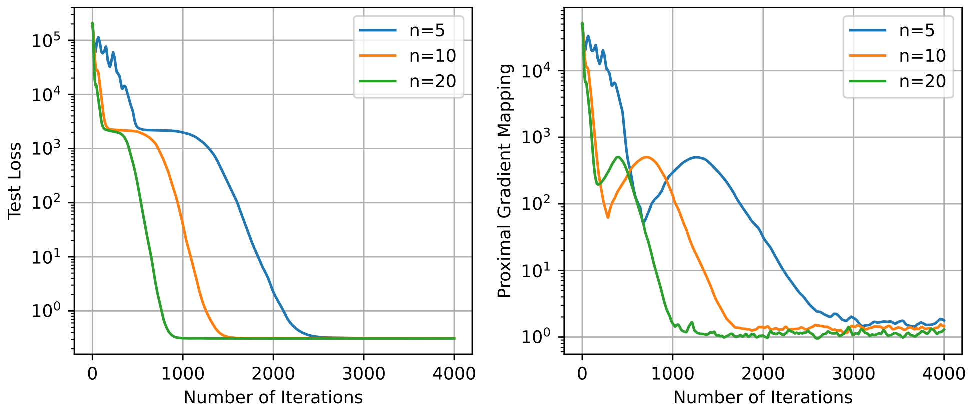


FIGURE 3.1. Linear-speedup performance of `Prox-DASA` for decentralized online sparse phase retrievel problems. (`Prox-DASA-GT` has relatively the same plots)

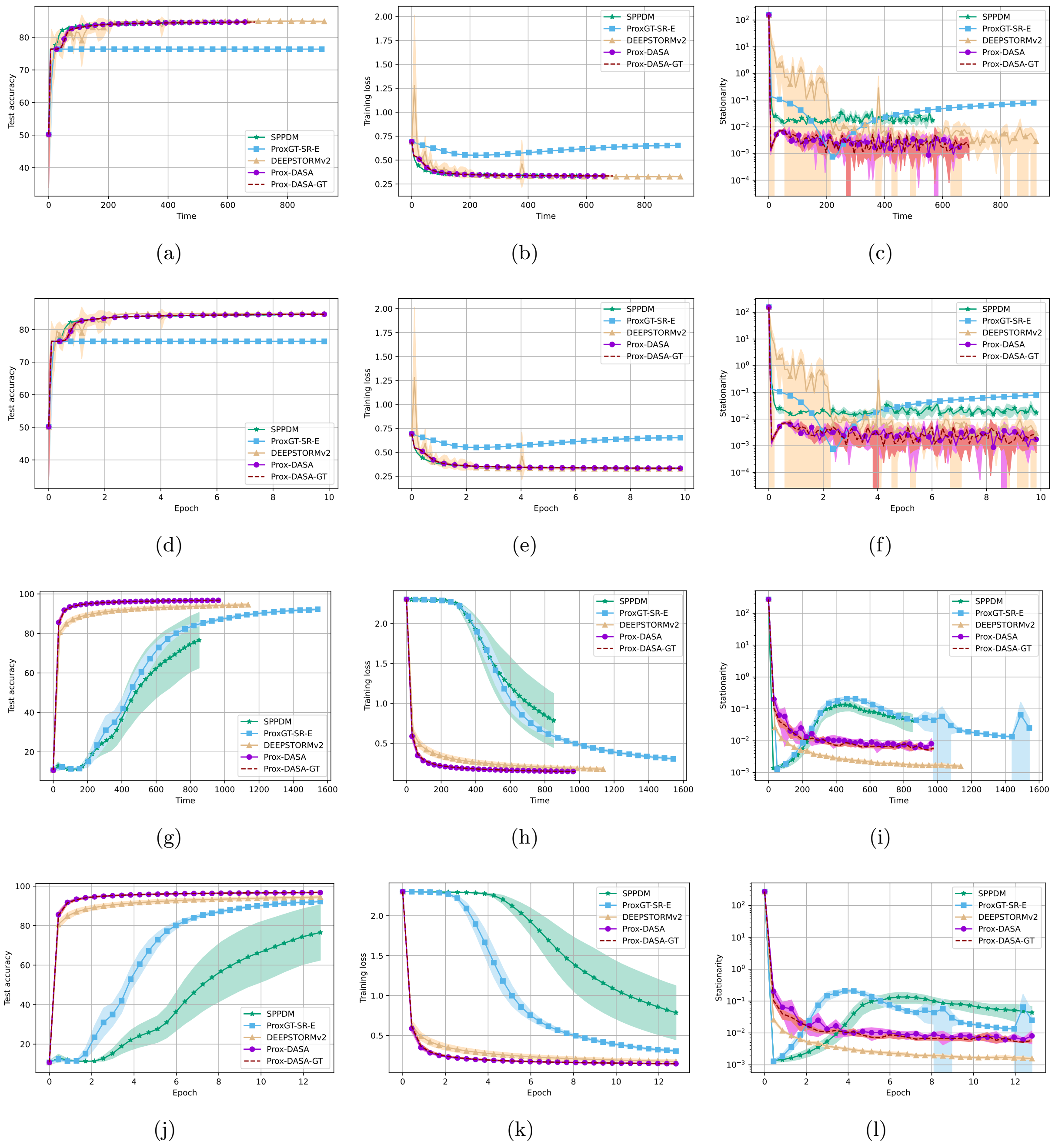
from two gaussian distributions. We employ a ring topology for the network where self-weighting and neighbor weights are set to be 1/3. We set the penalty parameter $\lambda = 0.01$, the total number of iterations $K = 10,000$, $\alpha_k = \sqrt{n/K}$, $\gamma = 0.01$, and the number of communication rounds per iteration $m = \lceil \frac{1}{1-\rho} \rceil$. We plot the test loss and the norm of proximal gradient mapping in the log scale against the number of iterations in Figure 3.1, which shows that our decentralized algorithms have an additional linear speed-up with respect to n. In other words, the algorithms become faster as more agents are added to the network.

3.4.2. Real-World Data. Following [**MBMXC22**], we consider solving the classification problem

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|\mathcal{D}_i|} \sum_{(x,y)\in\mathcal{D}_i} \ell_i(f(\theta, x), y) + \lambda \|\theta\|_1, \tag{3.8}$$

on a9a and MNIST datasets[1]. Here, ℓ_i denotes the cross entropy loss and f represents a neural network parameterized by θ with x being its input. $\mathcal{D}_i$ is the training set only available to agent i. The L_1 regularization term is used to impose a sparsity structure on the neural network. We use the code in [**MBMXC22**] for `SPPDM`, `ProxGT-SR-O/E`, `DEEPSTORM`, and then implement `Prox-DASA` and `Prox-DASA-GT` under their framework, which mainly utilizes PyTorch [**PGM+19**] and mpi4py [**DF21**]. We use a 2-layer perception model on a9a and the LeNet architecture [**L+15**] for the MNIST dataset. We have 8 agents which connect in the form of a ring for a9a and a random graph for MNIST. To

[1] Available at https://www.openml.org.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

(j)

(k)

(l)

FIGURE 3.2. Comparisons between SPPDM [**WZC**$^+$**21**], ProxGT-SR-E [**XDKK21**], DEEPSTORM [**MBMXC22**], Prox-DASA 2, and Prox-DASA-GT 3. The first two rows correspond to a9a and the last two rows correspond to MNIST. The results are averaged over 10 trials, and the shaded regions represent confidence intervals. The vertical axes in the third column are log-scale. It should be noted that ProxGT-SR-E maintains another hyperparameter q (see, e.g., Algorithm 4 and Theorem 3 in [**XDKK21**]) and computes gradients using a full batch every q iterations. For simplicity, we do not include that amount of epochs when we plot this figure. In other words, the real number of epochs required to obtain a point on ProxGT-SR is larger than plotted in the figures in the second and fourth rows. We include the plots that take q into account in Figure A.2.

demonstrate the performance of our algorithms in the constant batch size setting, the batch size is chosen to be 4 for a9a and 32 for MNIST for all algorithms. The number of communication rounds per iteration m is set to be 1 for all algorithms. We evaluate the model performance periodically during training, and then plot the results in Figure 3.2, from which we observe that both `Prox-DASA` and `Prox-DASA-GT` have considerably good performance with small variance in terms of test accuracy, training loss, and stationarity. In particular, it should be noted that although `DEEPSTORM` achieves better stationarity in Figure 3.2(l) and 3.2(i), training a neural network by using `DEEPSTORM` takes longer time than `Prox-DASA` and `Prox-DASA-GT` since it uses the momentum-based variance reduction technique, which requires **two forward-backward passes** (see, e.g., Eq. (10) and Algorithm 1 in [**MBMXC22**]) to compute the gradients in one iteration per agent while ours only require **one**, which saves a large amount of time (see Table A.1 in Appendix). We include further details of our experiments in Appendix A.1.

3.5. Discussion and Conclusion

In this work, we propose and analyze a class of single time-scale decentralized proximal algorithms (`Prox-DASA-(GT)`) for non-convex stochastic composite optimization in the form of (3.1). We show that our algorithms achieve linear speed-up with respect to the number of agents using an $\mathcal{O}(1)$ batch size per iteration under mild assumptions. Furthermore, we demonstrate the efficiency and effectiveness of our algorithms through extensive experiments, in which our algorithms achieve relatively better results with less training time using a small batch size comparing to existing methods.

CHAPTER 4

Conditional Gradient-Based Nested Averaged Stochastic Approximation

4.1. Introduction

We study projection-free algorithms for solving the following stochastic multi-level composition optimization problem

$$\min_{x\in\mathcal{X}} \quad F(x) := f_1 \circ \cdots \circ f_T(x), \tag{4.1}$$

where $f_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i-1}}, i = 1, \cdots, T$ $(d_0 = 1)$ are continuously differentiable functions, the composite function F is bounded below by $F^\star > -\infty$ and $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex set. We assume that the exact function values and derivatives of f_i's are not available. In particular, we assume that $f_i(y) = \mathbb{E}_{\xi_i}[G_i(y, \xi_i)]$ for some random variable ξ_i. Our goal is to solve the above optimization problem, given access to noisy evaluations of ∇f_i's and f_i's.

There are two main challenges to address in developing efficient projection-free algorithms for solving (4.1). First, note that denoting the transpose of the Jacobian of f_i by ∇f_i, the gradient of the objective function $F(x)$ in (4.1), is given by $\nabla F(x) = \nabla f_T(y_T)\nabla f_{T-1}(y_{T-1}) \cdots \nabla f_1(y_1)$, where $y_i = f_{i+1} \circ \cdots \circ f_T(x)$ for $1 \le i < T$, and $y_T = x$. Because of the nested nature of the gradient $\nabla F(x)$, obtaining an unbiased gradient estimator in the stochastic first-order setting, with controlled moments, becomes non-trivial. Using naive stochastic gradient estimators lead to oracle complexities that depend exponentially on T (in terms of the accuracy parameter). Next, even when $T = 1$ in the stochastic setting, projection-free algorithms like the conditional gradient method or its sliding variants invariably require increasing order of mini-batches[1] [**LZ16**, **RSPS16**, **HL16**, **QLX18**, **YSC19**], which make their practical implementation infeasible. In the general T-level setting, using naive stochastic gradient estimator would lead to mini-batch order that depends exponentially on T.

[1] We discuss in detail about some recent works that avoid requiring increasing mini-batches, albeit under stronger assumptions, in Section 4.1.3.

In this work, we propose a projection-free conditional gradient-type algorithm that achieves *level-independent* oracle complexities (i.e., the dependence of the complexities on the target accuracy is independent of T) using only *one sample* of $(\xi_i)_{1\leq i\leq T}$ in each iteration, thereby addressing both of the above challenges. Our algorithm uses moving-average based stochastic estimators of the gradient and function values, also used recently by [**GRW20**] and [**BGN22**], along with an inexact conditional gradient method used by [**BG22**] (which in turn is inspired by the sliding method by [**LZ16**]). In order to establish our oracle complexity results, we use a novel merit function based convergence analysis. To the best of our knowledge, such an analysis technique is used for the first time in the context of analyzing stochastic conditional gradient-type algorithms.

4.1.1. Motivating Examples. Problems of the form in (4.1) are generalizations of the standard constrained stochastic optimization problem, which is obtained when $T = 1$, and arise in several machine learning applications. Some examples include sparse additive modeling in non-parametric statistics [**WFL17**, Section 4.1], Bayesian optimization [**AF21**], model-agnostic meta-learning [**CSY21**, **FMO21**], distributionally robust optimization [**QGX$^+$21**], training graph neural networks [**CFKM20**], reinforcement learning [**WLF16**, Setion 1.1] and AUPRC maximization [**QLX$^+$21**, **WYZY22**, **QHZ$^+$22**]. Below, we provide a concrete motivating example from the field of risk-averse stochastic optimization [**RS06**].

The mean-deviation risk-averse optimization is given by the following optimization problem

$$\max_{x\in\mathcal{X}} \left\{ \mathbb{E}[U(x,\xi)] - \rho\mathbb{E}\left[\{\mathbb{E}[U(x,\xi)] - U(x,\xi)\}^2 \right]^{1/2} \right\}.$$

As noted by [**YWF19**], [**Rus21**] and [**BGN22**], the above problem is a stochastic 3-level composition optimization problem with

$$f_3 := \mathbb{E}[U(x,\xi)] \qquad f_2(z,x) := \mathbb{E}[\{z - U(x,\xi)\}^2] \qquad f_1((y_1,y_2)) := y_1 - \sqrt{y_2+\delta}.$$

Here, $\delta > 0$ is added to make the square root function smooth. In particular, we consider a semi-parametric data generating process given by a sparse single-index model of the form $b = g(\langle a, x^*\rangle) + \zeta$, where $g : \mathbb{R} \to \mathbb{R}$ is called the link function, $x^* \in \mathbb{R}^d$ is assumed to be a sparse vector and $\langle\cdot,\cdot\rangle$ represents the Euclidean inner-product between two vectors. Such single-index models are widely used in statistics, machine learning and economics [**RWC03**]. A standard choices of the link function g is the square function, in which case, the model is also called as the sparse phase retrieval model [**WGE17**]. Here, a is the input data which is assumed to be independent of the noise ζ. In this case, $\xi := (a, b)$ and the if we consider the squared-loss, then $U(x, \xi) := (b - (\langle a, x\rangle)^2)^2$ and is non-convex in x. The goal is to estimate the sparse index vector x^* in a risk-averse manner, as they are well-known to provide stable solutions [**YWF19**]. To encourage sparsity, the set $\mathcal{X}$ is the ℓ_1 ball [**Jag13**].

4.1.2. Preliminaries and Main Contributions. We now introduce the technical preliminaries required to state and highlight the main contributions of this work. Throughout this work, $\|\cdot\|$ denotes the Euclidean norm for vectors and the Frobenius norm for matrices. We first describe the set of assumptions on the objective functions and the constraint set.

ASSUMPTION 4.1 (Constraint set). *The set $\mathcal{X} \subset \mathbb{R}^d$ is convex and closed with $\max_{x,y\in\mathcal{X}} \|x-y\| \leq D_{\mathcal{X}}$.*

ASSUMPTION 4.2 (Smoothness). *All functions $f_1, \ldots, f_T$ and their derivatives are Lipschitz continuous with Lipschitz constants L_{f_i} and $L_{\nabla f_i}$, respectively.*

The above assumptions on the constraint set and the objective function are standard in the literature on stochastic optimization, and in particular in the analysis of conditional gradient algorithms and multi-level optimization; see, for example, [**LZ16**], [**YWF19**] and [**BGN22**]. We emphasize here that the above smoothness assumptions are made only on the functions $(f_i)_{1\leq i\leq T}$ and not on the stochastic functions $(G_i)_{1\leq i\leq T}$ (which would be a stronger assumption). Moreover, the Lipschitz continuity of f_i's are implied by the Assumption 4.1 and assuming the functions are continuously differentiable. However, for sake of illustration, we state both assumptions separately. In addition to these assumptions, we also make unbiasedness and bounded-variance assumptions on the stochastic first-order oracle. Due to its technical nature, we state the precise details later in Section 4.3 (see Assumption 4.3).

We next turn our attention to the convergence criterion that we use in this work to evaluate the performance of the proposed algorithm. Gradient-based algorithms iteratively solve sub-problems in the form of

$$\operatorname*{arg\,min}_{u\in\mathcal{X}} \left\{ \langle g, u\rangle + \frac{\beta}{2}\|u - x\|^2 \right\}, \tag{4.2}$$

for some $\beta > 0$, $g \in \mathbb{R}^d$ and $x \in \mathcal{X}$. Denoting the optimal solution of the above problem by $P_{\mathcal{X}}(x, g, \beta)$ and noting its optimality condition, one can easily show that

$$-\nabla F(\bar{x}) \in \mathcal{N}_{\mathcal{X}}(\bar{x}) + \mathcal{B}\Big(0, \|g - \nabla F(\bar{x})\| D_{\mathcal{X}} + \beta \|x - P_{\mathcal{X}}(x, g, \beta)\| (D_{\mathcal{X}} + \|\nabla F(\bar{x})\|/\beta)\Big),$$

where $\mathcal{N}_{\mathcal{X}}(\bar{x})$ is the normal cone to $\mathcal{X}$ at $\bar{x}$ and $\mathcal{B}(0, r)$ denotes a ball centered at the origin with radius r. Thus, reducing the radius of the ball in the above relation will result in finding an approximate first-order stationary point of the problem for non-convex constrained minimization problems. Motivated by this fact, we can define the gradient mapping at point $\bar{x} \in \mathcal{X}$ as

$$\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta) := \beta \left(\bar{x} - P_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta)\right) = \beta \left(\bar{x} - \mathbf{proj}_{\mathcal{X}} \left(\bar{x} - \frac{1}{\beta} \nabla F(\bar{x})\right)\right), \tag{4.3}$$

where $\mathbf{proj}_{\mathcal{X}}(y)$ denotes the Euclidean projection of the vector y onto the set $\mathcal{X}$. The gradient mapping is a classical measure has been widely used in the literature as a convergence criterion when solving nonconvex constrained problems [**N$^+$18**]. It plays an analogous role of the gradient for constrained optimization problems; in fact when the set $\mathcal{X} \equiv \mathbb{R}^d$ the gradient mapping just reduces to $\nabla F(\bar{x})$. It should be emphasized that while the gradient mapping cannot be computed in the stochastic setting, it still serves as a measure of convergence. Our main goal in this work is to find an ϵ-stationary solution to (4.1), in the sense described below.

Definition 4.1. *A point $\bar{x} \in \mathcal{X}$ generated by an algorithm for solving* (4.1) *is called an ϵ-stationary point, if we have $\mathbb{E}[\|\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta)\|^2] \le \epsilon$, where the expectation is taken over all the randomness involved in the problem.*

In the literature on conditional gradient methods for the nonconvex setting, the so-called Frank-Wolfe gap is also widely used to provide convergence analysis. The Frank-Wolfe Gap is defined formally as

$$g_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x})) := \max_{y \in \mathcal{X}} \langle \nabla F(\bar{x}), \bar{x} - y \rangle. \tag{4.4}$$

As pointed out by [**BG22**], the gradient mapping criterion and the Frank-Wolfe gap are related to each other in the following sense.

Proposition 4.1. [**BG22**] *Let $g_{\mathcal{X}}(\cdot)$ be the Frank-Wolfe gap defined in* (4.4) *and $\mathcal{G}_{\mathcal{X}}(\cdot)$ be the gradient mapping defined in* (4.3)*. Then, we have $\|\mathcal{G}_{\mathcal{X}}(x, \nabla F(x), \beta)\|^2 \leq g_{\mathcal{X}}(x, \nabla F(x)), \forall x \in \mathcal{X}$. Moreover, under Assumption 4.1, 4.2, we have $g_{\mathcal{X}}(x, \nabla F(x)) \leq \left[(1/\beta) \prod_{i=1}^{T} L_{f_i} + D_{\mathcal{X}}\right] \|\mathcal{G}_{\mathcal{X}}(x, \nabla F(x), \beta)\|$.*

For stochastic conditional gradient-type algorithms, the oracle complexity is measured in terms of number of calls to the Stochastic First-order Oracle (SFO) and the Linear Minimization Oracle (LMO) used to the solve the sub-problems (that are of the form of minimizing a linear function over the convex feasible set) arising in the algorithm. In this work, we hence measure the number of calls to SFO and LMO required by the proposed algorithm to obtain an ϵ-stationary solution in the sense of Definition 4.1. We now highlight our **main contributions**:

- We propose a projection-free conditional gradient-type method (Algorithm 5) for solving (4.1). In Theorem 4.1, we show that the SFO and LMO complexities of this algorithm, in order to achieve an ϵ-stationary solution in the sense of Definition 4.1, are of order $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-3})$, respectively.
- The above SFO and LMO complexities are in particular level-independent (i.e., the dependence of the complexities on the target accuracy is independent of T). The proposed algorithm is parameter-free and does not require any mini-batches, making it applicable for the online setting.
- When considering the case of $T \leq 2$, we present a simplified method (Algorithm 7 and 8) to obtain the same oracle complexities. Intriguingly, while this simplified method is still parameter-free for $T = 1$, it is not when $T = 2$ (see Theorem 4.2 and Remark 4.3.1). Furthermore, for the case of $T = 1$, we also establish high-probability bounds (see Theorem 4.3).

A summary of oracle complexities for stochastic conditional gradient-type algorithms is in Table 4.1.

4.1.3. Related Work. Conditional Gradient-Type Method. The conditional gradient algorithm [**FW56**, **LP66**], has had a renewed interest in the machine learning and optimization communities in the past decade; see [**Mig94**, **Jag13**, **HJN15**, **LJJ15**, **BS17**, **GKS21**] for a partial list of recent works. Considering the stochastic convex setup, [**HK12**, **HL16**] provided expected oracle complexity results for the stochastic conditional gradient algorithm. The complexities were further improved by a sliding procedure in [**LZ16**], based on Nesterov's acceleration method. [**RSPS16**, **YSC19**, **HL16**] considered variance reduced stochastic conditional gradient algorithms, and provided

Algorithm	**Criterion**	**# of levels**	**Batch size**	**SFO**	**LMO**
SPIFER-SFW [**YSC19**]	FW-gap	1	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$
1-SFW [**ZSM$^+$20**]	FW-gap	1	1	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$
SCFW [**ABTR21**]	FW-gap	2	1	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$
SCGS [**QLX18**]	GM	1	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
SGD+ICG [**BG22**]	GM	1	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
LiNASA+ICG (Algorithm 5)	GM	T	1	$\mathcal{O}_T(\epsilon^{-2})$	$\mathcal{O}_T(\epsilon^{-3})$

TABLE 4.1. Complexity results for stochastic conditional gradient type algorithms to find an ϵ-stationary solution in the nonconvex setting. FW-Gap and GM stands for Frank-Wolfe Gap (see (4.4)) and Gradient Mapping (see (4.3)) respectively. $\mathcal{O}_T$ hides constants in T. Existing one-sample based stochastic conditional gradient algorithms are either (i) not applicable to the case of general $T > 1$, or (ii) require strong assumptions [**ZSM$^+$20**], or (iii) are not truly online [**ABTR21**]; see Section 4.1.3 for detailed discussion. The results in [**BG22**] are actually presented for the zeroth-order setting; however the above stated first-order complexities follow immediately.

expected oracle complexities in the non-convex setting. [**QLX18**] analyzed the sliding algorithm in the non-convex setting and provided results for the gradient mapping criterion. *All of the above works require increasing orders of mini-batches to obtain their oracle complexity results.*

[**MHK20**] and [**ZSM$^+$20**] proposed a stochastic conditional gradient-type algorithm with moving-average gradient estimator for the convex and non-convex setting that uses only one-sample in each iteration. However, [**MHK20**] and [**ZSM$^+$20**] require several restrictive assumptions, which we explain next (focusing on [**ZSM$^+$20**] which considers the nonconvex case). Specifically, [**ZSM$^+$20**] requires that the stochastic function $G_1(x, \xi_1)$ has uniformly bounded function value, gradient-norm, and Hessian spectral-norm, and the distribution of the random vector ξ_1 has an absolutely continuous density p such that the norm of the gradient of $\log p$ and spectral norm of the Hessian of $\log p$ has finite fourth and second-moments respectively. In contrasts, we do not require such stringent assumptions. Next, all of the above works focus only on the case of $T = 1$. [**ABTR21**] considered stochastic conditional gradient algorithm for solving (4.1), with $T = 2$. However, [**ABTR21**] also makes stringent assumptions: (i) the stochastic objective functions $G_1(x, \xi_1)$ and $G_2(x, \xi_1)$ themselves have Lipschitz gradients almost surely and (ii) for a given instance of random vectors ξ_1 and ξ_2, one could query the oracle at the current and previous iterations, which makes the algorithm not to be truly online. See Table 4.1 for a summary.

Stochastic Multi-level Composition Optimization. Compositional optimization problems of the form in (4.1) have been considered as early as 1970s by [**Erm76**]. Recently, there has been a renewed interest on this problem. [**EN13**] and [**DPR17**] considered a sample-average approximation approach for solving (4.1) and established several asymptotic results. For the case of $T = 2$, [**WFL17**], [**WLF16**] and [**BGI$^+$17**] proposed and analyzed stochastic gradient descent-type algorithms in the smooth setting. [**DD19**] and [**DR18**] considered the non-smooth setting and established oracle complexity results. Furthermore, [**HZCH20**] proposed algorithms when the randomness between the two levels are not necessarily independent. For the general case of $T \geq 1$, [**YWF19**] proposed stochastic gradient descent-type algorithms and established oracle complexities established that depend exponentially on T and are hence sub-optimal. Indeed, large deviation and Central Limit Theorem results established in [**EN13**, **DPR17**], respectively, show that in the sample-average approximation setting, the $\arg\min$ of the problem in (4.1) based on n samples, converges at a level-independent rate (i.e., dependence of the convergence rate on the target accuracy is independent of T) to the true minimizer, under suitable regularity conditions.

[**GRW20**] proposed a single time-scale Nested Averaged Stochastic Approximation (NASA) algorithm and established optimal rates for the cases of $T = 1, 2$. For the general case of $T \geq 1$, [**BGN22**] proposed a linearized NASA algorithm and established level-independent and optimal convergence rates. Concurrently, [**Rus21**] considered the case when the function f_i are non-smooth and established asymptotic convergence results. [**ZX21**] also established non-asymptotic level-independent oracle complexities, however, under stronger assumptions than that in [**BGN22**]. Firstly, they assumed that for a fixed batch of samples, one could query the oracle on different points, which is not suited for the general online stochastic optimization setup. Next, they assume a much stronger mean-square Lipschitz smoothness assumption on the individual functions f_i and their gradients. Finally, they required mini-batches sizes that depend exponentially on T, which makes their method impractical. Concurrent to [**BGN22**], level-independent rates were also obtained for *unconstrained* problems by [**CSY21**], albeit, under the stronger assumption that the stochastic functions $G_i(x, \xi_i)$ are Lipschitz, almost surely. It is also worth mentioning that while some of the above papers considered constrained problems, the algorithms proposed and analyzed in the above works are not projection-free, which is the main focus of this work.

4.2. Methodology

In this section, we present our projection-free algorithm for solving problem (4.1). The method generates three random sequences, namely, approximate solutions $\{x^k\}$, average gradients $\{z^k\}$, and average function values $\{u^k\}$, defined on a certain probability space $(\Omega, \mathscr{F}, P)$. We let $\mathscr{F}_k$ to be the σ-algebra generated by $\{x^0, \ldots, x^k, z^0, \ldots, z^k, u_1^0, \ldots, u_1^k, \ldots, u_T^0, \ldots, u_T^k\}$. The overall method is given in Algorithm 5. In (4.7), the stochastic Jacobians $J_i^{k+1} \in \mathbb{R}^{d_i \times d_{i-1}}$, and the product $\prod_{i=1}^{T} J_{T+1-i}^{k+1}$ is calculated as $J_T^{k+1} J_{T-1}^{k+1} \cdots J_1^{k+1} \in \mathbb{R}^{d_T \times d_1} \equiv \mathbb{R}^{d_T \times 1}$. In (4.8), we use the notation $\langle \cdot, \cdot \rangle$ to represent both matrix-vector multiplication and vector-vector inner product. There are two aspects of the algorithm that we highlight specifically: (i) In addition to estimating the gradient of F, we also estimate a stochastic linear approximation of the inner functions f_i by a moving-average technique. In the multi-level setting we consider, it helps us to avoid the accumulation of bias, when estimating the f_i directly. Linearization techniques were used in the stochastic optimization since the work of [**Rus87**]. A similar approach was used in [**BGN22**] in the context of projected-based methods for solving (4.1). It is also worth mentioning that other linearization techniques have been used in [**DD19**] and [**DR18**] for estimating the stochastic inner function values for weakly convex two-level composition problems, and (ii) The ICG method given in Algorithm 6 is essentially applying *deterministic* conditional gradient method with the exact line search for solving the quadratic minimization subproblem in (4.2) with the estimated gradient z_k in (4.7). It was also used in [**BG22**] as a sub-routine and is motivated by the sliding approach of [**LZ16**].

4.3. Convergence Analysis

In this section, we present our main result on the oracle complexity of Algorithm 5. Before we proceed, we present our assumptions on the stochastic first-order oracle.

Assumption 4.3 (Stochastic First-Order Oracle). *Denote $u_{T+1}^k \equiv x^k$. For each k, u_{i+1}^k being the input, the stochastic oracle outputs $G_i^{k+1} \in \mathbb{R}^{d_i}$ and J_i^{k+1} such that given $\mathscr{F}_k$ and for any $i \in \{1, \ldots, T\}$*

(a) $\mathbb{E}[J_i^{k+1} | \mathscr{F}_k] = \nabla f_i(u_{i+1}^k)$, $\mathbb{E}[G_i^{k+1} | \mathscr{F}_k] = f_i(u_{i+1}^k)$,

(b) $\mathbb{E}[\|G_i^{k+1} - f_i(u_{i+1}^k)\|^2 | \mathscr{F}_k] \leq \sigma_{G_i}^2$, $\mathbb{E}[\|J_i^{k+1} - \nabla f_i(u_{i+1}^k)\|^2 | \mathscr{F}_k] \leq \sigma_{J_i}^2$,

Algorithm 5 Linearized NASA with Inexact Conditional Gradient Method (`LiNASA+ICG`)

```
Input: $x^0 \in \mathcal{X}$, $z^0 = 0 \in \mathbb{R}^d$, $u_i^0 \in \mathbb{R}^{d_i}, i = 1, \dots, T$, $\beta_k > 0$, $t_k > 0$, $\tau_k \in (0, 1]$, $\delta \geq 0$.
for $k = 0, 1, 2, \dots, N$ do
    1. Update the solution:
(4.5)        $\tilde{y}^k = \mathtt{ICG}(x^k, z^k, \beta_k, t_k, \delta),$
(4.6)        $x^{k+1} = x^k + \tau_k(\tilde{y}^k - x^k),$
    and compute stochastic Jacobians $J_i^{k+1}$, and function values $G_i^{k+1}$ at $u_{i+1}^k$ for $i = 1, \dots, T$.
    2. Update average gradients $z$ and function value estimates $u_i$ for each level $i = 1, \dots, T$
(4.7)        $z^{k+1} = (1 - \tau_k) z^k + \tau_k \prod_{i=1}^{T} J_{T+1-i}^{k+1},$
(4.8)        $u_i^{k+1} = (1 - \tau_k) u_i^k + \tau_k G_i^{k+1} + \langle J_i^{k+1}, u_{i+1}^{k+1} - u_{i+1}^k \rangle.$
end for
Output: $(x^R, z^R, u_1^R, \cdots, u_T^R)$, where $R$ is uniformly distributed over $\{1, 2, \dots, N\}$
```

Algorithm 6 Inexact Conditional Gradient Method (`ICG`)

```
Input: $(x, z, \beta, M, \delta)$
Set $w^0 = x$.
for $t = 0, 1, 2, \dots, M$ do
    1. Find $v^t \in \mathcal{X}$ with a quantity $\delta \geq 0$ such that
        $\langle z + \beta(w^t - x), v^t \rangle \leq \min_{v \in \mathcal{X}} \langle z + \beta(w^t - x), v \rangle + \frac{\beta D_{\mathcal{X}}^2 \delta}{t+2}.$
    2. Set $w^{t+1} = (1 - \mu_t) w^t + \mu_t v^t$ with $\mu_t = \min\left\{1, \frac{\langle \beta(x - w^t) - z, v^t - w^t \rangle}{\beta \|v^t - w^t\|^2}\right\}$.
end for
Output: $w^M$
```

(c) The outputs of the stochastic oracle at level i, G_i^{k+1} and J_i^{k+1}, are independent. The outputs of the stochastic oracle are independent between levels, i.e., $\{G_i^{k+1}\}_{i=1,\dots,T}$ are independent and so are $\{J_i^{k+1}\}_{i=1,\dots,T}$.

Parts (a) and (b) in Assumption 4.3 are standard unbiasedness and bounded variance assumptions on the stochastic gradient, common in the literature. Part (c) is essential to establish the convergence results in the multi-level case. Similar assumptions have been made, for example, in [**YWF19**] and [**BGN22**]. We also emphasize that unlike some prior works (see e.g., [**ZSM$^+$20**]), Assumption 4.3 allows the case of endogenous uncertainty, and we do not require the distribution of the random variables $(\xi_i)_{1 \leq i \leq T}$ to be independent of the distribution of the decision variables $(u_i)_{1 \leq i \leq T}$.

Remark. *Under Assumption 4.2, and 4.3, we can immediately conclude that* $\mathbb{E}[\|J_i^{k+1}\|^2|\mathscr{F}_k] = \mathbb{E}[\|J_i^{k+1} - \nabla f_i(u_{i+1}^k)\|^2|\mathscr{F}_k] + \|\nabla f_i(u_{i+1}^k)\|^2 \leq \sigma_{J_i}^2 + L_{f_i}^2 := \hat{\sigma}_{J_i}^2$. *In the sequel,* $\hat{\sigma}_{J_i}^2$ *will be used to simplify the presentation.*

We start with the merit function used in this work and its connection to the gradient mapping criterion. Our proof leverages the following merit function:

$$W_{\alpha,\gamma}(x, z, u) = F(x) - F^\star - \eta(x, z) + \alpha\|\nabla F(x) - z\|^2 + \sum_{i=1}^{T} \gamma_i \|f_i(u_{i+1}) - u_i\|^2, \tag{4.9}$$

where $\alpha, \{\gamma_i\}_{1\leq i\leq T}$ are positive constants and

$$\eta(x, z) = \min_{y\in\mathcal{X}} \left\{ H(y; x, z, \beta) := \langle z, y - x\rangle + \frac{\beta}{2}\|y - x\|^2 \right\}. \tag{4.10}$$

Compared to **[BGN22]**, we require the additional term $\|\nabla F(x) - z\|^2$, which turns out to be essential in our proof due to the `ICG` routine. The following proposition relates the merit function above to the gradient mapping.

Proposition 4.2. *Let* $\mathcal{G}_{\mathcal{X}}(\cdot)$ *be the gradient mapping defined in* (4.3) *and* $\eta(\cdot, \cdot)$ *be defined in* (4.10)*. For any pair of* (x, z) *and* $\beta > 0$*, we have* $\|\mathcal{G}_{\mathcal{X}}(x, \nabla F(x), \beta)\|^2 \leq -4\beta\eta(x, z) + 2\|\nabla F(x) - z\|^2$.

Proof. By expanding the square, and using the properties of projection operation, we have

$$\|\mathbf{proj}_{\mathcal{X}}(x - \frac{1}{\beta}z) - x\|^2 + \|\mathbf{proj}_{\mathcal{X}}(x - \frac{1}{\beta}z) - (x - \frac{1}{\beta}z)\|^2 \leq \|\bar{x} - (x - \frac{1}{\beta}z)\|^2 = \|\frac{1}{\beta}z\|^2.$$

Thus, we have $\eta(x, z) \leq -\frac{\beta}{2}\|\mathbf{proj}_{\mathcal{X}}(x - \frac{1}{\beta}z) - x\|^2$. The proof is completed immediately by noting that $\|\mathcal{G}(x, \nabla F(x), \beta)\|^2 \leq 2\beta^2\|\mathbf{proj}_{\mathcal{X}}(x - \frac{1}{\beta}z) - x\|^2 + 2\|\nabla F(x) - z\|^2$. □

We now present out main result on the oracle complexity of Algorithm5.

Theorem 4.1. *Under Assumption 4.1, 4.2, 4.3, let* $\{x^k, z^k, \{u_i^k\}_{1\leq i\leq T}\}_{k\geq 0}$ *be the sequence generated by Algorithm 5 with* $N \geq 1$ *and*

$$\beta_k \equiv \beta > 0, \qquad \tau_0 = 1,\ t_0 = 0, \quad \tau_k = \frac{1}{\sqrt{N}},\ t_k = \lceil\sqrt{k}\rceil, \quad \forall k \geq 1, \tag{4.11}$$

where β is an arbitrary positive constant. Provided that the merit function $W_{\alpha,\gamma}(x, z, u)$ is defined as (4.9) *with*

$$\alpha = \frac{\beta}{20L_{\nabla F}^2}, \quad \gamma_1 = \frac{\beta}{2}, \quad \gamma_j = \left(2\alpha + \frac{1}{4\alpha L_{\nabla F}^2}\right)(T-1)C_j^2 + \frac{\beta}{2}, \quad 2 \leq j \leq T, \tag{4.12}$$

we have,

$$\mathbb{E}\left[\|\mathcal{G}_{\mathcal{X}}(x^R, \nabla F(x^R), \beta)\|^2\right] \leq \frac{2(\beta + \frac{20L_{\nabla F}^2}{\beta})\left[2W_{\alpha,\gamma}(x^0, z^0, u^0) + \mathcal{B}(\beta, \sigma^2, L, D_{\mathcal{X}}, T, \delta)\right]}{\sqrt{N}}, \tag{4.13}$$

$$\mathbb{E}\left[\|f_i(u_{i+1}^R) - u_i^R\|^2\right] \leq \frac{2W_{\alpha,\gamma}(x^0, z^0, u^0) + \mathcal{B}(\beta, \sigma^2, L, D_{\mathcal{X}}, T, \delta)}{\beta\sqrt{N}}, \quad 1 \leq i \leq T. \tag{4.14}$$

where $u_{T+1} = x, \mathcal{B}(\beta, \sigma^2, L, D_{\mathcal{X}}, T, \delta) = 4\hat{\sigma}^2 + 32\beta D_{\mathcal{X}}^2(1+\delta)\left(\frac{3}{5} + \frac{5L_{\nabla F}^2}{\beta^2}\right)$, and $\hat{\sigma}^2$ is a constant depending on the parameters $(\beta, \sigma^2, L, D_{\mathcal{X}}, T)$ given in (B.26). *The expectation is taken with respect to all random sequences generated by the method and an independent random integer number R uniformly distributed over $\{1, \ldots, N\}$. That is to say, the number of calls to SFO and LMO to get an ϵ-stationary point is upper bounded by $\mathcal{O}_T(\epsilon^{-2}), \mathcal{O}_T(\epsilon^{-3})$ respectively.*

Remark. *The constant $\mathcal{B}(\beta, \sigma^2, L, D_{\mathcal{X}}, T, \delta)$ is $\mathcal{O}(T)$ given the definition of $\hat{\sigma}^2$ and the value of γ_j in* (4.12)*, which further implies that the total number of calls to SFO and LMO of Algorithm 5 for finding an ϵ-stationary point of* (4.1)*, are bounded by $\mathcal{O}(T^2\epsilon^{-2}) = \mathcal{O}_T(\epsilon^{-2})$ and $\mathcal{O}(T^3\epsilon^{-3}) = \mathcal{O}_T(\epsilon^{-3})$ respectively. Furthermore, it is worth noting that this complexity bound for Algorithm 5 is obtained without any dependence of the parameter β_k on Lipschitz constants due to the choice of arbitrary positive constant β in* (4.11)*, and τ_k, t_k depend only on the number of iterations N and k respectively. This makes Algorithm 5 parameter-free and easy to implement.*

Remark. *As discussed in Section 4.2, the* `ICG` *routine given in Algorithm 6 is a* deterministic *method with the estimated gradient z_k in* (4.7)*. The number of iterations, t_k, required to run Algorithm 6 is given by $t_k = \lceil\sqrt{k}\rceil$. That is, we require more precise solutions for the* `ICG` *routine, only for later outer iterations. Furthermore, due to the deterministic nature of the* `ICG` *routine, further advances in the analysis of deterministic conditional gradient methods under additional assumptions on the constraint set $\mathcal{X}$ (see, for example, [**GH15, GW21**]) could be leveraged to improve the overall LMO complexity.*

4.3.1. The special cases of $T = 1$ and $T = 2$. We now discuss several intriguing points regarding the choice of tuning parameter β, for the case of $T = 2$, and the more standard case of $T = 1$. Specifically, the linearization technique used in Algorithm 5 turns out to be not necessary for the case of $T = 2$ and $T = 1$ to obtain similar rates. However, without linearization, the choice of β is dependent on the problem parameters for $T = 2$. Whereas it turns out to be independent of the problem parameters (similar to Algorithm 5 and Theorem 4.1 which holds for all $T \geq 1$) for $T = 1$. As the outer function value estimates (i.e., u_1^{k+1} sequence) are not required for the convergence analysis, we remove them in Algorithms 7 and 8.

Algorithm 7 NASA with Inexact Conditional Gradient Method (`NASA+ICG`) for $T = 2$

Replace Step 2 of Algorithm 5 with the following:

2'. Update the average gradient z and the function value estimate u_2 respectively as:

$$z^{k+1} = (1 - \tau_k)z^k + \tau_k J_2^{k+1} J_1^{k+1} \quad \text{and} \quad u_2^{k+1} = (1 - \tau_k)u^k + \tau_k G_2^{k+1}$$

Algorithm 8 ASA with Inexact Conditional Gradient Method (`ASA+ICG`) for $T = 1$

Replace Step 2 of Algorithm 5 with the following:

2". Update the average gradient z as: $z^{k+1} = (1 - \tau_k)z^k + \tau_k J_1^{k+1}$.

THEOREM 4.2. *Let Assumptions 4.1, 4.2, 4.3 be satisfied by the optimization problem* (4.1). *Let $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$ be some constants depending on the parameters $(\beta, \sigma^2, L, D_{\mathcal{X}}, \delta)$, as defined in* (B.38) *and* (B.46). *Let $\tau_0 = 1, t_0 = 0$, $\tau_k = \frac{1}{\sqrt{N}}, t_k = \lceil \sqrt{k} \rceil, \forall k \geq 1$, where N is the total number of iterations.*

(a) Let $T = 2$, and let $\{x^k, z^k, u_2^k\}_{k\geq 0}$ be the sequence generated by Algorithm 7 with

$$\beta_k \equiv \beta \geq 6\rho L_{\nabla F} + (2\rho + \frac{2}{3\rho})L_{\nabla f_1} L_{f_2}^2, \quad \rho > 0. \tag{4.15}$$

Then, we have $\forall N \geq 1$,

$$\mathbb{E}\left[\|\mathcal{G}_{\mathcal{X}}(x^R, \nabla F(x^R), \beta)\|^2\right] \leq \frac{\mathcal{C}_1}{\sqrt{N}}, \quad \mathbb{E}\left[\|f_2(x^R) - u_2^R\|^2\right] \leq \frac{\mathcal{C}_2}{\sqrt{N}}.$$

(b) Let $T = 1$ and let $\{x^k, z^k\}_{k \geq 0}$ be the sequence generated by Algorithm 8 with $\beta_k \equiv \beta > 0$. Then, we have $\forall N \geq 1$,

$$\mathbb{E}\left[\|\mathcal{G}_{\mathcal{X}}(x^R, \nabla F(x^R), \beta)\|^2\right] \leq \frac{\mathcal{C}_3}{\sqrt{N}}.$$

All expectations are taken with respect to all random sequences generated by the respective algorithms and an independent random integer number R uniformly distributed over $\{1, \ldots, N\}$. In both cases, the number of calls to SFO and LMO to get an ϵ-stationary point is upper bounded by $\mathcal{O}(\epsilon^{-2}), \mathcal{O}(\epsilon^{-3})$ respectively.

Remark. *While we can obtain the same complexities without using the linear approximation of the inner function for $T = 2$, it seems necessary to have a parameter-free algorithm as the choice of β in* (4.15) *depends on the knowledge of the problem parameters. Indeed, the linearization term in* (4.8) *helps use to better exploit the Lipschitz smoothness of the gradients get an error bound in the order of $\tau_k^2\|d^k\|^2$ for estimating the inner function values. Without this term, we are only able to use the Lipschitz continuity of the inner functions and so the error estimate will increase to the order of $\tau_k\|d^k\|$. Hence, we need to choose a larger beta (as in* (4.15)*) to reduce $\|d^k\|$ and handle the error term without compromising the complexities. However, this is not the case for $T = 1$ as it can be seen as a two-level problem whose inner function is exactly known (the identity map). In this case, the choice of β is independent of the problem parameters with or without the linearization term.*

4.3.2. High-Probability Convergence for $T = 1$.

In this subsection, we establish an oracle complexity result with high-probability for the case of $T = 1$. We first provide a notion of (ϵ, δ)-stationary point and a related tail assumption on the stochastic first-order oracle below.

Definition 4.2. *A point $\bar{x} \in \mathcal{X}$ generated by an algorithm for solving* (4.1) *is called an (ϵ, δ)-stationary point, if we have $\|\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta)\|^2 \leq \epsilon$ with probability $1 - \delta$.*

Assumption 4.4. *Let $\Delta^{k+1} = \nabla F(x^k) - J_1^{k+1}$ for $k \geq 0$. For each k, given $\mathscr{F}_k$ we have $\mathbb{E}[\Delta^{k+1} | \mathscr{F}_k] = 0$ and $\|\Delta^{k+1}\| | \mathscr{F}_k$ is K-sub-Gaussian.*

The above assumption is commonly used in the literature; see [**HK14**,**HLPR19**,**LO20**,**ZCC$^+$18**]. We also refer to [**Ver18**] and Appendix B.4 for additional details. The high-probability bound for solving non-convex constrained problems by Algorithm 8 is given below.

THEOREM 4.3. *Let Assumptions 4.1, 4.2, 4.4 be satisfied by the optimization problem* (4.1) *with* $T = 1$. *Let* $\tau_0 = 1, t_0 = 0$, $\tau_k = \frac{1}{\sqrt{N}}, t_k = \lceil \sqrt{k} \rceil, \forall k \geq 1$, *where* N *is the total number of iterations. Let* $T = 1$ *and let* $\{x^k, z^k\}_{k \geq 0}$ *be the sequence generated by Algorithm 8 with* $\beta_k \equiv \beta > 0$. *Then, we have* $\forall N \geq 1, \delta > 0$, *with probability at least* $1 - \delta$,

$$\frac{1}{N} \sum_{k=1}^{N} \left\| \mathcal{G}_{\mathcal{X}}(x^k, \nabla F(x^k), \beta) \right\|^2 \leq \mathcal{O}\left(\frac{K^2 \log(1/\delta)}{\sqrt{N}} \right)$$

Therefore, the number of calls to SFO and LMO to get an (ϵ, δ)*-stationary point is upper bounded by* $\mathcal{O}(\epsilon^{-2} \log^2(1/\delta)), \mathcal{O}(\epsilon^{-3} \log^3(1/\delta))$ *respectively.*

REMARK. *To the best of our knowledge, the above result is (i) the first high-probability bound for one-sample stochastic conditional gradient-type algorithm for the case of* $T = 1$, *and (ii) the first high-probability bound for constrained stochastic optimization algorithms in the non-convex setting; see Appendix J of [**MDB21**].*

4.3.3. Proof Sketch of Main Results. In this section, we only present the proof sketch. The complete proofs are provided in the appendix. For convenience, let $u_{T+1} = x$, and we denote H_k as the function value of the subproblem at step k, y^k as the optimal solution of the subproblem i.e.,

$$H_k(y) := H(y; x^k, z^k, \beta_k), \quad y^k = \operatorname*{arg\,min}_{y \in \mathcal{X}} H_k(y). \tag{4.16}$$

Then, the proof of Theorem 4.1 proceeds via the following steps:

(1) We first leverage the merit function $W_k := W_{\alpha,\gamma}(x^k, z^k, u^k)$ defined in (4.9) with appropriate choices of α, γ for any $\beta > 0$ to obtain

$$\begin{aligned} W_{k+1} - W_k \leq & -\frac{\tau_k}{2} \left(\beta \left[\|d^k\|^2 + \sum_{i=1}^{T} \|f_i(u_{i+1}^k) - u_i^k\|^2 \right] + \frac{\beta}{20 L_{\nabla F}^2} \|\nabla F(x^k) - z^k\|^2 \right) \\ & + \mathbf{R}_k + \tau_k \left(\frac{12}{5} + \frac{20 L_{\nabla F}^2}{\beta^2} \right) \left(H_k(\tilde{y}^k) - H_k(y^k) \right), \quad \forall k \geq 0 \end{aligned}$$

where $\mathbf{R}_k$ is the residual term (see (B.15)) and $\mathbb{E}[\mathbf{R}_k | \mathscr{F}_k] \leq \hat{\sigma}^2 \tau_k^2$, as shown in Proposition B.1.

(2) Telescoping the above inequality, in Lemma B.6 we obtain the following:

$$\begin{aligned}&\sum_{k=1}^{N}\tau_k\left[\beta\left(\|d^k\|^2+\sum_{i=1}^{T}\|f_i(u_{i+1}^k)-u_i^k\|^2\right)+\frac{\beta}{20L_{\nabla F}^2}\|\nabla F(x^k)-z^k\|^2\right]\\ &\le 2W_0+2\sum_{k=0}^{N}\mathbf{R}_k+\left(\frac{24}{5}+\frac{40L_{\nabla F}^2}{\beta^2}\right)\sum_{k=0}^{N}\tau_k\left(H_k(\tilde{y}^k)-H_k(y^k)\right),\quad \forall N\ge 1.\end{aligned}$$

(3) To further control the error term $H_k(\tilde{y}^k)-H_k(y^k)$ introduced by the `ICG` method, we set t_k, the number of iterations in `ICG` method at step k, to $\lceil\sqrt{k}\rceil$. By Lemma B.3, we therefore have

$$H_k(\tilde{y}^k)-H_k(y^k)\le\frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{t_k+2}\le\frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{\sqrt{k}},\quad \forall k\ge 1.$$

Also, with the choice of $\tau_k=\frac{1}{\sqrt{N}}$ and $z^0=0$, we can conclude that

$$\sum_{k=0}^{N}\tau_k\left(H_k(\tilde{y}^k)-H_k(y^k)\right)\le\frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{\sqrt{N}}\sum_{k=1}^{N}\frac{1}{\sqrt{k}}\le 4\beta D_{\mathcal{X}}^2(1+\delta).$$

(4) Then, taking expectation of both sides and by the definition of random integer R, we have

$$\mathbb{E}\left[\beta\left(\|d^R\|^2+\sum_{i=1}^{T}\|f_i(u_{i+1}^R)-u_i^R\|^2\right)+\frac{\beta}{20L_{\nabla F}^2}\|\nabla F(x^R)-z^R\|^2\right]\le 2W_0+\mathcal{B},$$

$\forall N\ge 1$, where $\mathcal{B}$ is a constant depending on the problem parameters $(\beta,\sigma^2,L,D_{\mathcal{X}},T,\delta)$.

(5) As a result, we can obtain (4.13) and (4.14) by noting that $\forall k\ge 1$

$$\begin{aligned}\|\mathcal{G}(x^k,\nabla F(x^k),\beta)\|^2&\le 2\beta^2\|d^k\|^2+2\beta^2\left\|\mathbf{proj}_{\mathcal{X}}\left(x^k-\frac{1}{\beta}\nabla F(x^k)\right)-\mathbf{proj}_{\mathcal{X}}\left(x^k-\frac{1}{\beta}z^k\right)\right\|^2\\ &\le 2\beta^2\|d^k\|^2+2\|\nabla F(x^k)-z^k\|^2.\end{aligned}$$

where the second inequality follows the non-expansiveness of the projection operator.

The proofs of Theorems 4.2 and 4.3 follow the same argument with appropriate modifications. The high-probability convergence proof of Theorem 4.3 mainly consists of controlling the tail probability of the residual term $\mathbf{R}_k$ being large.

4.4. Numerical Experiments for $T=1$

To demonstrate the effectiveness and efficiency of proposed algorithms compared to 1-SFW **[ZSM+20]** for $T=1$, we consider the following matrix-valued single-index model **[YBL17]** with

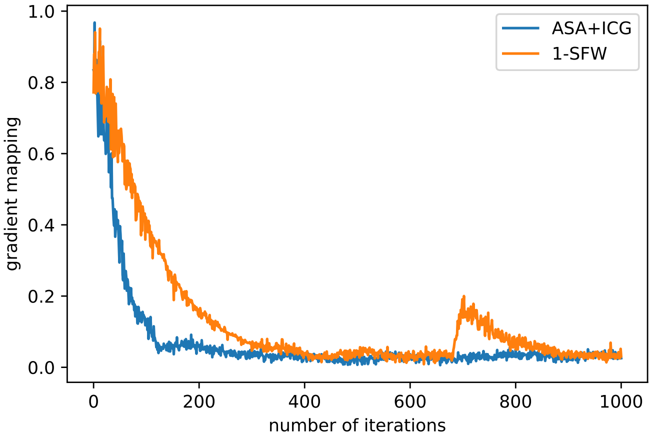


FIGURE 4.1. `ASA+ICG` vs. `1-SFW`

low-rank constraints:

$$y = |\langle A, B^\star\rangle_F|^2 + \epsilon, \quad \operatorname{rank}(B^\star) \le s,$$

where $A, B \in \mathbb{R}^{m\times n}$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius inner product, and s is some positive integer strictly less than m and n. To recover a low-rank matrix B, one can optimize the mean squared loss with nuclear norm constraint, in which the Frank-Wolfe update is much cheaper than the projection operator especially with large-scale matrices [**Jag13**]. Formally, our problem can be written as

$$\min\ F(B) = \mathbb{E}_{A,\epsilon}\left[(y - |\langle A, B\rangle_F|^2)^2\right] \quad \text{s.t. } \|B\|_\star \le s.$$

We evaluate the performance of `ASA+ICG` (Algorithm 8) and `1-SFW` on a toy example where $B^\star = vv^\top/\|vv^\top\|_\star$ is a 4 by 4 rank-1 matrix. The matrix A is generated as $A = I + E$ where $E_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.3)$. The stepsize parameter $\beta = 1$ for `ASA+ICG`, and all the parameters in `1-SFW` is set according to Theorem 2 in [**ZSM$^+$20**]. As the exact gradient of F is unavailable, we estimate the gradient mapping by using averaged stochastic gradients. In Figure 4.1, we plot the value of gradient mapping versus the number of iterations, which demonstrates the superior of our proposed method for $T = 1$ in the one-sample setting.

4.5. Discussion and Conclusion

In this work, we propose and analyze projection-free conditional gradient-type algorithms for constrained stochastic multi-level composition optimization of the form in (4.1). We show that the oracle complexity of the proposed algorithms is level-independent in terms of the target accuracy. Furthermore, our algorithm does not require any increasing order of mini-batches under standard unbiasedness and bounded second-moment assumptions on the stochastic first-order oracle, and is parameter-free. Some open questions for future research: (i) Considering the one-sample setting, either improving the LMO complexity from $\mathcal{O}(\epsilon^{-3})$ to $\mathcal{O}(\epsilon^{-2})$ for general closed convex constraint sets or establishing lower bounds showing that $\mathcal{O}(\epsilon^{-3})$ is necessary while keeping the SFO in the order of $\mathcal{O}(\epsilon^{-2})$, is extremely interesting; and (ii) Providing high-probability bounds for stochastic multi-level composition problems ($T > 1$) and under sub-Gaussian or heavy-tail assumptions (as in [**MDB21**, **LZW22**]) is interesting to explore.

CHAPTER 5

Stochastic Conditional Gradient Methods under Interpolation-like Conditions

5.1. Introduction

Consider the following constrained stochastic optimization problem:

$$\min_{x \in \Omega} \quad \{ f(x) := \mathbb{E}_{\xi} [F(x, \xi)] \} , \tag{5.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ and $\Omega \subset \mathbb{R}^d$ is a closed and convex set and ξ is a random vector characterizing the stochasticity in the problem. In a machine learning setup, the function F could be interpreted as the loss function associated with a sample ξ and the function f could represent the risk, which is defined as the expected loss. Such constrained stochastic optimization problems arise frequently in statistical machine learning applications. The conditional gradient algorithm, also called as the Frank-Wolfe algorithm, is an efficient method for solving constrained optimization problems of the form in (5.1) due to their projection-free nature [**Jag13**, **HJN15**, **FGM17**, **LPZZ17**, **BZK18**, **RDLS18**]. In each step of the conditional gradient method, it is only required to minimize a linear objective over the set Ω. This operation could be implemented efficiently for a variety of sets arising in statistical machine learning, compared to the operation of projecting onto the set Ω, which is required for example by the projected gradient method. Hence, the conditional gradient method has regained popularity in the last decade in the optimization and machine learning community.

There has been extensive work in the past decade on analyzing the stochastic conditional gradient algorithm for optimization problems of the form in (5.1); see for example [**GH13**, **HL16**, **LZ16**, **RSPS16**, **Gha19**]. However, existing works do not take into account certain favorable structures that are naturally available in modern over-parametrized machine learning problems. Specifically, it has been noted that modern machine learning models predict well on unseen data, despite fitting the training data perfectly [**ZBH$^+$16**, **HLVDMW17**, **LR18**, **MBB18**, **MRSY19**, **HMRT19**]. Examples include logistic regression or support vector machine with squared-hinge loss that are trained with linearly separable data [**VBS19**, **VML+19**, **MVL+20**] and deep neural networks [**VBS19**, **BBM18**]. From an optimization point of view, for the problem in (5.1) with $\Omega \equiv \mathbb{R}^d$, the above interpolation condition means that at the optimal point, the gradient is not only zero (or close to zero) with respect to the risk function f but is also almost surely equal to zero for the random loss function F. Such a scenario helps to reduce the stochasticity in the gradient estimation process which in turn results in improved complexity results for several stochastic optimization procedures. Indeed in the recent past, several works have provided improved rates for algorithms like stochastic gradient descent [**NWS14**, **MBB18**, **BBM18**, **GLQ+19**, **VBS19**, **VML+19**] and sub-sampled Newton's method [**MVL+20**]. In particular, for several settings, the above works demonstrate that the stochastic algorithm may perform as well as the corresponding deterministic counterpart. However, such works only study unconstrained optimization problems and do not have any consequences for constrained stochastic optimization problems of the form in (5.1).

Hence, in this work we consider the following question: *Can we obtain improvements in the oracle complexity of algorithms used for projection-free constrained stochastic optimization problems arising in the context of over-parametrized machine learning models, that are capable of perfectly interpolating the training data?* We give a positive answer to the above question by demonstrating that the stochastic conditional gradient method, a projection-free technique for solving constrained stochastic optimization problems, also enjoy improved oracle complexities when they are used to solve constrained stochastic optimization problems of the form in (5.1) under certain *interpolation-like* conditions. We elaborate on the specific form of improvement observed below. For stochastic conditional gradient algorithms, the oracle complexity is measured in terms of number of calls to the Stochastic First-order Oracle (SFO) and the Linear Minimization Oracle (LMO) used to the solve the subproblems (that are of the form of minimizing a linear function over the convex feasible set) arising in the algorithm. In this work, we make the following contribution to the literature on conditional gradient methods under interpolation-like assumptions (see Section 5.2 for the exact definitions) on the stochastic gradient:

(1) For the case of convex f in (5.1), we show that the number of calls to the SFO for the *vanilla* stochastic conditional gradient method and stochastic conditional gradient sliding methods are given respectively by $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-1.5})$. For comparison, without such assumptions, the

corresponding complexities are $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-2})$ respectively. The number of calls to the linear minimization oracle (LMO) is of the order $\mathcal{O}(\epsilon^{-1})$, in both cases.

(2) We also demonstrate similar improvements in the context of zeroth-order conditional gradient methods, where one only observes noisy evaluations of the function being optimized. Specifically, the number of calls to the stochastic zeroth-order oracle for the *vanilla* stochastic conditional gradient method and stochastic conditional gradient sliding methods are given respectively by $\mathcal{O}(d\epsilon^{-2})$ and $\mathcal{O}(d\epsilon^{-1.5})$, with the same LMO complexity as the first-order setting.

We emphasize that, notably the above improvements are achieved without incorporating any double-loop based existing variance reduction techniques, for example SVRF [**RSPS16**] or SPIDER-FW [**YSC19**]. It is also worth noting that [**DB19**, **Sch20**] argue that variance reduction techniques (at the least existing ones) are ineffective in the context of modern deep learning models which are invariably over-parametrized. We also remark that, in contrast to stochastic gradient methods for unconstrained optimization [**VBS19**, **BBM18**], the above improved results still do not match the corresponding deterministic rates highlighting the subtlety with projection-free optimization.

5.1.1. Related Work.

The conditional gradient method or the Frank-Wolfe method was proposed first by [**FW56**]. It has obtained renewed interest in the machine learning and optimization communities due to their projection-free nature. We refer the reader to [**Jag13**, **HJN15**, **LJJ15**, **BS17**, **GSK18**], for a partial list of recent works predominantly in the deterministic setting. For the stochastic setting that we consider, in each step of the conditional gradient method, the algorithm requires access to a stochastic first-order oracle (SFO) and a linear minimization oracle (LMO). The complexity of conditional gradient method is hence measured by the number of calls to both oracles.

Convex Setting. Considering the convex setup, [**HL16**] showed that to obtain an ϵ-optimal point, the number of calls to SFO and LMO are given respectively by $\mathcal{O}(1/\epsilon^3)$ and $\mathcal{O}(1/\epsilon^1)$. Furthermore, [**LZ16**] proposed Conditional Gradient Sliding (CGS), a modified Frank-Wolfe method by Nesterov's acceleration, that improves the number of calls to the SFO to $\mathcal{O}(1/\epsilon^3)$ while keeping the number of calls to LMO the same. The above methods require an increasing batch size, in each step, to obtain the above mentioned complexities. Recently, [**MHK18b**, **HKMS19**, **ZSM$^+$19**] addressed the issue of increasing batch size. But these works require potentially restrictive assumptions and in particular [**MHK18b**] and [**ZSM$^+$19**] also require an increased number of calls to the LMO. We highlight that [**YSC19**] used the SPIDER technique, and showed that one could improve the

	Extra Conditions	First-Order	
		SFO	LMO
Vanilla-SFW	–	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-1})$
SPIDER-SFW	Mean-square gradient-smoothness	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$
Vanilla-SFW	Interpolation-like	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$
Vanilla-SCGS	–	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$
SPIDER-SCGS	Mean-square gradient-smoothness	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$
Vanilla-SCGS	Interpolation-like	$\mathcal{O}(\epsilon^{-1.5})$	$\mathcal{O}(\epsilon^{-1})$

TABLE 5.1. Comparison of the oracle complexities of conditional gradient methods under various assumptions in the first-order setting. All methods require the gradient-smoothness condition: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|$. Mean-square gradient-smoothness refers to $\mathbb{E}_\xi\|\nabla F(x,\xi) - \nabla F(y,\xi)\|_2^2 \leq L\|x - y\|_2^2$ and is stronger than gradient-smoothness condition. SFO refers to number of calls to the stochastic first order oracle. LMO refers to the number of calls to the linear minimization oracle. The results for vanilla-SFW are from [**HL16**] and [**RSPS16**]. The results for vanilla-SCGS are from [**LZ16**]. The results for SPIDER-SFW and SPIDER-SCGS are from [**YSC19**]. The results highlighted in blue are our results. Finally, in this work, we also obtain complexities under the zeroth-order setting which are not highlighted in this table, for simplicity.

SFO complexity from $\mathcal{O}(1/\epsilon^3)$ to $\mathcal{O}(1/\epsilon^2)$ while mainting the number of calls to the LMO at $\mathcal{O}(1/\epsilon)$. However, to obtain this improved complexities, the SPIDER technique requires double-loop based variance reduction techniques, and hence they are harder to implement in practice compared to the vanilla methods. Furthermore, the SPIDER technique requires the stronger mean-square Lipschitz gradient assumption.

Comparison to 1-sample SFW from [ZSM$^+$19]. While above discussed methods require increasing batch-size with the number of iterations, we highlight that recently [**ZSM$^+$19**] proposed 1-sample SFW which does not require increasing batch size. The results in [**ZSM$^+$19**] for the 1-sample SFW method has an SFO complexity of $\mathcal{O}(1/\epsilon^2)$ for the convex setting, the same complexity we present in this work. However, the LMO complexity of 1-sample SFW is $\mathcal{O}(1/\epsilon^2)$, and they require additional smoothness assumption on F. We emphasize that increasing batch size setting is commonly used in the literature of conditional gradient methods and obtaining methods without this requirement under milder assumptions are interesting future work. In this work, we leverage the standard setting of SFW with increasing batch sizes, and focus on the improved oracle complexities of vanilla SFW methods under certain favorable structures.

Interpolation. In the interpolation regime, [**MBB18**] showed that mini-batch stochastic gradient descent (SGD) algorithm enjoys exponential rates of convergence for unconstrained strongly-convex optimization problems; see also [**SV09**,**NWS14**] for related earlier work. For the non-convex setting, [**BBM18**] analyze SGD for non-convex functions satisfying the Polyak-Lojasiewicz (PL) inequality ([**Pol63**]) under the interpolation condition and show that SGD can achieve a linear convergence rate. Later, [**VBS19**] introduced a more practical form of interpolation condition, and prove that the constant step-size SGD can obtain the optimal convergence rate for strongly-convex and smooth convex functions. They also show the first results in the non-convex setting that the constant step-size SGD can obtain the deterministic rate $\mathcal{O}(1/t)$ in the interpolation regime. Subsequently, [**MVL$^+$20**] investigate the regularized subsampled Newton method (R-SSN) and the stochastic BFGS algorithm under the interpolation-like conditions. Very recently, [**RBGM20**] showed that for non-convex problems, one could escape saddle-points and converge to local-minimizers faster under SGC condition. We emphasize that all the above works consider only unconstrained stochastic optimization problems, while we consider the more challenging constrained stochastic optimization problems.

5.2. Preliminaries and Assumptions

We now list and discuss the set of assumptions made in our work. We first list some regularity assumptions on the function f and the set Ω.

Assumption 5.1. *The function f has L-Lipschitz gradient ∇f, i.e., for any pair of points $x, y \in \Omega$, we have $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$, and the feasible set $\Omega \subset \mathbb{R}^d$ is bounded, i.e., $\max\limits_{x,y\in\Omega} \|x - y\| \leq D$.*

The above set of assumptions are standard in the analysis of stochastic conditional gradient methods and has been used in prior works in the literature; see for example [**GL13**]. We make the above assumptions for both the first-order setting. We also require the following smoothness assumption in the zeroth-order setting.

Assumption 5.2. *The function F has Lipschitz continuous gradient with constant L, almost surely for any ξ, i.e., for any $x, y \in \mathbb{R}^d$, i.e., almost surely we have $\|\nabla F(x,\xi) - \nabla F(y,\xi)\| \leq L\|x-y\|$.*

Note that the above assumption is stronger than the first statement of Assumption 5.1 and implies it. However, we only use Assumption 5.2 for the analysis of zeroth-order algorithms.

5.2.1. Growth Conditions in the Convex Constrained Setting. We now state the main interpolation-like assumptions that we make in our work when f is convex and provide the main intuition behind such an assumption.

Assumption 5.3 (Moment-based Weak Growth Condition). *Let x^* be the minimum point of f. We say that f satisfies the Moment-based Weak Growth Condition (WGC) with constant ρ, if for any point $x \in \Omega$, we have*

$$\mathbb{E}_\xi \|\nabla F(x,\xi)\|^2 \leq 2\rho L \left[f(x) - f(x^*)\right]. \tag{5.2}$$

Assumption 5.4 (Variance-based Weak Growth Condition). *Let x^* be the minimum point of f. We say that the function f satisfies the Variance-based Weak Growth Condition (WGC) with constant ρ, if for any point $x \in \Omega$, we have*

$$\mathbb{E}_\xi \|\nabla F(x,\xi) - \nabla f(x)\|^2 \leq 2\rho L \left[f(x) - f(x^*)\right]. \tag{5.3}$$

The above conditions are motivated by the so-called strong growth condition: $\mathbb{E}\|\nabla F(x,\xi)\|^2 \leq \rho\|\nabla f(x)\|^2$, used in **[VBS19]** for obtaining faster rates of convergence for stochastic gradient method in the unconstrained setting. Notice that in the interpolation setting, when $\nabla f(x^*) = 0$, we have $\nabla F(x^*,\xi) = 0$, almost surely. Thus, the strong growth condition is defined exactly to take advantage of this situation. Furthermore, in the smooth convex setting, **[VBS19]** showed that the strong-growth condition is equivalent to the moment-based weak growth condition in Assumption 5.3. However, the moment-based weak growth condition as proposed in **[VBS19]** is not directly suited for the constrained stochastic setting that we consider in this work. It is easy to construct examples for which there exists stationary point at the boundary of Ω with non-zero (stochastic) gradient, i.e., $\mathbb{E}\|\nabla F(x,\xi)\|^2$ could remain positive while the right hand side goes to 0 and hence the assumption is not satisfied. In order to resolve this issue, for the constrained setting, we relax the moment-based growth conditions to the variance-based versions. Note that we have

$$\mathbb{E}\|\nabla F(x,\xi)-\nabla f(x)\|^2=\mathbb{E}\|\nabla F(x,\xi)\|^2-\|\nabla f(x)\|^2\leq \mathbb{E}\|\nabla F(x,\xi)\|^2.$$

Thus variance-based growth conditions naturally become the substitute for the moment-based version in constrained problems and could hold even the moment-based conditions do not hold. As they are also motivated by the interpolation assumption, we refer to these conditions as interpolation-like conditions. Formally, under the variance-based growth conditions for a convex f, if we attain an optimal point $x^* \in \Omega$, the variance of the stochastic first-order oracle will be almost surely zero, i.e., $\nabla F(x^*,\xi)=\nabla f(x^*)$ almost surely. This property eventually leads to the improvements in the query complexity that we demonstrate. We emphasize that it is natural to construct counter-examples that violate Assumption 5.4. In those cases, the improved query complexities that we demonstrate are simply not applicable. Finally, we also have the following natural relationships between the two conditions.

Proposition 5.1. *The Weak Growth Conditions defined above have the following relations:*

(a) If f satisfies the Moment-based WGC (5.3) *with ρ, then f satisfies the Variance-based WGC* (5.4) *with ρ and there exists $x^* \in \Omega$ such that $\nabla f(x^*)=0$.*

(b) If f satisfies the Variance-based WGC (5.4) *with ρ and there exists $x^* \in \Omega$ such that $\nabla f(x^*)=0$, then f satisfies the Moment-based WGC* (5.3) *with $\rho+1$.*

5.2.2. Growth Conditions in the Zeroth-Order Constrained Setting.

In the zeroth-order setting, we only assume availability of the noisy function evaluations. This oracle setting is motivated by several applications where only noisy function queries of problem (5.1) is available, such as reinforcement learning [**SHC⁺17**, **CRS⁺18a**, **CRS⁺18b**], hyperparameter tuning [**SLA12**], and black-box attacks to deep networks [**CZS⁺17**, **SZK19**]. Hence, we use the Gaussian Stein's identity based random gradient estimator, a standard gradient estimator in the zeroth-order optimization literature [**GL13**, **DJWW15**, **NS17**, **BG21**]:

$$\bar{G}_\nu(x)=\frac{1}{b}\sum_{j=1}^{b}\frac{F(x+\nu u_j,\xi_j)-F(x,\xi_j)}{\nu}u_j,$$

where $u_1, \ldots, u_b$ are i.i.d. samples from $\mathcal{N}(0, \mathbf{I}_d)$. The above gradient estimator is a biased estimator of the true gradient $\nabla f(x)$, and was also used in [**BG21**], to develop zeroth-order conditional gradient descent algorithms.

While for the first-order setting, we use the relatively weaker variance-based conditions to obtain the improved bounds, in the zeroth-order setting, it turns out the stronger moment-based conditions are required. The reason is that the mean square error of the biased zeroth-order gradient estimator is bounded above by $\mathbb{E}\|\nabla F(x, \xi)\|^2$. Hence, to obtain improved rates, it makes it necessary to make assumptions on the moments of the stochastic gradient directly. We emphasize that this is required only for the constrained problems, since the moment-based conditions are equivalent to the variance-based conditions when there exists one zero-gradient point in the constraint set (see Proposition 5.1). In particular, we show in Appendix C.3 that a zeroth-order version of Theorem 3 from [**VBS19**], for stochastic gradient descent, to bound the gradient size in the nonconvex setting could be proved just under the variance-based growth conditions.

5.2.3. Motivating Examples. Before we present our main results in the next section, we briefly discuss some motivating examples of constrained stochastic optimization problems that arise in modern machine learning. In the convex setting, it is easy to see that kernel regression [**LR18**], squared-Hinge loss based linear SVM classifier or logistic regression on linearly separable data could be considered as operating in the over-parametrized regime and hence satisfy interpolation-like conditions [**VBS19**, **MVL$^+$20**].

However, without any constraints, such predictors might be biased against certain sensitive features like race or gender. One way to build fair predictors is to explicitly encode fairness constraints with respect to certain pre-defined sensitive features [**DOBD$^+$18**, **ABD$^+$18**]. Specifically, it was shown in [**ABD$^+$18**] that several standard and well-accepted notions of fairness in classification setting, including equalized odds [**HPS16**], demographic parity [**DIKL18**], balance for the negative class [**KMR16**], treatment equality [**BHJ$^+$18**] could be formulated as empirical risk minimization problems subjected linear inequality constraints. In this case, the problem is exactly of the form in (5.1) with Ω being a polytope. Furthermore, [**DOBD$^+$18**] also proposed a general approach for fair empirical risk minimization. Similar to [**ABD$^+$18**], the fundamental idea is to enforce constraints such that the conditional risk of a predictor is not varying much with respect to the sensitive features associated with the problem. Such formulations of fair empirical risk minimization in the interpolation regime also fall under the class of problems in (5.1).

Squared hinge loss with linearly separable data. As a concrete example, we extend the unconstrained examples presented in [**VBS19**] to the constrained setting we consider. Assuming a finite support of features and the linearly separable data, it has been shown that the squared-hinge loss satisfies SGC with $\rho = c/\tau^2$ where c is the cardinality of the support and τ is the margin (Lemma 1 in [**VBS19**]). In the above regime, the optimal classifier that minimizes the loss and achieves a stationary point with zero gradient is not always unique. In practice, to construct a fair classifier, enforcing constraints is a natural approach. Note that if there exists an $x^* \in \Omega$, by the convexity and the L-smoothness of f, we have

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - f(x^*)). \tag{5.4}$$

That is to say, for linearly separable data with margin τ and a finite support of size c, if there exists one $x^* \in \Omega$, the squared-hinge loss satisfies Assumption 5.3 with $\rho = c/\tau^2$.

5.3. Improved Complexities for Stochastic Conditional Gradient Methods

We now provide improved complexities for stochastic conditional gradient methods under the interpolation-like assumption in Section 5.2. For convenience, we first introduce the following mini-batch stochastic gradients with first-order and zeroth-order oracle access: at t-th iteration, we uniformly pick i.i.d. samples $\{\xi_{t,1}, \dots, \xi_{t,b_t}\}$ and estimate the gradient by

$$\tilde{\nabla}_t := \frac{1}{b_t}\sum_{i=1}^{b_t} \nabla F(x_{t-1}, \xi_{t,i}), \quad \bar{G}_\nu^t := \frac{1}{b_t}\sum_{j=1}^{b_t} \frac{F(x_{t-1} + \nu u_{t,j}, \xi_{t,j}) - F(x_{t-1}, \xi_{t,j})}{\nu} u_{t,j}$$

where $u_{t,1}, \dots, u_{t.b_t}$ are i.i.d. samples from $\mathcal{N}(0, \mathbf{I}_d)$.

5.3.1. Stochastic Frank-Wolfe. In this section, we studied the oracle complexity of the vanilla stochastic Frank-Wolfe algorithm under the weak interpolation-like conditions in Assumption 5.4 and 5.3.

Theorem 5.1. *Consider solving problem* (5.1)*, by Algorithm 9, under Assumption 5.1 with f being convex.*

Algorithm 9 Stochastic Frank-Wolfe

```
Input: $x_0 \in \Omega$, number of iterations $T$, $\gamma_t \in [0, 1]$, minibatch size $b_t$
for $t = 1, 2, \ldots, T$ do
    Compute the gradient $g_t$ as follows:
            Set $g_t = \tilde{\nabla}_t$ (for the first-order setting).
            Set $g_t = \bar{G}_\nu^t$ (for the zeroth-order setting).
    Compute $d_t = \arg\min_{d \in \Omega} \langle d, g_t \rangle$
    $x_t = x_{t-1} + \gamma_t (d_t - x_{t-1})$
end for
Output: $x_T$
```

(a) Assuming access to stochastic first-order oracle, under Assumption 5.4, setting

$$\gamma_t = \frac{4}{t+3}, \quad b_t = \lceil (t+3)/2 \rceil,$$

we have the following convergence rate:

$$\mathbb{E}[f(x_t) - f(x^*)] \le \frac{2(f(x_0) - f(x^*)) + 8(\rho + 1)LD^2}{t+3}.$$

Hence, the total number of calls to the stochastic first-order oracle and linear minimization oracle required to be solved to find an ϵ-optimal point of problem (5.1) *are, respectively, bounded by*

$$\mathcal{O}\left(\epsilon^{-2}\right), \quad \mathcal{O}\left(\epsilon^{-1}\right).$$

(b) Assuming access to stochastic zeroth-order oracle, under Assumptions 5.3 and 5.2, setting

$$\gamma_t = \frac{4}{t+3}, \quad b_t = (t+3)(d+4), \quad \nu = \frac{D}{(T+3)(d+6)^{3/2}}$$

we have

$$\mathbb{E}[f(x_t) - f(x^*)] \le \frac{2(f(x_0) - f(x^*)) + 8(\rho + \rho^{-1} + 1)LD^2}{t+3}.$$

Hence, the total number of calls to the stochastic zeroth-order oracle and linear minimization oracle required to be solved to find an ϵ-optimal point of problem (5.1) *are, respectively, bounded by*

$$\mathcal{O}\left(d\epsilon^{-2}\right), \quad \mathcal{O}\left(\epsilon^{-1}\right).$$

The above oracle complexities in the first-order setting, match the results obtained by [**YSC19**, **ZSM$^+$19**]. However, the above works require double-loop based variance reduction techniques which in turn require the stronger mean-square gradient-Lipschitz assumption. Furthermore, the use of the variance reduction technique results in the increased wall-clock running time of the algorithm. Our result here is applicable to the vanilla version of the stochastic conditional gradient method, as long as the problem satisfies the interpolation-like conditions observed in modern machine learning problems.

5.3.2. Stochastic Conditional Gradient Sliding. In this section, we analyze the complexity of the stochastic gradient sliding (SCGS) algorithm under the weak growth condition. The SCGS was first proposed and thoroughly analyzed in [**LZ16**]. It is a fundamental modification of the conditional gradient algorithm that achieved improved oracle complexities without relying on any variance reduction techniques. Below, we show that under the interpolation-like assumptions in Section 5.2, the oracle complexity of the SCGS could be further improved compared in both the first-order and zeroth-order methods.

Algorithm 10 Stochastic Conditional Gradient Sliding (SCGS)

```
Input: $x_0 \in \Omega$, $T$, $\beta_t \in \mathbb{R}_+$, $\gamma_t \in [0,1]$, $b_t$, $y_0 = x_0$
for $t = 1, 2, \dots, T$ do
    Set $z_t = (1-\gamma_t)x_{t-1} + \gamma_t y_{t-1}$
    Compute the gradient $g_t$ as follows:
            Set $g_t = \tilde{\nabla}_t$ (first-order).
            Set $g_t = \bar{G}_\nu^t$ (zeroth-order).
    Solve
            $y_t = \mathrm{ICG}(g_t, y_{t-1}, \beta_t, \eta_t)$
    by Algorithm 11
    Set $x_t = (1-\gamma_t)x_{t-1} + \gamma_t y_t$
end for
Output: $x_T$
```

Algorithm 11 Inexact Conditional Gradient Method (ICG)

```
Input: $g, u, \beta, \eta, u_1 = u, k = 1$
  1. Let $v_k$ be an optimal solution for the subproblem
     (5.5)  $\max_{v \in \Omega} \{h_k(v) = \langle g + \beta(u_k - u), u_k - v \rangle\}.$
  2. If $h_k(v_k) \leq \eta$, terminate and output $u_k$.
  3. $u_{k+1} = (1 - \alpha_k) u_k + \alpha_k v_k$ with
     $\alpha_k = \min \left\{ 1, \frac{\langle \beta(u - u_k) - g, v_t - u_t \rangle}{\beta \left\| v_k - u_k \right\|^2} \right\}.$
  4. Set $k \leftarrow k + 1$ and go to step 1.
```

Theorem 5.2. *Consider solving problem* (5.1)*, by Algorithm 10, under Assumption 5.1 with f being convex.*

(a) Assuming access to stochastic first-order oracle, under Assumption 5.4, setting

$$\beta_t = \frac{4L}{t+2}, \quad \gamma_t = \frac{3}{t+2}, \quad \eta_t = \frac{LD^2}{t(t+1)}, \quad b_t = \lceil 3\rho t(t+1) \rceil$$

we have

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{6LD^2}{(t+2)^2} + \frac{15LD^2 + 3\|\nabla f(x^*)\| D}{(t+1)(t+2)}.$$

Hence, the total number of calls to the stochastic first-order oracle and linear minimization oracle required to be solved to find an ϵ-optimal point of problem (5.1) *are, respectively, bounded by*

$$\mathcal{O}\left(\epsilon^{-1.5}\right), \quad \mathcal{O}\left(\epsilon^{-1}\right).$$

(b) Assuming access to stochastic zeroth-order oracle, in addition, with Assumption 5.3, 5.2, setting

$$\beta_t = \frac{4L}{t+2}, \quad \gamma_t = \frac{3}{t+2}, \quad \eta_t = \frac{LD^2}{t(t+1)}, \quad b_t = \lceil 6\rho(d+4)t(t+1) \rceil, \quad \nu = \frac{D}{(T+2)^2(d+6)^{3/2}},$$

we have

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{8LD^2}{(t+2)^2} + \frac{32LD^2}{(t+1)(t+2)}.$$

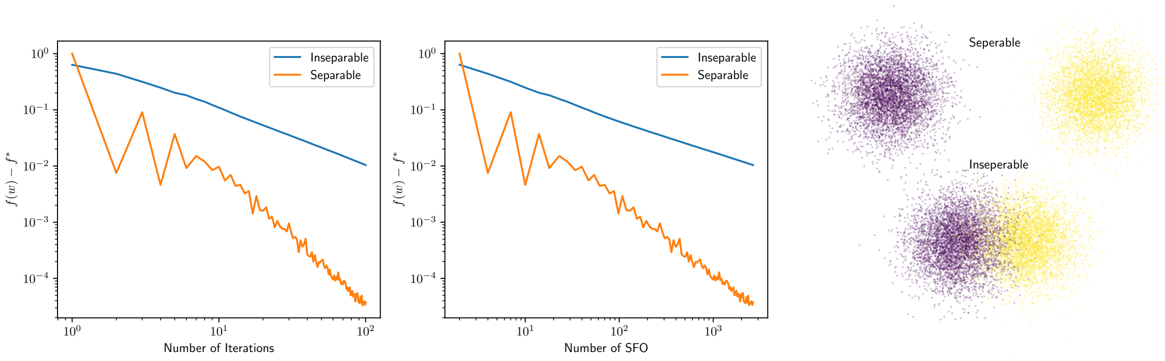


FIGURE 5.1. The convergence behaviors of SFW for linearly (in)-separable data. The right panel visualizes the first 2 dimensions of the synthetic data used for numerical analyses.

Hence, the total number of calls to the stochastic zeroth-order oracle and linear minimization oracle required to be solved to find an ϵ-optimal point of problem (5.1) *are, respectively, bounded by*

$$\mathcal{O}\left(d\epsilon^{-1.5}\right), \quad \mathcal{O}\left(\epsilon^{-1}\right).$$

To the best of our knowledge, the above complexity of $\mathcal{O}(\epsilon^{-1.5})$ is not achieved for any variance reduced versions of stochastic Frank-Wolfe methods. This improvement is solely obtained by the SCGS algorithm of [**LZ16**] under the interpolation-like assumptions which are natural in modern machine learning problems, without any variance reduction methods. We also highlight that, in the unconstrained setting, the stochastic gradient method performs as well as its deterministic counterpart. However, the above result still falls short of the corresponding deterministic complexity of conditional gradient sliding, which is of the order $\mathcal{O}(\epsilon^{-0.5})$ [**LZ16**]. This highlights the intrinsic difficulty associated with projection-free methods for constrained stochastic optimization problems.

5.4. Experiments

We generate synthetic binary classification datasets with two isotropic Gaussian blobs symmetric with respect to the origin, with the sample size $n = 100,000$ and the dimension $d = 500$. We ensure that two blobs are linearly separable with a positive margin for one dataset while the other has an overlap. We seek to find a hyperplane $w^\top x$ that minimizes the squared-hinge loss $f(w) = \frac{1}{n}\sum_{i=1}^n f_i(w) = \frac{1}{n}\sum_{i=1}^n \max(0, 1 - y_i \cdot w^\top x_i)^2$ satisfying the constraint $\|w\|_1 \leq 1$. Note that $f(w)$ satisfies the weak growth condition for linearly separable data in view of sampling only a mini-batch of gradient (with replacement) in each iteration, and the parameter $\rho = L_{\max}/L$; see Proposition 2 in [**VBS19**], and $L_{\max}$ is the largest Lipschitz constant for $\nabla f_i(w)$. In Figure 5.1, we plot the suboptimality $f(w) - f^*$ versus the number of iterations and the number of calls to the SFO. The results are obtained by averaging over 100 runs with random initialization w_0. We observe that SFW converges essentially faster for linearly separable data than the inseparable case.

5.5. Discussion and Conclusion

We briefly discuss extensions of our results to the nonconvex setting. Our proposed assumption is motivated by the notion of Frank-Wolfe gap [**DR70**, **Hea82**], which is defined as $\mathcal{G}_f(x) = \max_{y\in\Omega}\langle \nabla f(x), x - y\rangle$. With this, a nonconvex function f satisfies Constrained Growth Condition with constant ρ, if for any point $x \in \Omega$, $\mathbb{E}_\xi \left\|\nabla F(x,\xi) - \nabla f(x)\right\|^2 \leq 2\rho L \mathcal{G}_f(x)$. Note that if f is convex, then $\mathcal{G}_f(x) \geq f(x) - f(x^*)$. Hence, this generalizes Assumption 5.4 defined for the convex setting. Under this assumption in the nonconvex setting, it could be shown that the vanilla stochastic Frank-Wolfe algorithm can find an ϵ-stationary point of the problem within at most $\mathcal{O}\left(1/\epsilon^3\right)$ and $\mathcal{O}\left(1/\epsilon^2\right)$ number of calls to the SFO linear subproblem solver, respectively. However, although existence of functions satisfying the above asssumption could be shown, it is not clear if practical nonconvex functions appearing in machine learning context satisfy it. It would be extremely interesting to examine this as future work.

In a nutshell, considering convex constrained stochastic optimization problems, we show improved complexity bounds for the vanilla stochastic conditional gradient method under certain interpolation-like conditions that occur naturally in over-parametrized models that are common in machine learning. Our results do not require any double-loop based variance reduction techniques and is hence easily implementable. Furthermore, apart from the batch-size parameter for stochastic conditional gradient sliding method (Algorithm 10), the tuning parameters of the algorithm are independent of the parameter ρ characterizing the interpolation-like conditions.

APPENDIX A

Appendix of Chapter 3

A.1. Experimental Details

All experiments are conducted on a laptop with Intel Core i7-11370H Processor and Windows 11 operating system. The total iteration numbers for a9a and MNIST are 10000 and 3000 respectively. The graph that represents the network topology is set to be ring (or cycle in graph theory) for a9a and random graph (given by [**MBMXC22**]) for MNIST (See Figure A.1). To demonstrate the performance of our algorithms in a constant batch size setting, the batch sizes are chosen to be 4 for a9a and 32 for MNIST in all algorithms. We adjust the learning rates provided in the code of [**MBMXC22**] accordingly and select the ones that have the best performance. For `Prox-DASA` and `Prox-DASA-GT` we choose a diminishing stepsize sequence, namely, $\alpha_k = \min\left\{\alpha\sqrt{\frac{n}{k}}, 1\right\}$ for all $k \geq 0$. Note that the same complexity (up to logarithmic factors) bounds can be obtained by directly plugging in the aforementioned expressions for α_k in Section 4.3. Then we tune $\gamma \in \{1, 3, 10\}$ and $\alpha \in \{0.3, 1.0, 3.0\}$. The penalty parameter λ is chosen to be 0.0001 for all experiments.

We summarize the outputs of all experiments in Table A.1, from which we can tell `Prox-DASA` and `Prox-DASA-GT` achieve good performance in a relatively short amount of time. The stationarity is defined as $\|\mathcal{G}(\bar{x}^k, \nabla F(\bar{x}^k), 1)\|^2 + \|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2$, which is the same as that in [**MBMXC22**]. As mentioned in the caption of Figure 2 in the main paper, there is an extra hyperparameter q in `ProxGT-SR-E`, and we found that large q already works well for a9a experiment, but q has to be small in the MNIST experiment otherwise the final accuracy will be much smaller than the one presented in Table A.1. Hence in `ProxGT-SR-E` we choose $q = 1000$ for a9a and $q = 32$ for MNIST, and the plots that take this amount of epochs into account are in Figure A.2.

A.2. Proof of Theorem 3.1

We present the complete proof in this section. In the sequel, $\|\cdot\|$ denotes the ℓ_2-norm for vectors and Frobenius norm for matrices. $\|\cdot\|_2$ denotes the spectral norm for matrices. $\mathbf{1}$ represents

Table A.1. Comparisons between all algorithms

Algorithm	Accuracy	Training Loss	Stationarity	Communication time per iteration (s)	Computation time per iteration (s)	Total time per iteration (s)
				a9a		
`SPPDM`	84.64%	0.3340	0.0174	0.0260	0.0305	0.0565
`ProxGT-SR-E`	76.38%	0.6528	0.0797	0.0521	0.0394	0.0915
`DEEPSTORM v2`	**84.90%**	**0.3274**	0.0029	0.0525	0.0398	0.0923
`Prox-DASA`	84.71%	0.3338	**0.0017**	0.0360	0.0298	0.0658
`Prox-DASA-GT`	84.69%	0.3342	**0.0017**	0.0390	0.0301	0.0691
				MNIST		
`SPPDM`	76.54%	0.7854	0.0436	0.1587	0.1246	0.2833
`ProxGT-SR-E`	92.26%	0.3042	0.0250	0.1771	0.3368	0.5139
`DEEPSTORM v2`	94.52%	0.1759	**0.0016**	0.1758	0.2030	0.3788
`Prox-DASA`	96.74%	0.1469	0.0081	0.1912	0.1299	0.3211
`Prox-DASA-GT`	**96.84%**	**0.1460**	0.0058	0.1935	0.1317	0.3252

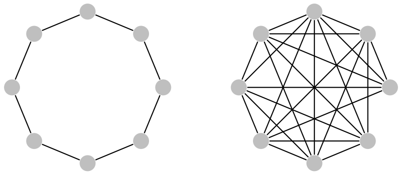
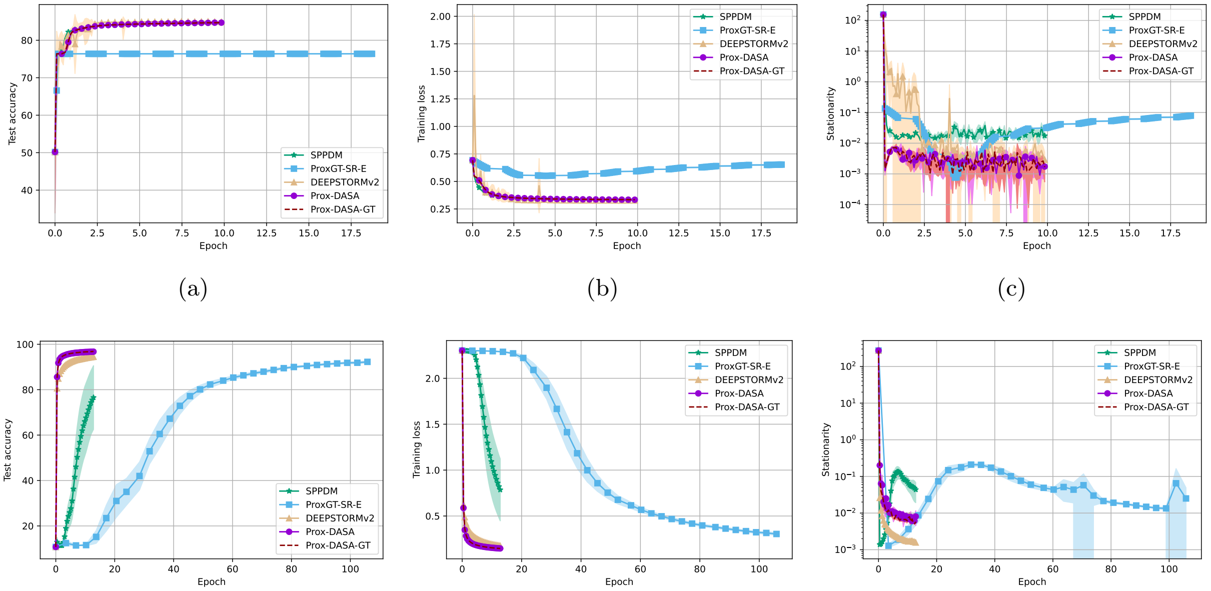


Figure A.1. Network topology. The left represents the ring topology and the right represents the random graph.



(a)

(b)

(c)

(d)

(e)

(f)

Figure A.2. Comparisons between `SPPDM` [**WZC+21**], `ProxGT-SR-E` [**XDKK21**], `DEEPSTORM` [**MBMXC22**], `Prox-DASA` 2, and `Prox-DASA-GT` 3. In each experiments `ProxGT-SR-E` computes 1 more epoch than other algorithms every q iterations. q is chosen to be 1000 for a9a and 32 for MNIST.

the all-one vector. We identify vectors at agent i in the subscript and use the superscript for the algorithm step. For example, the optimization variable of agent i at step k is denoted as x_i^k, and z_i^k is the corresponding dual variable. We use uppercase bold letters to represent the matrix that collects all the variables from agents (corresponding lowercase) as columns. To be specific,

$$\mathbf{X}_k = \left[x_1^k, \ldots, x_n^k\right], \quad \mathbf{Z}_k = \left[z_1^k, \ldots, z_n^k\right], \quad \mathbf{Y}_k = \left[y_1^k, \ldots, y_n^k\right], \quad \mathbf{V}_{k+1} = \left[v_1^{k+1}, \ldots, v_n^{k+1}\right].$$

We add an overbar to a letter to denote the average over all agents. For example,

$$\bar{x}^k = \frac{1}{n}\sum_{i=1}^n x_i^k = \frac{1}{n}\mathbf{X}_k\mathbf{1}, \quad \bar{\mathbf{X}}_k = [\bar{x}^k, \ldots, \bar{x}^k] = \bar{x}^k\mathbf{1}^\top = \frac{1}{n}\mathbf{X}_k\mathbf{1}\mathbf{1}^\top$$

Hence, the consensus errors for iterates $\{x_i^k\}$ and dual variables $\{z_i^k\}$ can be written as

$$\frac{1}{n}\sum_{i=1}^n \left\|x_i^k - \bar{x}^k\right\|^2 = \frac{1}{n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2, \quad \frac{1}{n}\sum_{i=1}^n \left\|z_i^k - \bar{z}^k\right\|^2 = \frac{1}{n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2.$$

We denote $L_{\nabla F} = \max_{1\le i\le n}\{L_{\nabla F_i}\}$ for ease of presentation. Our proof heavily relies on the merit function below:

$$W(\bar{x}^k, \bar{z}^k) = \underbrace{\Phi(\bar{x}^k) - \Phi_*}_{\text{function value gap}} + \underbrace{\Psi(\bar{x}^k) - \eta(\bar{x}^k, \bar{z}^k)}_{\text{primal convergence}} + \lambda \underbrace{\left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2}_{\text{dual convergence}}, \tag{A.1}$$

where

$$\eta(x, z) = \min_{y\in\mathbb{R}^d}\left\{\langle z, y - x\rangle + \frac{1}{2\gamma}\|y - x\|^2 + \Psi(y)\right\}. \tag{A.2}$$

A.2.1. Technical Lemmas.

Lemma A.1. *For any $p, q, r \in \mathbb{N}_+$ and matrix $\mathbf{A} \in \mathbb{R}^{p\times q}, \mathbf{B} \in \mathbb{R}^{q\times r}$, we have:*

$$\|\mathbf{A}\mathbf{B}\| \le \min\left(\|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|, \|\mathbf{A}\| \cdot \|\mathbf{B}^\top\|_2\right).$$

Lemma A.2. *Suppose $\mathbf{W}$ satisfies Assumption 3.1. For any $m \in \mathbb{N}_+$, we have*

$$\left\|\mathbf{W}^m - \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right\|_2 \le \rho^m$$

LEMMA A.3. *Suppose we are given three sequences* $\{a_n\}_{n=0}^\infty, \{b_n\}_{n=0}^\infty$, $\{\tau_n\}_{n=-1}^\infty$, *and a constant* r *satisfying*

$$a_{k+1} \leq r a_k + b_k,\ a_k \geq 0,\ b_k \geq 0,\ 0 = \tau_{-1} \leq \tau_{k+1} \leq \tau_k \leq 1, \tag{A.3}$$

for all $k \geq 0$. *Then for any* $K > 0$, *we have*

$$\sum_{k=0}^K \tau_k a_k \leq \frac{1}{1-r}\left(\tau_0 a_0 + \sum_{k=0}^K \tau_k b_k\right)$$

PROOF. Note that we have

$$(1-r)\sum_{k=0}^K \tau_k a_k \leq \sum_{k=0}^K \tau_k (a_k - a_{k+1} + b_k) = \sum_{k=0}^K (\tau_k - \tau_{k-1}) a_k - \tau_K a_{K+1} + \sum_{k=0}^K \tau_k b_k \leq \tau_0 a_0 + \sum_{k=0}^K \tau_k b_k,$$

where the inequalities use (A.3), and the equality uses summation by parts. □

LEMMA A.4. *Let* $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ *be a closed proper convex function.*

(a) Let $\eta(x,z)$ *be the function defined in* (A.2). *Then,* $\nabla\eta$ *is* C_γ*-Lipschitz continuous where*

$$C_\gamma = 2\sqrt{(1+\frac{1}{\gamma})^2 + (1+\frac{\gamma}{2})^2}. \tag{A.4}$$

(b) For $x, z \in \mathbb{R}^d$ *and* $\gamma \in \mathbb{R}$, *let* $y_+ = \mathbf{prox}_\Psi^\gamma(x - \gamma z) = \arg\min_{y\in\mathbb{R}^d}\left\{\langle z, y-x\rangle + \frac{1}{2\gamma}\|y-x\|^2 + \Psi(y)\right\}$, *then for any* $y \in \mathbb{R}^d$, *we have*

$$\Psi(y_+) - \Psi(y) \leq \left\langle z + \gamma^{-1}(y_+ - x), y - y_+\right\rangle$$

PROOF. We prove (a) at first. Recall that the Moreau envelope of a convex and closed function Ψ multiplied by a scalar γ is defined by

$$\mathrm{env}_{\gamma\Psi}(x) = \min_{y\in\mathbb{R}^d}\left\{\frac{1}{2\gamma}\|y-x\|^2 + \Psi(y)\right\},$$

and its gradient is given by $\nabla \mathrm{env}_{\gamma\Psi}(x) = \frac{1}{\gamma}(x - \mathbf{prox}_\Psi^\gamma(x))$ where $\mathbf{prox}_\Psi^\gamma(x) = \arg\min_{y\in\mathbb{R}^d}\left\{\frac{1}{2\gamma}\|y-x\|^2 + \Psi(y)\right\}$. Note that $\eta(x,z) = \mathrm{env}_{\gamma\Psi}(x - \gamma z) - \frac{\gamma}{2}\|z\|^2$. Therefore, the partial gradients of η are given by

$$\nabla_x \eta(x,z) = -z - \gamma^{-1}\left(\mathbf{prox}_\Psi^\gamma(x-\gamma z) - x\right), \quad \nabla_z \eta(x,z) = \mathbf{prox}_\Psi^\gamma(x - \gamma z) - x. \tag{A.5}$$

Hence, for any (x,z) and (x',z'),

$$\begin{aligned}\left\|\nabla\eta(x,z)-\nabla\eta(x',z')\right\| &\le \left\|\nabla_x\eta(x,z)-\nabla_x\eta(x',z')\right\|+\left\|\nabla_z\eta(x,z)-\nabla_z\eta(x',z')\right\| \\ &\le 2(1+1/\gamma)\left\|x-x'\right\|+(2+\gamma)\left\|z-z'\right\|\le C_\gamma\left\|(x,z)-(x',z')\right\|.\end{aligned}$$

To prove (b), denote the subdifferential of $\Psi(x)$ as $\partial\Psi(x)$. By the optimality condition, we have $\mathbf{0}$ is a subgradient of $H(y)=\langle z,y-x\rangle+\frac{1}{2\gamma}\|y-x\|^2+\Psi(y)$ at y_+, i.e.,

$$\mathbf{0}\in z+\gamma^{-1}(y_+-x)+\partial\Psi(y_+).$$

Hence, there exists a subgradient of $\Psi(y)$ at y_+, denoted by $\tilde\nabla\Psi(y_+)$, such that

$$\tilde\nabla\Psi(y_+)=-z-\gamma^{-1}(y_+-x).$$

Finally, by the convexity of Ψ, we have for any $y\in\mathbb{R}^d$,

$$\Psi(y)-\Psi(y_+)\ge\left\langle\tilde\nabla\Psi(y_+),y-y_+\right\rangle=\left\langle -z-\gamma^{-1}(y_+-x),y-y_+\right\rangle,$$

which completes the proof. □

A.2.2. Building Blocks of Main Proof. The following lemma connects the consensus error of $\mathbf{Y}$ to the consensus errors of $\mathbf{X}$ and $\mathbf{Z}$.

LEMMA A.5. *Let $y_+^k=\mathbf{prox}(\bar x^k-\gamma\bar z^k)$. Then for any $k\ge0$ and $\gamma>0$, we have*

$$\left\|y_+^k-\bar y^k\right\|^2+\frac{1}{n}\left\|\mathbf{Y}_k-\bar{\mathbf{Y}}_k\right\|^2=\frac{1}{n}\sum_{i=1}^n\left\|y_i^k-y_+^k\right\|^2\le\frac{2}{n}\left\{\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2+\gamma^2\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right\}.$$

PROOF. By the non-expansiveness of proximal operator, we have

$$\|y_i^k-y_+^k\|\le\|x_i^k-\bar x^k-\gamma\left(z_i^k-\bar z^k\right)\|\le\|x_i^k-\bar x^k\|+\gamma\|z_i^k-\bar z^k\|.$$

Hence we know the consensus error of y can be bounded

$$\begin{aligned}&\frac{1}{n}\|\mathbf{Y}_k-\bar{\mathbf{Y}}^k\|^2=\frac{1}{n}\sum_{i=1}^n\|y_i^k-\bar y^k\|^2=\frac{1}{n}\sum_{i=1}^n\|y_i^k-y_+^k+\frac{1}{n}\sum_{j=1}^n(y_+^k-y_j^k)\|^2\\ =&\frac{1}{n}\sum_{i=1}^n\|y_i^k-y_+^k\|^2-\|\frac{1}{n}\sum_{j=1}^n\left(y_j^k-y_+^k\right)\|^2\le\frac{1}{n}\sum_{i=1}^n\|y_i^k-y_+^k\|^2\end{aligned}$$

$$\leq \frac{2}{n}\left\{\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2 + \gamma^2 \|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2\right\}$$

where the third equality uses the fact that

$$\frac{1}{n}\sum_{i=1}^{n}\left\| v_i - \left(\frac{1}{n}\sum_{j=1}^{n} v_j\right)\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\|v_i\|^2 - \left\|\frac{1}{n}\sum_{j=1}^{n} v_j\right\|^2$$

for any vectors v_i $(1 \leq i \leq n)$. □

The following technical lemma explicitly characterizes the consensus error.

Lemma A.6 (Conensus Error of Algorithm 2: `Prox-DASA`). *Suppose Assumptions 3.1, 3.4, 3.5, 3.6, and 3.7 hold. Let* $\varrho(m) = \frac{(1+\rho^{2m})\rho^{2m}}{(1-\rho^{2m})^2}$, *and* ρ, m *and* α_k *satisfy*

$$\varrho(m)\alpha_k^2 \leq \min\left\{\frac{1}{8}, \frac{1}{24L_{\nabla F}^2\gamma^2}\right\},\ 0 = \alpha_{-1} \leq \alpha_{k+1} \leq \alpha_k \leq 1 \tag{A.6}$$

for any $k \geq 0$. *Then in Algorithm 2 for any* $p \geq 0$, *we have*

$$\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2\right] \leq 4\gamma^2(\sigma^2 + 3L_{\nabla F}^2\nu^2)\varrho(m)\sum_{k=0}^{K}\alpha_k^{p+2},$$

$$\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2\right] \leq 4(\sigma^2 + 3L_{\nabla F}^2\nu^2)\varrho(m)\sum_{k=0}^{K}\alpha_k^{p+2},.$$

Proof. By Assumption 3.1, the iterates in Algorithm 2 satisfy

$$\begin{aligned}
&\mathbf{X}_{k+1} = (1-\alpha_k)\mathbf{X}_k\mathbf{W}^m + \alpha_k\mathbf{Y}_k\mathbf{W}^m,\ \bar{x}^{k+1} = (1-\alpha_k)\bar{x}^k + \alpha_k\bar{y}^k,\\
&\mathbf{Z}_{k+1} = (1-\alpha_k)\mathbf{Z}_k\mathbf{W}^m + \alpha_k\mathbf{V}_{k+1}\mathbf{W}^m,\ \bar{z}^{k+1} = (1-\alpha_k)\bar{z}^k + \alpha_k\bar{v}^{k+1}.
\end{aligned} \tag{A.7}$$

Hence, for the consensus error of iterates $\{x_i^k\}$, we have

$$\begin{aligned}
&\left\|\mathbf{X}_{k+1} - \bar{\mathbf{X}}_{k+1}\right\|^2\\
=&\left\|\Big((1-\alpha_k)\left(\mathbf{X}_k - \bar{\mathbf{X}}_k\right) + \alpha_k\left(\mathbf{Y}_k - \bar{\mathbf{Y}}_k\right)\Big)\left(\mathbf{W}^m - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right\|^2\\
\leq&\left\{\left(1 + \frac{1-\rho^{2m}}{2\rho^{2m}}\right)(1-\alpha_k)^2\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 + \left(1 + \frac{2\rho^{2m}}{1-\rho^{2m}}\right)\alpha_k^2\left\|\mathbf{Y}_k - \bar{\mathbf{Y}}_k\right\|^2\right\}\rho^{2m}\\
\leq&\frac{(1+\rho^{2m})}{2}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 + \frac{(1+\rho^{2m})\rho^{2m}}{1-\rho^{2m}}\alpha_k^2\left\|\mathbf{Y}_k - \bar{\mathbf{Y}}_k\right\|^2,
\end{aligned} \tag{A.8}$$

where the first inequality uses Lemma A.1 and A.2. Combining (A.6), (A.8), and Lemma A.5, we have

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{X}_{k+1}-\bar{\mathbf{X}}_{k+1}\|^2\right] &\leq \frac{(1+\rho^{2m})}{2}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+\frac{(1-\rho^{2m})}{4}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2+\gamma^2\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]\\ &=\frac{(3+\rho^{2m})}{4}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+\frac{(1-\rho^{2m})\gamma^2}{4}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]\end{aligned}$$

Using Lemma A.3 in the above inequality with $\tau_k=\frac{\alpha_k^p}{n}$ for any fixed $p\geq 0$ we know

$$\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]\leq\sum_{k=0}^{K}\frac{\gamma^2\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]. \tag{A.9}$$

Similarly to (A.8), we can obtain the following results on the consensus error of dual variables $\{z_i^k\}$:

$$\left\|\mathbf{Z}_{k+1}-\bar{\mathbf{Z}}_{k+1}\right\|^2\leq\frac{(1+\rho^{2m})}{2}\left\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\right\|^2+\frac{(1+\rho^{2m})\rho^{2m}}{1-\rho^{2m}}\alpha_k^2\left\|\mathbf{V}_{k+1}-\bar{\mathbf{V}}_{k+1}\right\|^2, \tag{A.10}$$

Using (A.6) and Lemma A.3 in (A.10) with $\tau_k=\frac{\alpha_k^p}{n}$, we have

$$\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]\leq 2\varrho(m)\sum_{k=0}^{K}\frac{\alpha_k^{p+2}}{n}\mathbb{E}\left[\|\mathbf{V}_{k+1}-\bar{\mathbf{V}}_{k+1}\|^2\right]. \tag{A.11}$$

To bound $\|\mathbf{V}_{k+1}-\bar{\mathbf{V}}_{k+1}\|$ we first notice that

$$\begin{aligned}v_i^{k+1}-\bar{v}^{k+1}=&v_i^{k+1}-\mathbb{E}\left[v_i^{k+1}|\mathscr{F}_k\right]-\frac{1}{n}\sum_{j=1}^{n}(v_j^{k+1}-\mathbb{E}\left[v_j^{k+1}|\mathscr{F}_k\right])\\ &+\mathbb{E}\left[v_i^{k+1}|\mathscr{F}_k\right]-\nabla F_i(\bar{x}^k)+\nabla F_i(\bar{x}^k)-\nabla F(\bar{x}^k)+\nabla F(\bar{x}^k)-\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[v_j^{k+1}|\mathscr{F}_k\right]\\ &=\left(1-\frac{1}{n}\right)(v_i^{k+1}-\mathbb{E}\left[v_i^{k+1}|\mathscr{F}_k\right])-\frac{1}{n}\sum_{j\neq i}(v_j^{k+1}-\mathbb{E}\left[v_j^{k+1}|\mathscr{F}_k\right])\\ &+\left(1-\frac{1}{n}\right)\left(\nabla F_i(x_i^k)-\nabla F_i(\bar{x}^k)\right)+\nabla F_i(\bar{x}^k)-\nabla F(\bar{x}^k)+\frac{1}{n}\sum_{j\neq i}\left(\nabla F_j(\bar{x}^k)-\nabla F_i(x_j^k)\right)\end{aligned}$$

which gives

$$\begin{aligned}&\mathbb{E}\left[\|v_i^{k+1}-\bar{v}^{k+1}\|^2\right]\\ =&\left(1-\frac{1}{n}\right)^2\mathbb{E}\left[\|v_i^{k+1}-\mathbb{E}\left[v_i^{k+1}|\mathscr{F}_k\right]\|^2\right]+\frac{1}{n^2}\sum_{j\neq i}^{n}\mathbb{E}\left[\|v_j^{k+1}-\mathbb{E}\left[v_j^{k+1}|\mathscr{F}_k\right]\|^2\right]\end{aligned}$$

$$+\left\|\left(1-\frac{1}{n}\right)\left(\nabla F_i(x_i^k)-\nabla F_i(\bar{x}^k)\right)+\nabla F_i(\bar{x}^k)-\nabla F(\bar{x}^k)+\frac{1}{n}\sum_{j\neq i}\left(\nabla F_j(\bar{x}^k)-\nabla F_i(x_j^k)\right)\right\|^2$$

$$\leq\sigma^2+3L_{\nabla F}^2\left(\left(1-\frac{1}{n}\right)^2\|x_i^k-\bar{x}^k\|^2+\nu^2+\frac{1}{n}\sum_{j\neq i}\|x_j^k-\bar{x}^k\|^2\right),$$

where the first equality uses Assumption 3.5, and the second inequality uses Cauchy-Schwarz inequality, Assumptions 3.2, 3.6, and 3.7. Hence we have

$$\mathbb{E}\left[\|\mathbf{V}_{k+1}-\bar{\mathbf{V}}_{k+1}\|^2\right]\leq 6L_{\nabla F}^2\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+n\sigma^2+3nL_{\nabla F}^2\nu^2. \tag{A.12}$$

Combining (A.11) and (A.12), we have

$$\begin{aligned}\sum_{k=0}^K\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]\leq&2\varrho(m)\sum_{k=0}^K\left\{\frac{6L_{\nabla F}^2\alpha_k^{p+2}}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+(\sigma^2+3L_{\nabla F}^2\nu^2)\sum_{k=0}^K\alpha_k^{p+2}\right\}\\ \leq&\sum_{k=0}^K\left\{12\varrho(m)\alpha_k^2L_{\nabla F}^2\gamma^2\right\}\frac{\alpha_k^p}{n\gamma^2}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+2(\sigma^2+3L_{\nabla F}^2\nu^2)\varrho(m)\sum_{k=0}^K\alpha_k^{p+2}\\ \leq&\sum_{k=0}^K\frac{\alpha_k^p}{2n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]+2(\sigma^2+3L_{\nabla F}^2\nu^2)\varrho(m)\sum_{k=0}^K\alpha_k^{p+2},\end{aligned}\tag{A.13}$$

where the second inequality uses (A.6). By (A.9) and (A.13) we can finally obtain that

$$\sum_{k=0}^K\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]\leq 4\gamma^2(\sigma^2+3L_{\nabla F}^2\nu^2)\varrho(m)\sum_{k=0}^K\alpha_k^{p+2},, \tag{A.14}$$

$$\sum_{k=0}^K\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]\leq 4(\sigma^2+3L_{\nabla F}^2\nu^2)\varrho(m)\sum_{k=0}^K\alpha_k^{p+2},. \tag{A.15}$$

□

Lemma A.7 (Conensus Error of Algorithm 3: `Prox-DASA-GT`). *Suppose Assumptions 3.1, 3.4, 3.6 and 3.5 hold. Let* $\varrho(m)=\frac{(1+\rho^{2m})\rho^{2m}}{(1-\rho^{2m})^2}$*, and* ρ, m *and* α_k *satisfy*

$$\varrho(m)\alpha_k^2\leq\frac{1}{8},\ \varrho(m)\alpha_k\leq\frac{1}{9L_{\nabla F}\gamma},\ 0=\alpha_{-1}\leq\alpha_{k+1}\leq\alpha_k\leq 1 \tag{A.16}$$

for any $k \geq 0$, and the initialization satisfies $u_i^0 = v_i^0 = 0$ for all i. Then in Algorithm 3 for any $p \geq 0$ we have

$$\sum_{k=0}^{K} \frac{\alpha_k^p}{n} \mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2\right] \leq 40\gamma^2 \varrho(m)^2 \sum_{k=0}^{K} \alpha_k^{p+2} \left\{ L_{\nabla F}^2 \alpha_k^2 \mathbb{E}\left[\|\bar{x}^k - \bar{y}^k\|^2\right] + 2\sigma^2 \right\},$$
$$\sum_{k=0}^{K} \frac{\alpha_k^p}{n} \mathbb{E}\left[\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2\right] \leq 40\varrho(m)^2 \sum_{k=0}^{K} \alpha_k^{p+2} \left\{ L_{\nabla F}^2 \alpha_k^2 \mathbb{E}\left[\|\bar{x}^k - \bar{y}^k\|^2\right] + 2\sigma^2 \right\}.$$

Proof. The updates in Algorithm 3 take the form:

$$\begin{aligned} &\mathbf{X}_{k+1} = (1-\alpha_k)\mathbf{X}_k \mathbf{W}^m + \alpha_k \mathbf{Y}_k \mathbf{W}^m, \ \bar{x}^{k+1} = (1-\alpha_k)\bar{x}^k + \alpha_k \bar{y}^k, \\ &\mathbf{U}_{k+1} = \mathbf{U}_k \mathbf{W}^m + (\mathbf{V}_{k+1} - \mathbf{V}_k)\mathbf{W}^m, \ \bar{u}^{k+1} = \bar{u}^k + \bar{v}^{k+1} - \bar{v}^k, \\ &\mathbf{Z}_{k+1} = (1-\alpha_k)\mathbf{Z}_k \mathbf{W}^m + \alpha_k \mathbf{U}_k \mathbf{W}^m, \ \bar{z}^{k+1} = (1-\alpha_k)\bar{z}^k + \alpha_k \bar{u}^k. \end{aligned} \tag{A.17}$$

Setting $u_i^0 = v_i^0$, we can prove by induction that $\bar{u}^k = \bar{v}^k$. To analyze the consensus error of $\mathbf{U}_k$, we first notice:

$$\begin{aligned} &\mathbf{U}_{k+1} - \bar{\mathbf{U}}_{k+1} \\ =&\left(\mathbf{U}_k - \bar{\mathbf{U}}_k + \mathbf{V}_{k+1} - \mathbf{V}_k - \bar{\mathbf{V}}^{k+1} + \bar{\mathbf{V}}^k\right)\left(\mathbf{W}^m - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right) \\ =&\left(\mathbf{U}_k - \bar{\mathbf{U}}_k + (\mathbf{V}_{k+1} - \mathbf{V}_k)\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right)\left(\mathbf{W}^m - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right) \end{aligned}$$

which gives

$$\begin{aligned} &\|\mathbf{U}_{k+1} - \bar{\mathbf{U}}_{k+1}\|^2 \\ \leq&\left\{\left(1 + \frac{1-\rho^{2m}}{2\rho^{2m}}\right)\left\|\mathbf{U}_k - \bar{\mathbf{U}}_k\right\|^2 + \left(1 + \frac{2\rho^{2m}}{1-\rho^{2m}}\right)\|\mathbf{V}_{k+1} - \mathbf{V}_k\|^2\right\}\rho^{2m} \\ =&\frac{(1+\rho^{2m})}{2}\left\|\mathbf{U}_k - \bar{\mathbf{U}}_k\right\|^2 + \frac{(1+\rho^{2m})\rho^{2m}}{1-\rho^{2m}}\|\mathbf{V}_{k+1} - \mathbf{V}_k\|^2. \end{aligned}$$

Using Lemma A.3, we know for any $k \geq 0$ and $p \geq 0$,

$$\sum_{k=0}^{K} \alpha_k^p \|\mathbf{U}_k - \bar{\mathbf{U}}_k\|^2 \leq 2\varrho(m) \sum_{k=0}^{K} \alpha_k^p \|\mathbf{V}_{k+1} - \mathbf{V}_k\|^2. \tag{A.18}$$

Note that we also have

$$\mathbf{V}_{k+1} - \mathbf{V}_k = \mathbf{V}_{k+1} - \mathbb{E}\left[\mathbf{V}_{k+1}|\mathscr{F}_k\right] - \left(\mathbf{V}_k - \mathbb{E}\left[\mathbf{V}_k|\mathscr{F}_{k-1}\right]\right)$$
$$+ \mathbb{E}\left[\mathbf{V}_{k+1}|\mathscr{F}_k\right] - \nabla\mathbf{F}(\bar{x}^k) + \nabla\mathbf{F}(\bar{x}^k) - \nabla\mathbf{F}(\bar{x}^{k-1}) + \nabla\mathbf{F}(\bar{x}^{k-1}) - \mathbb{E}\left[\mathbf{V}_k|\mathscr{F}_{k-1}\right]$$

where we overload the notation and define $\nabla\mathbf{F}(x) = [\nabla F_1(x), ..., \nabla F_n(x)]$. Hence we know

(A.19)

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{V}_{k+1} - \mathbf{V}_k\|^2\right] \\
\leq& 5\Bigg\{\mathbb{E}\left[\|\mathbf{V}_{k+1} - \mathbb{E}\left[\mathbf{V}_{k+1}|\mathscr{F}_k\right]\|^2\right] + \mathbb{E}\left[\|\mathbf{V}_k - \mathbb{E}\left[\mathbf{V}_k|\mathscr{F}_{k-1}\right]\|^2\right] + \mathbb{E}\left[\sum_{i=1}^n \|\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k)\|^2\right] \\
&+ \mathbb{E}\left[\sum_{i=1}^n \|\nabla F_i(\bar{x}^k) - \nabla F_i(\bar{x}^{k-1})\|^2\right] + \mathbb{E}\left[\sum_{i=1}^n \|\nabla F_i(x_i^{k-1}) - \nabla F_i(\bar{x}^{k-1})\|^2\right]\Bigg\} \\
\leq& 5\left(2n\sigma^2 + L_{\nabla F}^2\mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2 + \|\mathbf{X}^{k-1} - \bar{\mathbf{X}}^{k-1}\|^2 + n\alpha_{k-1}^2\|\bar{x}^{k-1} - \bar{y}^{k-1}\|^2\right]\right)
\end{aligned}$$

where the first inequality uses Cauchy-Schwarz inequality, and the second inequality uses Lipschitz continuity of ∇f_i and (A.17). For simplicity we set $x_i^{-1} = y_i^{-1} = 0$ for all i so that it is easy to check the above inequality holds for all $k \geq 0$. Using (A.18) and (A.19) we know:

(A.20)

$$\begin{aligned}
&\sum_{k=0}^K \frac{\alpha_k^p}{n}\|\mathbf{U}_k - \bar{\mathbf{U}}_k\|^2 \\
\leq& \frac{10\varrho(m)}{n}\sum_{k=0}^K \alpha_k^p\left(2n\sigma^2 + L_{\nabla F}^2\mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2 + \|\mathbf{X}^{k-1} - \bar{\mathbf{X}}^{k-1}\|^2 + n\alpha_{k-1}^2\|\bar{x}^{k-1} - \bar{y}^{k-1}\|^2\right]\right).
\end{aligned}$$

(A.21)

$$\leq \frac{20L_{\nabla F}^2\varrho(m)}{n}\sum_{k=0}^K \alpha_k^p\mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2\right] + 10L_{\nabla F}^2\varrho(m)\sum_{k=0}^K \alpha_k^{p+2}\mathbb{E}\left[\|\bar{x}^k - \bar{y}^k\|^2\right] + 20\sigma^2\varrho(m)\sum_{k=0}^K \alpha_k^p,$$

where the third inequality uses (A.16). For other consensus error terms we follow the same proof in Lemma A.6 to get

$$\left\|\mathbf{X}_{k+1} - \bar{\mathbf{X}}_{k+1}\right\|^2 \leq \frac{(1+\rho^{2m})}{2}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 + \frac{(1+\rho^{2m})\rho^{2m}}{1-\rho^{2m}}\alpha_k^2\left\|\mathbf{Y}_k - \bar{\mathbf{Y}}_k\right\|^2, \tag{A.22}$$

$$\left\|\mathbf{Y}_k - \bar{\mathbf{Y}}_k\right\|^2 \leq 2(\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 + \gamma^2\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2), \tag{A.23}$$

$$\left\|\mathbf{Z}_{k+1} - \bar{\mathbf{Z}}_{k+1}\right\|^2 \leq \frac{(1+\rho^{2m})}{2}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2 + \frac{(1+\rho^{2m})\rho^{2m}}{1-\rho^{2m}}\alpha_k^2\left\|\mathbf{U}_k - \bar{\mathbf{U}}_k\right\|^2. \tag{A.24}$$

Hence we know (A.9) still holds:

$$\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]\le\sum_{k=0}^{K}\frac{\gamma^2\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]. \tag{A.25}$$

Applying Lemma (A.3) in (A.24) with $\tau_k=\frac{\alpha_k^p}{n}$, we have

$$\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]\le 2\varrho(m)\sum_{k=0}^{K}\frac{\alpha_k^{p+2}}{n}\mathbb{E}\left[\|\mathbf{U}_k-\bar{\mathbf{U}}_k\|^2\right]. \tag{A.26}$$

The above two inequalities together with (A.21) and (A.16) imply

$$\begin{aligned}&\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]\le 2\varrho(m)\gamma^2\sum_{k=0}^{K}\frac{\alpha_k^{p+2}}{n}\mathbb{E}\left[\|\mathbf{U}_k-\bar{\mathbf{U}}_k\|^2\right]\\ \le&\sum_{k=0}^{K}\left\{40L_{\nabla F}^2\gamma^2\varrho(m)^2\alpha_k^2\right\}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+20\gamma^2\varrho(m)^2\sum_{k=0}^{K}\alpha_k^{p+2}\left\{L_{\nabla F}^2\alpha_k^2\mathbb{E}\left[\|\bar{x}^k-\bar{y}^k\|^2\right]+2\sigma^2\right\}\\ \le&\frac{1}{2}\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+20\gamma^2\varrho(m)^2\sum_{k=0}^{K}\alpha_k^{p+2}\left\{L_{\nabla F}^2\alpha_k^2\mathbb{E}\left[\|\bar{x}^k-\bar{y}^k\|^2\right]+2\sigma^2\right\},\end{aligned}$$

which gives

$$\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]\le 40\gamma^2\varrho(m)^2\sum_{k=0}^{K}\alpha_k^{p+2}\left\{L_{\nabla F}^2\alpha_k^2\mathbb{E}\left[\|\bar{x}^k-\bar{y}^k\|^2\right]+2\sigma^2\right\}. \tag{A.27}$$

Combining (A.16), (A.21), (A.26), and (A.27), we obtain that

$$\begin{aligned}&\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]\le 2\varrho(m)\sum_{k=0}^{K}\frac{\alpha_k^{p+2}}{n}\mathbb{E}\left[\|\mathbf{U}_k-\bar{\mathbf{U}}_k\|^2\right]\\ \le&\frac{1}{2\gamma^2}\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+20\varrho(m)^2\sum_{k=0}^{K}\alpha_k^{p+2}\left\{L_{\nabla F}^2\alpha_k^2\mathbb{E}\left[\|\bar{x}^k-\bar{y}^k\|^2\right]+2\sigma^2\right\},\\ \le&40\varrho(m)^2\sum_{k=0}^{K}\alpha_k^{p+2}\left\{L_{\nabla F}^2\alpha_k^2\mathbb{E}\left[\|\bar{x}^k-\bar{y}^k\|^2\right]+2\sigma^2\right\}.\end{aligned}$$

□

Lemma A.8 (Basic Inequalities of Dual Convergence).

$$\delta^k=\frac{\nabla F(\bar{x}^k)-\nabla F(\bar{x}^{k+1})}{\alpha_k}+\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(x_i^k)-\nabla F(\bar{x}^k),\quad \bar{\Delta}^{k+1}=\bar{v}^{k+1}-\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(x_i^k). \tag{A.28}$$

Under Assumption 3.2, we have

$$\begin{aligned}\|\bar{z}^{k+1}-\nabla F(\bar{x}^{k+1})\|^2 \leq (1-\alpha_k)\left\|\bar{z}^k-\nabla F(\bar{x}^k)\right\|^2 + 2L_{\nabla F}^2\alpha_k\left\|\bar{x}^k-\bar{y}^k\right\|^2+\alpha_k^2\left\|\bar{\Delta}^{k+1}\right\|^2 \\ + \frac{2L_{\nabla F}^2\alpha_k}{n}\left\|\mathbf{X}_k-\bar{\mathbf{X}}_k\right\|^2+2\left\langle\alpha_k\bar{\Delta}^{k+1},(1-\alpha_k)\left(\bar{z}^k-\nabla F(\bar{x}^k)\right)+\alpha_k\delta^k\right\rangle,\end{aligned} \tag{A.29}$$

and

$$\begin{aligned}\left\|\bar{z}^{k+1}-\bar{z}^k\right\|^2 \leq \alpha_k^2\Bigg\{2\left\|\nabla F(\bar{x}^k)-\bar{z}^k\right\|^2+\frac{2L_{\nabla F}^2}{n}\left\|\mathbf{X}_k-\bar{\mathbf{X}}_k\right\|^2+\left\|\bar{\Delta}^{k+1}\right\|^2 \\ +2\left\langle\bar{\Delta}^{k+1},\frac{1}{n}\sum_{i=1}^n\nabla F_i(x_i^k)-\bar{z}^k\right\rangle\Bigg\}.\end{aligned} \tag{A.30}$$

Proof. By definitions in (A.28), we have

$$\bar{z}^{k+1}-\nabla F(\bar{x}^{k+1})=(1-\alpha_k)\left(\bar{z}^k-\nabla F(\bar{x}^k)\right)+\alpha_k\delta^k+\alpha_k\bar{\Delta}^{k+1},$$

Hence, we can get

$$\begin{aligned}&\left\|\bar{z}^{k+1}-\nabla F(\bar{x}^{k+1})\right\|^2\\ &=\left\|(1-\alpha_k)\left(\bar{z}^k-\nabla F(\bar{x}^k)\right)+\alpha_k\delta^k\right\|^2+\alpha_k^2\left\|\bar{\Delta}^{k+1}\right\|^2+2\left\langle\alpha_k\bar{\Delta}^{k+1},(1-\alpha_k)\left(\bar{z}^k-\nabla F(\bar{x}^k)\right)+\alpha_k\delta^k\right\rangle\\ &\leq(1-\alpha_k)\left\|\bar{z}^k-\nabla F(\bar{x}^k)\right\|^2+\alpha_k\left\|\delta^k\right\|^2+\alpha_k^2\left\|\bar{\Delta}^{k+1}\right\|^2+2\left\langle\alpha_k\bar{\Delta}^{k+1},(1-\alpha_k)\left(\bar{z}^k-\nabla F(\bar{x}^k)\right)+\alpha_k\delta^k\right\rangle\end{aligned}$$

where the inequality uses the convexity of $\|\cdot\|^2$. In addition, we have

$$\begin{aligned}\left\|\delta^k\right\|^2&\leq 2\left\|\frac{\nabla F(\bar{x}^k)-\nabla F(\bar{x}^{k+1})}{\alpha_k}\right\|^2+2\left\|\frac{1}{n}\sum_{i=1}^n\left(\nabla F_i(x_i^k)-\nabla F_i(\bar{x}^k)\right)\right\|^2\\ &\leq 2L_{\nabla F}^2\left\|\bar{x}^k-\bar{y}^k\right\|^2+\frac{2L_{\nabla F}^2}{n}\left\|\mathbf{X}_k-\bar{\mathbf{X}}_k\right\|^2,\end{aligned}$$

which completes the proof of (A.29). The inequality (A.30) can be proved similarly by noting that

$$\begin{aligned}&\left\|\bar{z}^{k+1}-\bar{z}^k\right\|^2=\alpha_k^2\left\|-\bar{z}^k+\bar{v}^{k+1}\right\|^2\\ &=\alpha_k^2\left\|(\nabla F(\bar{x}^k)-\bar{z}^k)+\left(\frac{1}{n}\sum_{i=1}^n\left(\nabla F_i(x_i^k)-\nabla F_i(\bar{x}^k)\right)\right)+\alpha_k\bar{\Delta}^{k+1}\right\|^2\end{aligned}$$

$$= \alpha_k^2\Big\{ \left\| (\nabla F(\bar{x}^k) - \bar{z}^k) + \left(\frac{1}{n}\sum_{i=1}^n \Big(\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k)\Big)\right)\right\|^2 + \left\|\bar{\Delta}^{k+1}\right\|^2 + 2\left\langle \bar{\Delta}^{k+1}, \frac{1}{n}\sum_{i=1}^n \nabla F_i(x_i^k) - \bar{z}^k \right\rangle \Big\}.$$

□

Lemma A.9. *Under Assumption 3.3,*

$$\Psi(\bar{y}^k) - \Psi(y_+^k) \le \left\langle \bar{z}^k + \gamma^{-1}(\bar{y}^k - \bar{x}^k), y_+^k - \bar{y}^k \right\rangle + \frac{\gamma}{2n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2 + \frac{\gamma^{-1}}{2n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2. \tag{A.31}$$

Proof. By the convexity of Ψ and part (b) of Lemma A.4, we have

$$\begin{aligned}\Psi(\bar{y}^k) - \Psi(y_+^k) &\overset{\text{cvx}}{\le} \frac{1}{n}\sum_{i=1}^n \Big(\Psi(y_i^k) - \Psi(y_+^k)\Big) \overset{\text{Lemma A.4 (b)}}{\le} \frac{1}{n}\sum_{i=1}^n \left\langle z_i^k + \gamma^{-1}(y_i^k - x_i^k), y_+^k - y_i^k \right\rangle \\ &= \left\langle \bar{z}^k + \gamma^{-1}(\bar{y}^k - \bar{x}^k), y_+^k - \bar{y}^k \right\rangle + \frac{1}{n}\sum_{i=1}^n \left\langle z_i^k - \bar{z}^k + \gamma^{-1}(y_i^k - \bar{y}^k + \bar{x}^k - x_i^k), \bar{y}^k - y_i^k \right\rangle \\ &\le \left\langle \bar{z}^k + \gamma^{-1}(\bar{y}^k - \bar{x}^k), y_+^k - \bar{y}^k \right\rangle + \frac{\gamma}{2n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2 + \frac{1}{2n\gamma}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2.\end{aligned}$$

The equality above comes from the fact that for sequences $\{a_i\}_{1\le i\le n}, \{b_i\}_{1\le i\le n} \in \mathbb{R}^d$, we have

$$\sum_{i=1}^n \left\langle a_i - \frac{1}{n}\sum_{i=1}^n a_i, b_i - \frac{1}{n}\sum_{i=1}^n b_i \right\rangle = \sum_{i=1}^n \langle a_i, b_i\rangle - \left(\frac{1}{n}\sum_{i=1}^n a_i\right)\left(\frac{1}{n}\sum_{i=1}^n b_i\right).$$

The last inequality above is obtained by Young's inequalities:

$$\begin{aligned}\left\langle z_i^k - \bar{z}^k, \bar{y}^k - y_i^k \right\rangle &\le \frac{\gamma}{2}\left\|z_i^k - \bar{z}^k\right\|^2 + \frac{1}{2\gamma}\left\|y_i^k - \bar{y}^k\right\|^2, \\ \gamma^{-1}\left\langle \bar{x}^k - x_i^k, \bar{y}^k - y_i^k \right\rangle &\le \frac{1}{2\gamma}\left\|x_i^k - \bar{x}^k\right\|^2 + \frac{1}{2\gamma}\left\|y_i^k - \bar{y}^k\right\|^2.\end{aligned}$$

□

Lemma A.10 (Basic Lemma of Merit Function Difference). *Let $W(\bar{x}^k, \bar{z}^k)$ be the merit function defined in* (A.1) *with $\lambda = \frac{\gamma^{-1}}{8L_{\nabla F}^2}$. Under Assumption 3.2, 3.3, for any $k \ge 0$, setting $\alpha_k \le \min\{\frac{\gamma^{-1}}{8L_{\nabla F}}, \frac{\gamma^{-1}}{8C_\gamma}, \frac{\gamma^{-1}}{32C_\gamma L_{\nabla F}^2}\}$, we have*

$$W(\bar{x}^{k+1}, \bar{z}^{k+1}) - W(\bar{x}^k, \bar{z}^k) \le -\alpha_k\left\{\Theta^k + \Upsilon^k + \alpha_k\Lambda^k + r^{k+1}\right\},$$

where

(A.32)
$$\Theta^k = \left\{ \frac{\gamma^{-1}}{4}\|\bar{x}^k - \bar{y}^k\|^2 + \frac{\lambda}{4}\left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2 \right\}, \quad \Lambda^k = \left\{ \frac{C_\gamma + 2\lambda}{2}\left\|\bar{\Delta}^{k+1}\right\|^2 \right\},$$
$$\Upsilon^k = \left\{ \frac{2\gamma(1+4\gamma^2 L_{\nabla F}^2)}{n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2 + \frac{2\left(\gamma^{-1} + 3\gamma L_{\nabla F}^2\right)}{n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 \right\},$$
$$r^{k+1} = \left\langle \bar{\Delta}^{k+1}, \bar{x}^k - y_+^k + C_\gamma \alpha_k \left(\frac{1}{n}\sum_{i=1}^n \nabla F_i(x_i^k) - \bar{z}^k \right) + 2\lambda\left((1-\alpha_k)\left(\bar{z}^k - \nabla F(\bar{x}^k)\right) + \alpha_k \delta^k \right) \right\rangle.$$

Proof. By the smoothness of F and η, we have

$$F(\bar{x}^{k+1}) - F(\bar{x}^k)$$

(A.33)
$$\leq \left\langle \nabla F(\bar{x}^k), \bar{x}^{k+1} - \bar{x}^k \right\rangle + \frac{L_{\nabla F}}{2}\|\bar{x}^{k+1} - \bar{x}^k\|^2 = -\alpha_k \left\langle \nabla F(\bar{x}^k), \bar{x}^k - \bar{y}^k \right\rangle + \frac{L_{\nabla F}\alpha_k^2}{2}\|\bar{x}^k - \bar{y}^k\|^2$$

$$\eta(\bar{x}^k, \bar{z}^k) - \eta(\bar{x}^{k+1}, \bar{z}^{k+1})$$
$$\leq \left\langle -\bar{z}^k - \gamma^{-1}(y_+^k - \bar{x}^k), \bar{x}^k - \bar{x}^{k+1} \right\rangle + \left\langle y_+^k - \bar{x}^k, \bar{z}^k - \bar{z}^{k+1} \right\rangle + \frac{C_\gamma}{2}\left(\|\bar{x}^{k+1} - \bar{x}^k\|^2 + \|\bar{z}^{k+1} - \bar{z}^k\|^2 \right)$$
$$= 2\alpha_k \left\langle \bar{z}^k, y_+^k - \bar{x}^k \right\rangle + \gamma^{-1}\alpha_k \|\bar{x}^k - y_+^k\|^2 + \alpha_k \left\langle \bar{v}^{k+1}, \bar{x}^k - \bar{y}^k \right\rangle$$

(A.34)
$$+ \alpha_k \left\langle \bar{z}^k + \gamma^{-1}(y_+^k - \bar{x}^k) + \bar{v}^{k+1}, \bar{y}^k - y_+^k \right\rangle + \frac{C_\gamma}{2}\left(\alpha_k^2 \|\bar{x}^k - \bar{y}^k\|^2 + \|\bar{z}^{k+1} - \bar{z}^k\|^2 \right).$$

Since y_+^k is the minimizer of a $1/\gamma$-strongly convex function, i.e.,

$$\left\langle \bar{z}^k, y_+^k - \bar{x}^k \right\rangle + \frac{1}{2\gamma}\|y_+^k - \bar{x}^k\|^2 + \Psi(y_+^k) \leq \Psi(\bar{x}^k) - \frac{1}{2\gamma}\|y_+^k - \bar{x}^k\|^2,$$

which together with (A.34) gives

$$\eta(\bar{x}^k, \bar{z}^k) - \eta(\bar{x}^{k+1}, \bar{z}^{k+1})$$
$$\leq -\gamma^{-1}\alpha_k \|\bar{x}^k - y_+^k\|^2 + \alpha_k \left\langle \bar{v}^{k+1}, \bar{x}^k - \bar{y}^k \right\rangle + \alpha_k \left\langle \bar{z}_k + \gamma^{-1}(y_+^k - \bar{x}^k) + \bar{v}^{k+1}, \bar{y}^k - y_+^k \right\rangle$$
$$+ 2\alpha_k \left(\Psi(\bar{x}^k) - \Psi(y_+^k) \right) + \frac{C_\gamma}{2}\left(\|\bar{x}^{k+1} - \bar{x}^k\|^2 + \|\bar{z}^{k+1} - \bar{z}^k\|^2 \right). \tag{A.35}$$

By the convexity of Ψ, we have

$$\Psi(\bar{x}^{k+1}) - \Psi(\bar{x}^k) \leq (1-\alpha_k)\Psi(\bar{x}^k) + \alpha_k \Psi(\bar{y}^k) - \Psi(\bar{x}^k) = \alpha_k \left(\Psi(\bar{y}^k) - \Psi(\bar{x}_i^k)\right). \tag{A.36}$$

Combining (A.33), (A.35), and (A.36), we have

$$\begin{aligned}
&\Big[\Phi(\bar{x}^{k+1}) + \Psi(\bar{x}^{k+1}) - \eta(\bar{x}^{k+1}, \bar{z}^{k+1})\Big] - \Big[\Phi(\bar{x}^k) + \Psi(\bar{x}^k) - \eta(\bar{x}^k, \bar{z}^k)\Big] \\
\leq & -\gamma^{-1}\alpha_k \|\bar{x}^k - y_+^k\|^2 + \alpha_k \left\langle \bar{v}^{k+1} - \nabla F(\bar{x}^k), \bar{x}^k - \bar{y}^k \right\rangle + 2\alpha_k(\Psi(\bar{y}^k) - \Psi(y_+^k)) \\
& + \alpha_k \left\langle \bar{z}^k + \gamma^{-1}(y_+^k - \bar{x}^k) + \bar{v}^{k+1}, \bar{y}^k - y_+^k \right\rangle + \frac{(L_{\nabla F} + C_\gamma)\alpha_k^2}{2}\|\bar{x}^k - \bar{y}^k\|^2 + \frac{C_\gamma}{2}\|\bar{z}^{k+1} - \bar{z}^k\|^2.
\end{aligned} \tag{A.37}$$

Removing non-smooth terms in (A.37) using (A.31) in Lemma A.9, and re-organizing (A.37) using the decomposition that $\bar{z}^{k+1} - \bar{z}^k = \alpha_k(-\bar{z}^k + \bar{v}^{k+1}) = \alpha_k(\nabla F(\bar{x}^k) - \bar{z}^k) + \alpha_k(\frac{1}{n}\sum_{i=1}^n (\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k))) + \alpha_k \bar{\Delta}^{k+1}$, we can get

$$\begin{aligned}
&\Big[\Phi(\bar{x}^{k+1}) + \Psi(\bar{x}^{k+1}) - \eta(\bar{x}^{k+1}, \bar{z}^{k+1})\Big] - \Big[\Phi(\bar{x}^k) + \Psi(\bar{x}^k) - \eta(\bar{x}^k, \bar{z}^k)\Big] \\
\leq & \underbrace{\gamma^{-1}\alpha_k \Big\{ -\|\bar{x}^k - y_+^k\|^2 + \left\langle (y_+^k - \bar{y}^k) + (\bar{x}^k - \bar{y}^k), \bar{y}^k - y_+^k \right\rangle \Big\}}_{\varkappa_1} \\
& + \underbrace{\alpha_k \left\langle \frac{1}{n}\sum_{i=1}^n \left(\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k)\right), \bar{x}^k - y_+^k \right\rangle}_{\varkappa_2} + \underbrace{\alpha_k \left\langle \nabla F(\bar{x}^k) - \bar{z}^k, \bar{y}^k - y_+^k \right\rangle}_{\varkappa_3} + \alpha_k \left\langle \bar{\Delta}^{k+1}, \bar{x}^k - y_+^k \right\rangle \\
& \frac{(L_{\nabla F} + C_\gamma)\alpha_k^2}{2}\|\bar{x}^k - \bar{y}^k\|^2 + \underbrace{\frac{C_\gamma}{2}\|\bar{z}^{k+1} - \bar{z}^k\|^2}_{\varkappa_4} + \frac{\gamma\alpha_k}{n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2 + \frac{\gamma^{-1}\alpha_k}{n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2.
\end{aligned}$$

To further simplify the above inequalities, we analyze the terms $\varkappa_1, \varkappa_2, \varkappa_3, \varkappa_4$ separately as follows:

$$\begin{aligned}
\varkappa_1 =& \gamma^{-1}\alpha_k \left\{ -\left\|\bar{x}^k - \bar{y}^k\right\|^2 - \left\langle \bar{x}^k - \bar{y}^k, \bar{y}^k - y_+^k \right\rangle - 2\left\|\bar{y}^k - y_+^k\right\|^2 \right\} \leq -\frac{7\gamma^{-1}\alpha_k}{8}\left\|\bar{x}^k - \bar{y}^k\right\|^2, \\
\varkappa_2 \leq& 2\gamma\alpha_k \left\| \frac{1}{n}\sum_{i=1}^n \left(\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k)\right) \right\|^2 + \frac{\gamma^{-1}\alpha_k}{8}\left\|\bar{x}^k - y_+^k\right\|^2 \\
\leq& \frac{2\gamma\alpha_k L_{\nabla F}^2}{n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 + \frac{\gamma^{-1}\alpha_k}{4}\left\|\bar{x}^k - \bar{y}^k\right\|^2 + \frac{\gamma^{-1}\alpha_k}{4}\left\|\bar{y}^k - y_+^k\right\|^2, \\
\varkappa_3 \leq& \frac{\lambda\alpha_k}{2}\left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2 + \frac{\lambda^{-1}\alpha_k}{2}\left\|\bar{y}^k - y_+^k\right\|^2,
\end{aligned}$$

$$\varkappa_4 \leq \frac{C_\gamma \alpha_k^2}{2} \left\{ 2 \left\| \nabla F(\bar{x}^k) - \bar{z}^k \right\|^2 + \frac{2L_{\nabla F}^2}{n} \left\| \mathbf{X}_k - \bar{\mathbf{X}}_k \right\|^2 + \left\| \bar{\Delta}^{k+1} \right\|^2 + 2 \left\langle \bar{\Delta}^{k+1}, \frac{1}{n} \sum_{i=1}^n \nabla F_i(x_i^k) - \bar{z}^k \right\rangle \right\}.$$

Combining the above results with (A.29) in Lemma A.8 and the definition of $W(\bar{x}^k, \bar{z}^k)$ in (A.1), we have

$$\begin{aligned} & W(\bar{x}^{k+1}, \bar{z}^{k+1}) - W(\bar{x}^k, \bar{z}^k) \leq \alpha_k \left\{ -\frac{5}{8}\gamma^{-1} + \frac{(L_{\nabla F} + C_\gamma)\alpha_k}{2} + 2\lambda L_{\nabla F}^2 \right\} \|\bar{x}^k - \bar{y}^k\|^2 \\ & + \alpha_k \left\{ -\frac{\lambda}{2} + C_\gamma \alpha_k \right\} \left\| \nabla F(\bar{x}^k) - \bar{z}^k \right\|^2 + \frac{C_\gamma \alpha_k^2}{2} \left\| \bar{\Delta}^{k+1} \right\|^2 + \frac{(\gamma^{-1} + 2\lambda^{-1})\alpha_k}{4} \left\| y_+^k - \bar{y}^k \right\|^2 \\ & + \frac{\gamma \alpha_k}{n} \left\| \mathbf{Z}_k - \bar{\mathbf{Z}}_k \right\|^2 + \frac{\left(\gamma^{-1} + 2\gamma L_{\nabla F}^2 + 2\lambda L_{\nabla F}^2 + C_\gamma L_{\nabla F}^2 \alpha_k\right) \alpha_k}{n} \left\| \mathbf{X}_k - \bar{\mathbf{X}}_k \right\|^2 \end{aligned}$$

(A.38)

$$+ \alpha_k \underbrace{\left\langle \bar{\Delta}^{k+1}, \bar{x}^k - y_+^k + C_\gamma \alpha_k \left(\frac{1}{n} \sum_{i=1}^n \nabla F_i(x_i^k) - \bar{z}^k \right) + 2\lambda \left((1 - \alpha_k) \left(\bar{z}^k - \nabla F(\bar{x}^k) \right) + \alpha_k \delta^k \right) \right\rangle}_{r^{k+1}}.$$

In addition, from Lemma A.5, we already know

$$\left\| y_+^k - \bar{y}^k \right\|^2 \leq \frac{2}{n} \left\{ \|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2 + \gamma^2 \|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2 \right\}.$$

Finally, choosing α_k such that $\alpha_k \leq \min\{\frac{\gamma^{-1}}{8L_{\nabla F}}, \frac{\gamma^{-1}}{8C_\gamma}, \frac{\gamma^{-1}}{32C_\gamma L_{\nabla F}^2}\}$ and $\lambda = \frac{\gamma^{-1}}{8L_{\nabla F}^2}$, we can re-organize the terms in (A.38) as follows and complete the proof.

$$\begin{aligned} & W(\bar{x}^{k+1}, \bar{z}^{k+1}) - W(\bar{x}^k, \bar{z}^k) \\ \leq & -\alpha_k \underbrace{\left\{ \frac{\gamma^{-1}}{4} \|\bar{x}^k - \bar{y}^k\|^2 + \frac{\lambda}{4} \left\| \nabla F(\bar{x}^k) - \bar{z}^k \right\|^2 \right\}}_{\Theta^k} + \alpha_k^2 \underbrace{\left\{ \frac{C_\gamma + 2\lambda}{2} \left\| \bar{\Delta}^{k+1} \right\|^2 \right\}}_{\Lambda^k} + \alpha_k r^k \\ & + \alpha_k \underbrace{\left\{ \frac{2\gamma(1 + 4\gamma^2 L_{\nabla F}^2)}{n} \left\| \mathbf{Z}_k - \bar{\mathbf{Z}}_k \right\|^2 + \frac{2\left(\gamma^{-1} + 3\gamma L_{\nabla F}^2\right)}{n} \left\| \mathbf{X}_k - \bar{\mathbf{X}}_k \right\|^2 \right\}}_{\Upsilon^k}. \end{aligned}$$

(A.39)

□

A.3. Discussion on Different Types of Consensus Errors

In this section, we briefly discuss two different functions that measure the consensus violation of vectors among agents. Suppose agent i has $x_i \in \mathbb{R}^d$, our consensus error can be viewed as

$$f(x_1, ..., x_n) = \frac{1}{n}\sum_{i=1}^{n} \|x_i - \bar{x}\|^2,$$

where $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$, while `SPPDM` in [**WZC$^+$21**] defines (see Eq. (4a), (4b), (5a), (5b), and (41) in [**WZC$^+$21**])

$$\begin{aligned} g_W(x_1, ..., x_n) &= \sum_{i\sim j, 1\le i<j\le n} \|x_i - x_j\|^2 \\ &= \frac{1}{2}\sum_{i=j \text{ or } i\sim j} \left(\|x_i - \bar{x}\|^2 + \|x_j - \bar{x}\|^2 - 2\langle x_i - \bar{x}, x_j - \bar{x}\rangle\right) \end{aligned} \tag{A.40}$$

over a connected network whose weighted adjacency matrix (i.e., mixing matrix) is W, and the stationarity therein is defined by using g_W. $i \sim j$ means agents i and j are neighbors. Note that in general the relationship between f and g_W largely depends on W. We consider several special cases:

- W is a complete graph. By (A.40) we have

$$g_W(x_1, ..., x_n) = n\sum_{i=1}^{n}\|x_i - \bar{x}\|^2 - \left\langle \sum_{i=1}^{n}(x_i - \bar{x}), \sum_{j=1}^{n}(x_j - \bar{x})\right\rangle = n^2 f(x_1, ..., x_n).$$

- W is a cycle. By (A.40) we have

$$g_W(x_1, ..., x_n) \le \sum_{i\sim j, 1\le i<j\le n} 2\left(\|x_i - \bar{x}\|^2 + \|x_j - \bar{x}\|^2\right) = 4nf(x_1, ..., x_n).$$

- W is a simple path such that i and $i+1$ are adjacent for all $1 \le i \le n-1$, and $x_i = i \in \mathbb{R}$. Note that in this case, we can directly obtain $g_W(x_1, ..., x_n) = n - 1$. For f we have

$$f(x_1, ..., x_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n+1}{2} - i\right)^2 = \Theta(n^2),$$

  which implies $g_W = \Theta(\frac{f}{n})$.

We know from the above examples that the order (in terms of n) of g_W/f can range from $\frac{1}{n}$ to n^2. Hence these two types of consensus error are not comparable if no additional assumptions are given, and thus we only include `SPPDM` in the experiments and do not compare their complexity results to ours.

APPENDIX B

Appendix of Chapter 4

The supplementary materials are organized as follows. Appendix 4.1.1 provides motivating examples for stochastic multilevel optimization. Appendix B.1 introduces the essential technical lemmas to complete the proof. We present the whole proofs of Theorem 4.1 and Theorem 4.2 in Appendix B.2 and B.3. Finally, we present the high-probability convergence analysis particularly for the case when $T = 1$ in Appendix B.4.

B.1. Technical Lemmas

LEMMA B.1. (Smoothness of Composite Functions [**BGN22**]) *Assume that Assumption 4.2 holds.*

a) Define $F_i(x) = f_i \circ f_{i+1} \circ \cdots \circ f_T(x)$. Under , the gradient of F_i is Lipschitz continuous with the constant

$$L_{\nabla F_i} = \sum_{j=i}^{T} \left[L_{\nabla f_j} \prod_{l=i}^{j-1} L_{f_l} \prod_{l=j+1}^{T} L_{f_l}^2 \right].$$

b) Define

$$R_1 = L_{\nabla f_1} L_{f_2} \cdots L_{f_T}, \qquad R_j = L_{f_1} \cdots L_{f_{j-1}} L_{\nabla f_j} L_{f_{j+1}} \cdots L_{f_T} / L_{f_j}, \quad 2 \le j \le T-1,$$
$$C_2 = R_1, \qquad C_j = \sum_{i=1}^{j-2} R_i \left(\prod_{l=i+1}^{j-1} L_{f_l} \right), \quad 3 \le j \le T \tag{B.1}$$

and let $u_{T+1} = x$. Then, for $T \ge 2$, we have

$$\left\| \nabla F(x) - \prod_{i=1}^{T} \nabla f_{T+1-i}(u_{T+2-i}) \right\| \le \sum_{j=2}^{T} C_j \| f_j(u_{j+1}) - u_j \|. \tag{B.2}$$

LEMMA B.2. (Smoothness of $\eta(\cdot,\cdot)$ [**GRW20**]) *For fixed $\beta > 0$ and, $\eta(x,z)$ defined in (4.10), the gradient of $\eta(x,z)$ w.r.t. (x,z) is Lipschitz continuous with the constant $L_{\nabla \eta} = 2\sqrt{(1+\beta)^2 + \left(1 + \frac{1}{2\beta}\right)^2}$.*

Lemma B.3. (Convergnece of `ICG` [**Jag13**]) *Let $\tilde{y}^k$ be the vector output by Algorithm 6 at step k, and y^k be the optimal solution of the subproblem 4.16, then under Assumption 4.1*

$$\frac{\beta}{2}\|\tilde{y}^k - y^k\|^2 \leq H_k(\tilde{y}^k) - H_k(y^k) \leq \frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{t_k+2}$$

where δ defined in Algorithm 6 is the quality of the linear minimization procedure.

Proof of Lemma B.3. The result is obtained by applying Theorem 1 in [**Jag13**] to H_k and noting that the curvature constant $C_{H_k} = \beta D_{\mathcal{X}}^2, \forall k \geq 0$. □

B.2. Proof of Theorem 4.1

To establish the rate of convergence for Algorithm 5 in Theorem 4.1, we first present Lemma B.4 and Lemma B.5 regarding the basic recursion on the errors in estimating the inner function values and the order of $\mathbb{E}[\|u_i^{k+1} - u_i^k\|^2|\mathscr{F}_k]$. The proofs follow [**BGN22**] with minor modifications. We present the complete proofs below for the reader's convenience.

Lemma B.4. *Let $\{x^k\}_{k\geq 0}$ and $\{u_i^k\}_{k\geq 0}$ be generated by Algorithm 5 and $u_{T+1} = x$. Define, $1 \leq i \leq T$,*

$$\begin{aligned} \Delta_{G_i}^{k+1} &:= f_i(u_{i+1}^k) - G_i^{k+1}, \quad \Delta_{J_i}^{k+1} := \nabla f_i(u_{i+1}^k) - J_i^{k+1}, \\ e_i^k &:= f_i(u_{i+1}^{k+1}) - f_i(u_{i+1}^k) - \langle \nabla f_i(u_{i+1}^k), u_{i+1}^{k+1} - u_{i+1}^k\rangle. \end{aligned} \tag{B.3}$$

Under Assumption 4.2, we have, for $1 \leq i \leq T$,

$$\begin{aligned} \|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\|^2 &\leq (1-\tau_k)\|f_i(u_{i+1}^k) - u_i^k\|^2 + \tau_k^2\|\Delta_{G_i}^{k+1}\|^2 + \dot{r}_i^{k+1} \\ &+ \left[4L_{f_i}^2 + L_{\nabla f_i}\|f_i(u_{i+1}^k) - u_i^k\| + \|\Delta_{J_i}^{k+1}\|^2\right]\|u_{i+1}^{k+1} - u_{i+1}^k\|^2, \end{aligned} \tag{B.4}$$

and

$$\|u_i^{k+1} - u_i^k\|^2 \leq \tau_k^2\left[2\|f_i(u_{i+1}^k) - u_i^k\|^2 + \|\Delta_{G_i}^{k+1}\|^2\right] + 2\|J_i^{k+1}\|^2\|u_{i+1}^{k+1} - u_{i+1}^k\|^2 + \ddot{r}_i^{k+1} \tag{B.5}$$

where

$$\begin{aligned} \dot{r}_i^{k+1} &:= 2\tau_k\langle \Delta_{G_i}^{k+1}, e_i^k + (1-\tau_k)(f_i(u_{i+1}^k) - u_i^k) + {\Delta_{J_i}^{k+1}}^\top(u_{i+1}^{k+1} - u_{i+1}^k)\rangle \\ &\quad + 2\langle {\Delta_{J_i}^{k+1}}^\top(u_{i+1}^{k+1} - u_{i+1}^k), e_i^k + (1-\tau_k)(f_i(u_{i+1}^k) - u_i^k)\rangle, \\ \ddot{r}_i^{k+1} &:= \tau_k\langle -\Delta_{G_i}^{k+1}, \tau_k(f_i(u_{i+1}^k) - u_i^k) + {J_i^{k+1}}^\top(u_{i+1}^{k+1} - u_{i+1}^k)\rangle. \end{aligned} \tag{B.6}$$

Proof. We first prove part (B.4). By the definitions in (B.3), (B.6), for any $1 \le i \le T$, we have

$$\begin{aligned}
&\|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\|^2 \\
=&\|e_i^k + f_i(u_{i+1}^k) + \nabla f_i(u_{i+1}^k)^\top (u_{i+1}^{k+1} - u_{i+1}^k) - (1-\tau_k)u_i^k - \tau_k G_i^{k+1} - {J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\|^2 \\
=&\|e_i^k + {\Delta_{J_i}^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k) + (1-\tau_k)(f_i(u_{i+1}^k) - u_i^k) + \tau_k \Delta_{G_i}^{k+1}\|^2 \\
=&\|{\Delta_{J_i}^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\|^2 + \|e_i^k + (1-\tau_k)(f_i(u_{i+1}^k) - u_i^k)\|^2 + \tau_k^2 \|\Delta_{G_i}^{k+1}\|^2 + \dot{r}_i^{k+1} \\
\le&\|e_i^k + (1-\tau_k)(f_i(u_{i+1}^k) - u_i^k)\|^2 + \tau_k^2 \|\Delta_{G_i}^{k+1}\|^2 + \|\Delta_{J_i}^{k+1}\|^2 \|u_{i+1}^{k+1} - u_{i+1}^k\|^2 + \dot{r}_i^{k+1} \\
\le&(1-\tau_k)\|f_i(u_{i+1}^k) - u_i^k\|^2 + \|e_i^k\|^2 + 2(1-\tau_k)\|e_i^k\|\|f_i(u_{i+1}^k) - u_i^k\| + \tau_k^2 \|\Delta_{G_i}^{k+1}\|^2 \\
&\quad + \|\Delta_{J_i}^{k+1}\|^2 \|u_{i+1}^{k+1} - u_{i+1}^k\|^2 + \dot{r}_i^{k+1}.
\end{aligned}$$

Furthermore, with Assumption 4.2, we have

$$\|e_i^k\| \le \frac{L_{\nabla f_i}}{2}\|u_{i+1}^{k+1} - u_{i+1}^k\|^2, \qquad \|e_i^k\|^2 \le 4L_{f_i}^2 \|u_{i+1}^{k+1} - u_{i+1}^k\|^2, \tag{B.7}$$

which leads to (B.4). To show (B.5), with the update rule given by (4.8) and the definitions in (B.3), we have, for $1 \le i \le T$,

$$\begin{aligned}
&\|u_i^{k+1} - u_i^k\|^2 \\
=&\|\tau_k(G_i^{k+1} - u_i^k) + \langle J_i^{k+1}, u_{i+1}^{k+1} - u_{i+1}^k\rangle\|^2 \\
=&\tau_k^2 \|G_i^{k+1} - u_i^k\|^2 + \|{J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\|^2 + 2\tau_k \langle G_i^{k+1} - u_i^k, {J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\rangle \\
=&\tau_k^2 \|G_i^{k+1} - u_i^k\|^2 + \|{J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\|^2 + 2\tau_k \langle f_i(u_{i+1}^k) - u_i^k, {J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\rangle \\
&\quad + 2\tau_k \langle -\Delta_{G_i}^{k+1}, {J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\rangle \\
\le&\tau_k^2 \|G_i^{k+1} - u_i^k\|^2 + 2\|J_i^{k+1}\|^2 \|u_{i+1}^{k+1} - u_{i+1}^k\|^2 + \tau_k^2 \|f_i(u_{i+1}^k) - u_i^k\|^2 \\
&\quad + 2\tau_k \langle -\Delta_{G_i}^{k+1}, {J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\rangle \\
=&2\tau_k^2 \|f_i(u_{i+1}^k) - u_i^k\|^2 + \tau_k^2 \|\Delta_{G_i}^{k+1}\|^2 + 2\|J_i^{k+1}\|^2 \|u_{i+1}^{k+1} - u_{i+1}^k\|^2 \\
&\quad + 2\tau_k \langle -\Delta_{G_i}^{k+1}, \tau_k(f_i(u_{i+1}^k) - u_i^k) + {J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\rangle.
\end{aligned}$$

where the inequality comes from the fact that $\|{J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\|^2 \le \|J_i^{k+1}\|^2 \|u_{i+1}^{k+1} - u_{i+1}^k\|^2$ and $2\tau_k \langle f_i(u_{i+1}^k) - u_i^k, {J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\rangle \le \|{J_i^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\|^2 + \tau_k^2 \|f_i(u_{i+1}^k) - u_i^k\|^2$. □

LEMMA B.5. *Let $u_{T+1} = x$. Under Assumption 4.2, 4.3, and with the choice of $\tau_0 = 1$, we have, for $1 \le i \le T$ and $k \ge 0$,*

$$\mathbb{E}[\|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\|^2 | \mathscr{F}_k] \le \sigma_{G_i}^2 + (4L_{f_i}^2 + \sigma_{J_i}^2)c_{i+1}, \tag{B.8}$$

$$\mathbb{E}[\|u_i^{k+1} - u_i^k\|^2 | \mathscr{F}_k] \le c_i \tau_k^2, \tag{B.9}$$

where

$$c_i := 3\sigma_{G_i}^2 + 2(4L_{f_i}^2 + \sigma_{J_i}^2 + \hat{\sigma}_{J_i}^2)c_{i+1}, \qquad c_{T+1} = D_{\mathcal{X}}^2. \tag{B.10}$$

PROOF. By the update rule given in (4.8) and the definitions in (B.3), for $1 \le i \le T$ and $k \ge 0$, we have

$$f_i(u_{i+1}^{k+1}) - u_i^{k+1} = (1 - \tau_k)(f_i(u_{i+1}^k) - u_i^k) + \mathbf{D}_{k,i},$$

where $\mathbf{D}_{k,i} := e_i^k + \tau_k \Delta_{G_i}^{k+1} + {\Delta_{J_i}^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)$. With the convexity of $\|\cdot\|^2$, we can further obtain

$$\|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\|^2 \le (1 - \tau_k)\|f_i(u_{i+1}^k) - u_i^k\|^2 + \frac{1}{\tau_k}\|\mathbf{D}_{k,i}\|^2, \quad \forall k \ge 0. \tag{B.11}$$

Moreover, under Assumption 4.3, we have, for $1 \le i \le T$ and $k \ge 0$,

$$\begin{aligned} \mathbb{E}[\|\mathbf{D}_{k,i}\|^2 | \mathscr{F}_k] &= \mathbb{E}[\|e_i^k\|^2 | \mathscr{F}_k] + \tau_k^2 \mathbb{E}[\|\Delta_{G_i}^{k+1}\|^2 | \mathscr{F}_k] + \mathbb{E}[\|{\Delta_{J_i}^{k+1}}^\top (u_{i+1}^{k+1} - u_{i+1}^k)\|^2 | \mathscr{F}_k] \\ &\le \tau_k^2 \mathbb{E}[\|\Delta_{G_i}^{k+1}\|^2 | \mathscr{F}_k] + \left(4L_{f_i}^2 + \mathbb{E}[\|\Delta_{J_i}^{k+1}\|^2 | \mathscr{F}_k]\right) \mathbb{E}[\|u_{i+1}^{k+1} - u_{i+1}^k\|^2 | \mathscr{F}_k] \\ &\le \tau_k^2 \sigma_{G_i}^2 + \left(4L_{f_i}^2 + \sigma_{J_i}^2\right) \mathbb{E}[\|u_{i+1}^{k+1} - u_{i+1}^k\|^2 | \mathscr{F}_k]. \end{aligned} \tag{B.12}$$

where the second inequality follows from (B.7). Setting $i = T$ in the inequality above and noting that $u_{T+1}^k = x^k$, we have

$$\mathbb{E}[\|\mathbf{D}_{k,T}\|^2 | \mathscr{F}_k] \le \tau_k^2 \left[\sigma_{G_T}^2 + (4L_{f_T}^2 + \sigma_{J_T}^2) D_{\mathcal{X}}^2\right], \quad \forall k \ge 0.$$

Thus, with the choice of $\tau_0 = 1$, we obtain

$$\mathbb{E}[\|f_T(x^k) - u_T^k\|^2 | \mathscr{F}_k] \le \sigma_{G_T}^2 + (4L_{f_T}^2 + \sigma_{J_T}^2) D_{\mathcal{X}}^2, \quad \forall k \ge 1.$$

Taking expectation of both sides of (B.5) conditioning on $\mathscr{F}_k$, and under Assumption 4.3, we obtain

$$\mathbb{E}[\|u_i^{k+1} - u_i^k\|^2 | \mathscr{F}_k] \le \tau_k^2 \mathbb{E}\left[2\|f_i(x^k) - u_i^k\|^2 + \|\Delta_{G_i}^{k+1}\|^2 + \frac{2}{\tau_k^2}\|J_i^{k+1}\|^2 \|u_{i+1}^{k+1} - u_{i+1}^k\|^2 \middle| \mathscr{F}_k\right]. \tag{B.13}$$

Setting $i = T$ in the inequality above, we have

$$\mathbb{E}[\|u_T^{k+1} - u_T^k\|^2|\mathscr{F}_k] \leq \tau_k^2 \left[3\sigma_{G_T}^2 + 2(4L_{f_T}^2 + \sigma_{J_T}^2 + \hat{\sigma}_{J_T}^2)D_{\mathcal{X}}^2\right], \quad \forall k \geq 1.$$

This completes the proof of (B.8) and (B.9) when $i = T$. We now use backward induction to complete the proof. By the above result, the base case of $i = T$ holds. Assume that (B.9) hold when $i = j$ for some $1 < j \leq T$, i.e., $\mathbb{E}[\|u_j^{k+1} - u_j^k\|^2|\mathscr{F}_k] \leq c_j\tau_k^2, \forall k \geq 0$. Then, setting $i = j-1$ in (B.12), we obtain

$$\mathbb{E}[\|\mathbf{D}_{k,j-1}\|^2|\mathscr{F}_k] \leq \tau_k^2 \left[\sigma_{G_{j-1}}^2 + (4L_{f_{j-1}}^2 + \sigma_{J_{j-1}}^2)c_j\right], \quad \forall k \geq 0.$$

Furthermore, with (B.11) and the choice of $\tau_0 = 1$, we have

$$\mathbb{E}[\|f_{j-1}(u_j^{k+1}) - u_{j-1}^{k+1}\|^2|\mathscr{F}_k] \leq \sigma_{G_{j-1}}^2 + (4L_{f_{j-1}}^2 + \sigma_{J_{j-1}}^2)c_j, \quad \forall k \geq 0.$$

which together with (B.13), imply that

$$\mathbb{E}[\|u_{j-1}^{k+1} - u_{j-1}^k\|^2|\mathscr{F}_k] \leq c_{j-1}\tau_k^2, \quad \forall k \geq 0.$$

□

We now leverage the merit function defined in (4.9) and provide a basic inequality for establishing convergence analysis of Algorithm 5 in Lemma B.6. In Proposition B.1, we show the boundedness of the term $\mathbf{R}_k$ appearing on the right hand side of (B.14) in expectation. These two results form the crucial steps in establishing the convergence analysis of Algorithm 5.

Lemma B.6. *Let $\{x^k, z^k, u^k\}_{k\geq 0}$ be the sequence generated by Algorithm 5, the merit function $W_{\alpha,\gamma}(\cdot,\cdot,\cdot)$ be defined in* (4.9) *with positive constants $\{\alpha, \{\gamma_i\}_{1\leq i\leq T}\}$, and $u_{T+1} = x$. Under Assumption 4.2, for any $\beta > 0$, let*

$$\beta_k \equiv \beta, \quad \alpha = \frac{\beta}{20L_{\nabla F}^2}, \quad \gamma_1 = \frac{\beta}{2}, \quad \gamma_j = \left(2\alpha + \frac{1}{4\alpha L_{\nabla F}^2}\right)(T-1)C_j^2 + \frac{\beta}{2}, \quad 2 \leq j \leq T,$$

where C_j's are defined in (B.1). *Then,* $\forall N \geq 0$

$$\begin{aligned}
&\sum_{k=0}^{N} \tau_k \left(\beta \left[\|d^k\|^2 + \sum_{i=1}^{T} \|f_i(u_{i+1}^k) - u_i^k\|^2 \right] + \frac{\beta}{20 L_{\nabla F}^2} \|\nabla F(x^k) - z^k\|^2 \right) \\
&\qquad \leq 2W_0 + 2 \sum_{k=0}^{N} \mathbf{R}_k + \left(\frac{24}{5} + \frac{40 L_{\nabla F}^2}{\beta^2} \right) \sum_{k=0}^{N} \tau_k \left(H_k(\tilde{y}^k) - H_k(y^k) \right),
\end{aligned} \tag{B.14}$$

where $d^k := y^k - x^k$, $H_k(\cdot), y^k$ *are defined in* (4.16), *and*

$$\begin{aligned}
\mathbf{R}_k :=& \sum_{i=1}^{T} \gamma_i \left[4L_{f_i}^2 + L_{\nabla f_i} \|f_i(u_{i+1}^k) - u_i^k\| + \|\Delta_{J_i}^{k+1}\|^2 \right] \|u_{i+1}^{k+1} - u_{i+1}^k\|^2 \\
&+ \tau_k^2 \left[\frac{L_{\nabla F} + L_{\nabla \eta}}{2} D_{\mathcal{X}}^2 + \sum_{i=1}^{T} \gamma_i \|\Delta_{G_i}^{k+1}\|^2 + \alpha \|\Delta^{k+1}\|^2 \right] \\
&+ \tau_k \left[\langle d^k, \Delta^{k+1} \rangle + \sum_{i=1}^{T} \gamma_i \dot{r}_i^{k+1} + 2\alpha \, \dddot{r}^{k+1} \right] + \frac{L_{\nabla \eta}}{2} \|z^{k+1} - z^k\|^2, \\
\Delta^{k+1} :=& \prod_{i=1}^{T} \nabla f_{T+1-i}(u_{T+2-i}^k) - \prod_{i=1}^{T} J_{T-i+1}^{k+1}, \\
\dddot{r}^{k+1} :=& \langle \Delta^{k+1}, (1 - \tau_k)[\nabla F(x^k) - z^k] + \tau_k [\nabla F(x^k) - \prod_{i=1}^{T} \nabla f_{T+1-i}(u_{T+2-i}^k)] \\
&+ \nabla F(x^{k+1}) - \nabla F(x^k) \rangle,
\end{aligned} \tag{B.15}$$

$\Delta_{G_i}^{k+1}$, $\Delta_{J_i}^{k+1}$ *are defined in* (B.3), *and* $\dot{r}_i^{k+1}$ *is defined in* (B.6).

Proof. We first bound $F(x^{k+1}) - F(x^k)$. By the Lipschitzness of ∇F (Lemma B.1), we have

$$\begin{aligned}
&F(x^{k+1}) - F(x^k) \leq \langle \nabla F(x^k), x^{k+1} - x^k \rangle + \frac{L_{\nabla F} \tau_k^2}{2} \|\tilde{d}^k\|^2 \\
&= \tau_k \langle \nabla F(x^k), d^k \rangle + \tau_k \langle \nabla F(x^k) - z^k, \tilde{y}^k - y^k \rangle - \tau_k \langle \beta d^k, \tilde{y}^k - y^k \rangle \\
&\qquad + \tau_k \langle z^k + \beta d^k, \tilde{y}^k - y^k \rangle + \frac{L_{\nabla F} \tau_k^2}{2} \|\tilde{d}^k\|^2 \\
&\leq \tau_k \langle \nabla F(x^k), d^k \rangle + \tau_k \|\nabla F(x^k) - z^k\| \|\tilde{y}^k - y^k\| + \tau_k \langle z^k + \beta d^k, \tilde{y}^k - y^k \rangle \\
&\qquad + \tau_k \beta \|d^k\| \|\tilde{y}^k - y^k\| + \frac{L_{\nabla F} \tau_k^2}{2} \|\tilde{d}^k\|^2.
\end{aligned} \tag{B.16}$$

We then provide a bound for $\eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1})$. By the lipschitzness of $\nabla\eta$ (Lemma B.2) with the partial gradients of $\nabla\eta$ given by

$$\nabla_x \eta(x^k, z^k) = -z^k - \beta d^k, \quad \nabla_z \eta(x^k, z^k) = d^k,$$

we have

$$\begin{aligned}
&\eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1}) \\
&\leq \big\langle z^k + \beta d^k, x^{k+1} - x^k \big\rangle - \big\langle d^k, z^{k+1} - z^k \big\rangle + \frac{L_{\nabla\eta}}{2}\Big[\|x^{k+1} - x^k\|^2 + \|z^{k+1} - z^k\|^2\Big] \\
&= \tau_k \big\langle 2z^k + \beta d^k, d^k \big\rangle + \tau_k \big\langle z^k + \beta d^k, \tilde{d}^k - d^k \big\rangle - \tau_k \big\langle d^k, \prod_{i=1}^{T} J_{T-i+1}^{k+1} \big\rangle \\
&\qquad + \frac{L_{\nabla\eta}}{2}\Big[\tau_k^2 \|\tilde{d}^k\|^2 + \|z^{k+1} - z^k\|^2\Big],
\end{aligned} \tag{B.17}$$

where the second equality comes from (4.6) and (4.7). Due to the optimality condition of in the definition of y^k, we have $\big\langle z^k + \beta d^k, x - y^k \big\rangle \geq 0$ for all $x \in \mathcal{X}$, which together with the choice of $x = x^k$ implies that

$$\langle z^k, d^k \rangle + \beta \|d^k\|^2 \leq 0. \tag{B.18}$$

Thus, combining (B.17) with (B.18), we obtain

$$\begin{aligned}
\eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1}) \leq -\beta\tau_k \|d^k\|^2 + \tau_k \big\langle z^k + \beta d^k, \tilde{y}^k - y^k \big\rangle - \tau_k \big\langle d^k, \prod_{i=1}^{T} J_{T-i+1}^{k+1} \big\rangle \\
+ \frac{L_{\nabla\eta}}{2}\Big[\tau_k^2 \|\tilde{d}^k\|^2 + \|z^{k+1} - z^k\|^2\Big].
\end{aligned} \tag{B.19}$$

In addition, by Lemma B.2, we have

$$\langle d^k, \nabla F(x^k) - \prod_{i=1}^{T} \nabla f_{T+1-i}(u_{T+2-i}^k) \rangle \leq \sum_{j=2}^{T} C_j \|d^k\| \|f_j(u_{j+1}^k) - u_j^k\|. \tag{B.20}$$

Then combing (B.16), (B.19), (B.20), we have

$$
\begin{aligned}
&[F(x^{k+1}) - \eta(x^{k+1}, z^{k+1})] - [F(x^k) - \eta(x^k, z^k)] \\
&\le \tau_k \Big\{ -\beta\|d^k\|^2 + \sum_{j=2}^{T} C_j \|d^k\| \|f_j(u_{j+1}^k) - u_j^k\| + \langle d^k, \Delta^{k+1}\rangle + 2\langle z^k + \beta d^k, \tilde{y}^k - y^k\rangle \\
&\quad + \Big[\beta\|d^k\| + \|\nabla F(x^k) - z^k\|\Big] \|\tilde{y}^k - y^k\| \Big\} + \frac{L_{\nabla F} + L_{\nabla\eta}}{2}\tau_k^2 \|\tilde{d}^k\|^2 + \frac{L_{\nabla\eta}}{2}\|z^{k+1} - z^k\|^2.
\end{aligned}
\tag{B.21}
$$

Furthermore, defining

$$
\varkappa^k := \nabla F(x^k) - \prod_{i=1}^{T} \nabla f_{T+1-i}(u_{T+2-i}^k), \qquad \bar{\varkappa}^k := \frac{\nabla F(x^{k+1}) - \nabla F(x^k)}{\tau_k},
$$

and by the update rule given by (4.7), we have

$$
\begin{aligned}
&\|\nabla F(x^{k+1}) - z^{k+1}\|^2 \\
&= \|(1-\tau_k)[\nabla F(x^k) - z^k] + \tau_k[\varkappa^k + \bar{\varkappa}^k + \Delta^{k+1}]\|^2 \\
&= \|(1-\tau_k)[\nabla F(x^k) - z^k] + \tau_k[\varkappa^k + \bar{\varkappa}^k]\|^2 + \tau_k^2\|\Delta^{k+1}\|^2 + 2\tau_k \dddot{r}^{k+1} \\
&\le (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + 2\tau_k \Big[\|\varkappa^k\|^2 + \|\bar{\varkappa}^k\|^2\Big] + \tau_k^2\|\Delta^{k+1}\|^2 + 2\tau_k \dddot{r}^{k+1} \\
&\le (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + \tau_k^2\|\Delta^{k+1}\|^2 \\
&\quad + 2\tau_k \left[(T-1)\sum_{j=2}^{T} C_j^2 \|f_j(u_{j+1}) - u_j\|^2 + 2L_{\nabla F}^2(\|d^k\|^2 + \|\tilde{y}^k - y^k\|^2) + \dddot{r}^{k+1}\right].
\end{aligned}
\tag{B.22}
$$

where $\dddot{r}^{k+1} := \langle \Delta^{k+1}, (1-\tau_k)[\nabla F(x^k) - z^k] + \tau_k[\varkappa^k + \bar{\varkappa}^k]\rangle$ and the last inequality comes from two fact that $\|\bar{\varkappa}^k\|^2 \le 2L_{\nabla F}^2(\|d^k\|^2 + \|\tilde{y}^k - y^k\|^2)$ and

$$
\|\varkappa^k\|^2 = \left\|\nabla F(x^k) - \prod_{i=1}^{T} \nabla f_{T+1-i}(u_{T+2-i}^k)\right\|^2 \le (T-1)\sum_{j=2}^{T} C_j^2 \|f_j(u_{j+1}) - u_j\|^2.
$$

The above upper bound for the term $\|\varkappa^k\|^2$ is obtained by leveraging Lemma B.2 and the fact that $(\sum_{i=1}^n a_i) \le n\sum_{i=1}^n a_i^2$ for non-negative sequence $(a_i)_{1\le i\le n}$.

Moreover, by Lemma B.4, we have, for $1 \le i \le T$,

$$
\begin{aligned}
\|f_i(u_{i+1}^{k+1}) - u_i^{k+1}\|^2 - \|f_i(u_{i+1}^k) - u_i^k\|^2 &\le -\tau_k\|f_i(u_{i+1}^k) - u_i^k\|^2 + \tau_k^2\|\Delta_{G_i}^{k+1}\|^2 + \dot{r}_i^{k+1} \\
&\quad + \Big[4L_{f_i}^2 + L_{\nabla f_i}\|f_i(u_{i+1}^k) - u_i^k\| + \|\Delta_{J_i}^{k+1}\|^2\Big] \|u_{i+1}^{k+1} - u_{i+1}^k\|^2,
\end{aligned}
\tag{B.23}
$$

Finally, multiplying both sides of (B.23) by γ_i for $i = 1, \ldots, T$ and both sides of (B.22) by α, adding them to (B.21), rearranging the terms, and noting that $\langle z^k + \beta d^k, \tilde{y}^k - y^k \rangle = H_k(\tilde{y}^k) - H_k(y^k) - (\beta/2)\|\tilde{y}^k - y^k\|^2$ due to the quadratic structure of H_k and $\|\tilde{d^k}\|^2 \leq D_{\mathcal{X}}^2$, we obtain

$$W_{k+1} - W_k \leq \tau_k \mathbf{A}_k + \mathbf{R}_k \tag{B.24}$$

where $\mathbf{R}_k$ is defined in (B.15) and

$$\begin{aligned}
\mathbf{A}_k := & \left(-\beta + 4\alpha L_{\nabla F}^2\right) \|d^k\|^2 + \sum_{j=2}^{T} \left(-\gamma_j + 2\alpha(T-1)C_j^2\right) \|f_j(u_{j+1}^k) - u_j^k\|^2 \\
& - \gamma_1 \|f_1(u_2^k) - u_1^k\|^2 - \alpha \|\nabla F(x^k) - z^k\|^2 + \sum_{j=2}^{T} C_j \|d^k\| \|f_j(u_{j+1}) - u_j\| \\
& + \Big(\beta\|d^k\| + \|\nabla F(x^k) - z^k\|\Big) \|\tilde{y}^k - y^k\| + \left(4\alpha L_{\nabla F}^2 - \beta\right) \|\tilde{y}^k - y^k\|^2 \\
& + 2\left(H_k(\tilde{y}^k) - H_k(y^k)\right).
\end{aligned}$$

We can further provide a simplified upper bound for $\mathbf{A}_k$. By Young's inequality, we have

$$\begin{aligned}
\beta\|d^k\|\|\tilde{y}^k - y^k\| &\leq \frac{\beta}{4}\|d^k\|^2 + \beta\|\tilde{y}^k - y^k\|^2, \\
\|\nabla F(x^k) - z^k\|\|\tilde{y}^k - y^k\| &\leq \frac{\alpha}{2}\|\nabla F(x^k) - z^k\|^2 + \frac{1}{2\alpha}\|\tilde{y}^k - y^k\|^2 \\
C_j\|d^k\|\|f_j(u_{j+1}) - u_j\| &\leq \frac{\alpha L_{\nabla F}^2}{T-1}\|d^k\|^2 + \frac{(T-1)C_j^2}{4\alpha L_{\nabla F}^2}\|f_j(u_{j+1}) - u_j\|^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbf{A}_k \leq & \left(-\frac{3\beta}{4} + 5\alpha L_{\nabla F}^2\right) \|d^k\|^2 - \gamma_1\|f_1(u_2^k) - u_1^k\|^2 - \frac{\alpha}{2}\|\nabla F(x^k) - z^k\|^2 \\
& + \sum_{j=2}^{T} \left(-\gamma_j + \left(2\alpha + \frac{1}{4\alpha L_{\nabla F}^2}\right)(T-1)C_j^2\right) \|f_j(u_{j+1}^k) - u_j^k\|^2 \\
& + \left(4\alpha L_{\nabla F}^2 + \frac{1}{2\alpha}\right) \|\tilde{y}^k - y^k\|^2 + 2\left(H_k(\tilde{y}^k) - H_k(y^k)\right)
\end{aligned}$$

For any $\beta > 0$, let

$$\alpha = \frac{\beta}{20L_{\nabla F}^2}, \qquad \gamma_1 = \frac{\beta}{2}, \qquad \gamma_j = \left(2\alpha + \frac{1}{4\alpha L_{\nabla F}^2}\right)(T-1)C_j^2 + \frac{\beta}{2}, \quad 2 \leq j \leq T$$

Then, we have

$$\begin{aligned}\mathbf{A}_k \leq -\frac{\beta}{2}\left(\|d^k\|^2 + \sum_{i=1}^{T}\|f_i(u_{i+1}^k) - u_i^k\|^2\right) - \frac{\beta}{40L_{\nabla F}^2}\|\nabla F(x^k) - z^k\|^2 \\ + \left(\frac{12}{5} + \frac{20L_{\nabla F}^2}{\beta^2}\right)\left(H_k(\tilde{y}^k) - H_k(y^k)\right).\end{aligned} \tag{B.25}$$

As a result of (B.24) and (B.25), we can further obtain

$$\begin{aligned}&\tau_k\left(\beta\left[\|d^k\|^2 + \sum_{i=1}^{T}\|f_i(u_{i+1}^k) - u_i^k\|^2\right] + \frac{\beta}{20L_{\nabla F}^2}\|\nabla F(x^k) - z^k\|^2\right) \\ &\leq 2W_k - 2W_{k+1} + 2\mathbf{R}_k + \tau_k\left(\frac{24}{5} + \frac{40L_{\nabla F}^2}{\beta^2}\right)\left(H_k(\tilde{y}^k) - H_k(y^k)\right),\end{aligned}$$

which immediately implies (B.14) by telescoping. □

Proposition B.1. *Let* $\mathbf{R}_k$ *be defined in* (B.15) *and* $\tau_0 = 1$*. Then, under Assumption 4.3, we have*

$$\mathbb{E}[\mathbf{R}_k|\mathscr{F}_k] \leq \hat{\sigma}^2\tau_k^2, \quad \forall k \geq 1,$$

where

$$\begin{aligned}\hat{\sigma}^2 := &\sum_{i=1}^{T}\gamma_i\left(\left[4L_{\nabla f_i}^2 + L_{\nabla f_i}\sqrt{\sigma_{G_i}^2 + (4L_{f_i}^2 + \sigma_{J_i}^2)c_{i+1}} + \sigma_{J_i}^2\right]c_{i+1} + \sigma_{G_i}^2\right) \\ &+ (\alpha + 2L_\eta)\prod_{i=1}^{T}\hat{\sigma}_{J_i}^2 + \frac{L_{\nabla F} + L_\eta}{2}D_{\mathcal{X}}^2.\end{aligned} \tag{B.26}$$

Proof. Note that under Assumption 4.3, we have, for $1 \leq i \leq T$,

$$\begin{aligned}&\mathbb{E}[\Delta^{k+1}|\mathscr{F}_k] = 0, \quad \mathbb{E}[\dot{r}_i^{k+1}|\mathscr{F}_k] = 0, \quad \mathbb{E}[\dddot{r}^{k+1}|\mathscr{F}_k] = 0, \\ &\mathbb{E}[\|\Delta_{G_i}^{k+1}\|^2|\mathscr{F}_k] \leq \sigma_{G_i}^2, \quad \mathbb{E}[\|\Delta_{J_i}^{k+1}\|^2|\mathscr{F}_k] \leq \sigma_{J_i}^2,\end{aligned}$$

and

$$\mathbb{E}[\|\Delta^{k+1}\|^2|\mathscr{F}_k] \leq \mathbb{E}\left[\left\|\prod_{i=1}^{T}J_{T-i+1}^{k+1}\right\|^2\middle|\mathscr{F}_k\right] \leq \prod_{i=1}^{T}\mathbb{E}\left[\left\|J_{T-i+1}^{k+1}\right\|^2\middle|\mathscr{F}_k\right] \leq \prod_{i=1}^{T}\hat{\sigma}_{J_i}^2.$$

In addition, by Lemma B.4 and Hölder's inequality. we have $\mathbb{E}[\|u_i^{k+1}-u_i^k\|^2|\mathscr{F}_k]\le c_i\tau_k^2$ and

$$\begin{aligned}
&\mathbb{E}[\|f_i(u_{i+1}^{k+1})-u_i^{k+1}\|\|u_i^{k+1}-u_i^k\|^2|\mathscr{F}_k]\\
&\le \mathbb{E}[\|f_i(u_{i+1}^{k+1})-u_i^{k+1}\||\mathscr{F}_k]\mathbb{E}[\|u_i^{k+1}-u_i^k\|^2|\mathscr{F}_k]\\
&\le \left(\mathbb{E}[\|f_i(u_{i+1}^{k+1})-u_i^{k+1}\||\mathscr{F}_k]\right)^{\frac{1}{2}}\mathbb{E}[\|u_i^{k+1}-u_i^k\|^2|\mathscr{F}_k]\\
&\le c_i\sqrt{\sigma_{G_i}^2+(4L_{f_i}^2+\sigma_{J_i}^2)c_{i+1}}\;\tau_k^2.
\end{aligned}$$

Lastly, from eq.(28) of Proposition 2.1 in **[BGN22]**, we have for any $k\ge 1$,

$$\mathbb{E}[\|z^{k+1}-z^k\|^2|\mathscr{F}_k]\le 4\tau_k^2\prod_{i=1}^T\hat{\sigma}_{J_i}^2.$$

The proof is completed by combing all above observations with the expression of $\mathbf{R}_k$ in (B.15). □

Proof of Theorem 4.1. We now present the proof of Theorem 4.1. Note that by Lemma B.6 and given values of α,γ in (4.12), we obtain

$$\begin{aligned}
&\sum_{k=1}^N\tau_k\left[\beta\left(\|d^k\|^2+\sum_{i=1}^T\|f_i(u_{i+1}^k)-u_i^k\|^2\right)+\frac{\beta}{20L_{\nabla F}^2}\|\nabla F(x^k)-z^k\|^2\right]\\
&\le 2W_{\alpha,\gamma}(x^0,z^0,u^0)+2\sum_{k=0}^N\mathbf{R}_k+\left(\frac{24}{5}+\frac{40L_{\nabla F}^2}{\beta^2}\right)\sum_{k=0}^N\tau_k\left(H_k(\tilde{y}^k)-H_k(y^k)\right),
\end{aligned}$$

Taking expectation of both sides and noting that $\mathbb{E}[\mathbf{R}_k|\mathscr{F}_k]\le\hat{\sigma}^2\tau_k^2$ by Proposition B.1, we have

$$\begin{aligned}
&\sum_{k=1}^N\tau_k\mathbb{E}\left[\rho\left(\|d^k\|^2+\sum_{i=1}^T\|f_i(u_{i+1}^k)-u_i^k\|^2\right)+\alpha\|\nabla F(x^k)-z^k\|^2\middle|\mathscr{F}_{k-1}\right]\\
&\le 2W_{\alpha,\gamma}(x^0,z^0,u^0)+2\hat{\sigma}^2\sum_{k=0}^N\tau_k^2+\left(\frac{24}{5}+\frac{40L_{\nabla F}^2}{\beta^2}\right)\sum_{k=0}^N\tau_k\left(H_k(\tilde{y}^k)-H_k(y^k)\right).
\end{aligned}\tag{B.27}$$

Then, setting τ_k,t_k to be values in (4.11) and noting that by Lemma B.3, we have

$$H_k(\tilde{y}^k)-H_k(y^k)\le\frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{t_k+2}\le\frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{\sqrt{k}},\quad\forall k\ge 1.$$

Also, with the choice of $z^0=0$, we have $y^0=\tilde{y}^0=x^0$. Thus, we can conclude that

$$\sum_{k=0}^N\tau_k\left(H_k(\tilde{y}^k)-H_k(y^k)\right)\le\frac{2\beta D_{\mathcal{X}}^2(1+\delta)}{\sqrt{N}}\sum_{k=1}^N\frac{1}{\sqrt{k}}\le 4\beta D_{\mathcal{X}}^2(1+\delta).$$

which together with (B.27) immediately imply that $\forall N \geq 1$,

$$\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\mathbb{E}\left[\beta\left(\|d^k\|^2+\sum_{j=1}^{T}\|f_j(u_{j+1}^k)-u_j^k\|^2\right)+\frac{\beta}{20L_{\nabla F}^2}\|\nabla F(x^k)-z^k\|^2\middle|\mathscr{F}_{k-1}\right]$$

$$\leq 2W_{\alpha,\gamma}(x^0,z^0,u^0)+\mathcal{B}(\beta,\sigma^2,L,D_{\mathcal{X}},T,\delta).$$

where

$$\mathcal{B}(\beta,\sigma^2,L,D_{\mathcal{X}},T,\delta)=4\hat{\sigma}^2+32\beta D_{\mathcal{X}}^2(1+\delta)\left(\frac{3}{5}+\frac{5L_{\nabla F}^2}{\beta^2}\right),$$

and $\hat{\sigma}^2$ is given in (B.26). As a result, we can obtain (4.13) and (4.14) by the definition of random integer R and

$$\|\mathcal{G}(x^k,\nabla F(x^k),\beta)\|^2 \leq 2\beta^2\|d^k\|^2+2\beta^2\left\|\mathbf{proj}_{\mathcal{X}}\left(x^k-\frac{1}{\beta}\nabla F(x^k)\right)-\mathbf{proj}_{\mathcal{X}}\left(x^k-\frac{1}{\beta}z^k\right)\right\|^2$$

$$\leq 2\beta^2\|d^k\|^2+2\|\nabla F(x^k)-z^k\|^2.$$

□

B.3. Proofs for Section 4.3.1

B.3.1. Proof of Theorem 4.2 for $T=2$. To show the rate of convergence for Algorithm 7, we simplify the merit function in the analysis of the multi-level problems and leverage the following function:

$$W_{\alpha,\gamma}(x^k,z^k,u^k)=F(x^k)-F^\star-\eta(x^k,z^k)+\alpha\|\nabla F(x^k)-z^k\|^2+\gamma\|f_2(x^k)-u_2^k\|^2, \tag{B.28}$$

where α,γ are positive constants, $\eta(\cdot,\cdot)$ is defined in (4.10). We now present the analogue of Lemma B.6 for Algorithm 7. The proof follows similar steps as that proof of Lemma B.6 with slight modifications, and hence we will skip some arguments already presented before.

Lemma B.7. *Let $\{x^k,z^k,u_2^k\}_{k\geq 0}$ be the sequence generated by Algorithm 7 and the merit function $W_{\alpha,\gamma}(\cdot,\cdot,\cdot)$ be defined in* (B.28) *with*

$$\alpha=\frac{\rho}{L_{\nabla F}},\qquad \gamma=3\rho L_{\nabla f_1},\qquad \rho>0.$$

Under Assumptions 4.2 with $T=2$, setting $\beta_k \equiv \beta \geq 6\rho L_{\nabla F} + (2\rho + \frac{2}{3\rho})L_{\nabla f_1} L_{f_2}^2$, we have $\forall N \geq 0$

$$\rho \sum_{k=0}^{N} \tau_k \Big(L_{\nabla F}\|d^k\|^2 + L_{\nabla f_1}\|f_2(x^k) - u_2^k\|^2 + \frac{1}{L_{\nabla F}}\|\nabla F(x^k) - z^k\|^2 \Big)$$

$$\leq 2W_0 + 2\sum_{k=0}^{N} \mathbf{R}_k + \left(4 + \frac{2(8\rho + 1/\rho)L_{\nabla F} + 24\rho L_{\nabla f_1} L_{f_2}^2}{\beta}\right) \sum_{k=0}^{N} \tau_k \left(H_k(\tilde{y}^k) - H_k(y^k) \right)$$

where $d^k = y^k - x^k$, $H_k(\cdot), y^k$ are defined in (4.16), and

$$\begin{aligned}
\mathbf{R}_k :=& \tau_k^2 \left[\frac{L_{\nabla F} + L_{\nabla \eta}}{2} D_{\mathcal{X}}^2 + \gamma \|\Delta_{G_2}^{k+1}\|^2 + \alpha \|\Delta^{k+1}\|^2 \right] + \frac{L_\eta}{2}\|z^{k+1} - z^k\|^2 \\
&+ \tau_k \langle d^k, \Delta^{k+1} \rangle + \gamma \dot{r}^{k+1} + \alpha \ddot{r}^{k+1}, \\
\Delta^{k+1} :=& \nabla f_2(x^k)\nabla f_1(u_2^k) - J_2^{k+1} J_1^{k+1}, \quad \Delta_{G_2}^{k+1} := f_2(x^k) - G_2^{k+1} \\
\dot{r}^{k+1} :=& 2\tau_k \langle \Delta_{G_2}^{k+1}, f_2(x^{k+1}) - f_2(x^k) + (1 - \tau_k)(f_2(x^k) - u_2^k) \rangle, \\
\ddot{r}^{k+1} :=& 2\tau_k \langle \Delta^{k+1}, (1 - \tau_k)[\nabla F(x^k) - z^k] + \tau_k[\nabla F(x^k) - \nabla f_2(x^k)\nabla f_1(u^k)] \\
&+ \nabla F(x^{k+1}) - \nabla F(x^k) \rangle.
\end{aligned} \tag{B.29}$$

Proof of Lemma B.7. 1. By the Lipschitzness of ∇F (Lemma B.1), we have

$$\begin{aligned}
F(x^{k+1}) - F(x^k) \leq& \tau_k \langle \nabla F(x^k), d^k \rangle + \tau_k \|\nabla F(x^k) - z^k\| \|\tilde{y}^k - y^k\| \\
&+ \tau_k \beta \|d^k\| \|\tilde{y}^k - y^k\| + \tau_k \langle z^k + \beta d^k, \tilde{y}^k - y^k \rangle + \frac{L_{\nabla F}\tau_k^2 \|\tilde{d}^k\|^2}{2}.
\end{aligned} \tag{B.30}$$

2. Also, by the Lipschitzness of $\nabla \eta$ (Lemma B.2) and the optimality condition of in the definition of y^k, we have

$$\begin{aligned}
\eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1}) \leq& -\beta \tau_k \|d^k\|^2 + \tau_k \big\langle z^k + \beta d^k, \tilde{y}^k - y^k \big\rangle \\
&- \tau_k \big\langle d^k, \nabla f_2(x^k)\nabla f_1(u_2^k) \big\rangle + \tau_k \big\langle d^k, \Delta^{k+1} \big\rangle + \frac{L_{\nabla \eta}}{2}\Big[\tau_k^2 \|\tilde{d}^k\|^2 + \|z^{k+1} - z^k\|^2 \Big].
\end{aligned} \tag{B.31}$$

3. In addition, by the Lipschitzness of f_2 and ∇f_1, we have

$$\begin{aligned}
\langle d^k, \nabla F(x^k) - \nabla f_2(x^k)\nabla f_1(u_2^k) \rangle =& \big\langle d^k, \nabla f_2(x^k)^\top \big[\nabla f_1(f_2(x^k)) - \nabla f_1(u_2^k) \big] \big\rangle \\
\leq& L_{\nabla f_1} L_{f_2} \|d^k\| \|f_2(x^k) - u_2^k\|.
\end{aligned} \tag{B.32}$$

4. Moreover, by the update rule, we have

$$\begin{aligned}\|f_2(x^{k+1}) - u_2^{k+1}\|^2 &= \|f_2(x^{k+1}) - f_2(x^k) + (1-\tau_k)[f_2(x^k) - u_2^k] + \tau_k \Delta_{G_2}^{k+1}\|^2 \\ &= \|(1-\tau_k)[f_2(x^k) - u_2^k] + f_2(x^{k+1}) - f_2(x^k)\|^2 + \tau_k^2\|\Delta_{G_2}^{k+1}\|^2 + \dot{r}^{k+1} \\ &\le (1-\tau_k)\|f_2(x^k) - u_2^k\|^2 + 2\tau_k L_{f_2}^2(\|d^k\|^2 + \|\tilde{y}^k - y^k\|^2) + \tau_k^2\|\Delta_{G_2}^{k+1}\|^2 + \dot{r}^{k+1}\end{aligned} \tag{B.33}$$

where $\dot{r}^{k+1} := 2\tau_k\langle \Delta_{G_2}^{k+1}, f_2(x^{k+1}) - f_2(x^k) + (1-\tau_k)(f_2(x^k) - u_2^k)\rangle$ and the last inequality follows Jensen's inequality for the convex function $\|\cdot\|^2$ as well as

$$\left\|\frac{1}{\tau_k}\left[f_2(x^{k+1}) - f_2(x^k)\right]\right\|^2 \le L_{f_2}^2\|\tilde{d}^k\|^2 \le 2L_{f_2}^2(\|d^k\|^2 + \|\tilde{y}^k - y^k\|^2).$$

5. Defining

$$e^k := \frac{1}{\tau_k}\Big[\nabla F(x^{k+1}) - \nabla F(x^k)\Big] + \nabla F(x^k) - \nabla f_2(x^k)\nabla f_1(u^k),$$

and by the update rule, we have

$$\begin{aligned}\|\nabla F(x^{k+1}) - z^{k+1}\|^2 &= \|(1-\tau_k)[\nabla F(x^k) - z^k] + \tau_k[e^k + \Delta^{k+1}]\|^2 \\ &= \|(1-\tau_k)[\nabla F(x^k) - z^k] + \tau_k e^k\|^2 + \tau_k^2\|\Delta^{k+1}\|^2 + \ddot{r}^{k+1} \\ &\le (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + \tau_k\|e^k\|^2 + \tau_k^2\|\Delta^{k+1}\|^2 + \ddot{r}^{k+1}\end{aligned} \tag{B.34}$$

where $\ddot{r}^{k+1} := 2\tau_k\langle\Delta^k, (1-\tau_k)[\nabla F(x^k) - z^k] + \tau_k e^k\rangle$. We can further upper bound the term $\|e^k\|^2$ by

$$\begin{aligned}\|e^k\|^2 &\le 2L_{\nabla F}^2\|\tilde{d}^k\|^2 + 2L_{\nabla f_1}^2 L_{f_2}^2\|f_2(x^k) - u^k\|^2 \\ &\le 4L_{\nabla F}^2(\|d^k\|^2 + \|\tilde{y}^k - y^k\|^2) + 2L_{\nabla f_1}^2 L_{f_2}^2\|f_2(x^k) - u^k\|^2\end{aligned} \tag{B.35}$$

6. By combing (B.30), (B.31), (B.32), (B.33), (B.34), (B.35), rearranging the terms, and noting that $\langle z^k + \beta d^k, \tilde{y}^k - y^k\rangle = H_k(\tilde{y}^k) - H_k(y^k) - (\beta/2)\|\tilde{y}^k - y^k\|^2$ and $\|\tilde{d}^k\| \le D_{\mathcal{X}}$, we obtain

$$W_{k+1} - W_k \le \tau_k \mathbf{A}_k + \mathbf{R}_k \tag{B.36}$$

where $\mathbf{R}_k$ is defined in (B.29) and

$$\begin{aligned}\mathbf{A}_k := &\left(-\beta + 4\alpha L_{\nabla F}^2 + 2\gamma L_{f_2}^2\right)\|d^k\|^2 + \left(-\gamma + 2\alpha L_{\nabla f_1}^2 L_{f_2}^2\right)\|f_2(x^k) - u_2^k\|^2 \\ &+ L_{\nabla f_1} L_{f_2}\|d^k\|\|f_2(x^k) - u_2^k\| - \alpha\|\nabla F(x^k) - z^k\|^2 \\ &+ \Big(\beta\|d^k\| + \|\nabla F(x^k) - z^k\|\Big)\|\tilde{y}^k - y^k\| \\ &+ \left(4\alpha L_{\nabla F}^2 + 2\gamma L_{f_2}^2 - \beta\right)\|\tilde{y}^k - y^k\|^2 + 2\Big(H_k(\tilde{y}^k) - H_k(y^k)\Big).\end{aligned}$$

We then provide a simplified upper bound for $\mathbf{A}_k$. By the Young's inequality, we have

$$\begin{aligned}\beta\|d^k\|\|\tilde{y}^k - y^k\| &\le \frac{\beta}{4}\|d^k\|^2 + \beta\|\tilde{y}^k - y^k\|^2, \\ \|\nabla F(x^k) - z^k\|\|\tilde{y}^k - y^k\| &\le \frac{\alpha}{2}\|\nabla F(x^k) - z^k\|^2 + \frac{1}{2\alpha}\|\tilde{y}^k - y^k\|^2.\end{aligned}$$

In addition, we reparametrize $\alpha = \frac{\rho}{L_{\nabla F}}$. Noting that by Lemma B.1 with $T = 2$

$$\frac{L_{\nabla f_1}^2 L_{f_2}^2}{L_{\nabla F}} = \frac{L_{\nabla f_1}^2 L_{f_2}^2}{L_{\nabla f_1} L_{f_2}^2 + L_{f_1} L_{\nabla f_2}} \le L_{\nabla f_1},$$

we therefore have

$$\begin{aligned}\mathbf{A}_k \le &\left(-\frac{3\beta}{4} + 4\rho L_{\nabla F} + 2\gamma L_{f_2}^2\right)\|d^k\|^2 + \left(-\gamma + 2\rho L_{\nabla f_1}\right)\|f_2(x^k) - u_2^k\|^2 \\ &+ L_{\nabla f_1} L_{f_2}\|d^k\|\|f_2(x^k) - u_2^k\| - \frac{\rho}{2L_{\nabla F}}\|\nabla F(x^k) - z^k\|^2 \\ &+ \left(4\rho L_{\nabla F} + 2\gamma L_{f_2}^2 + \frac{L_{\nabla F}}{2\rho}\right)\|\tilde{y}^k - y^k\|^2 + 2\Big(H_k(\tilde{y}^k) - H_k(y^k)\Big)\end{aligned}$$

Then, setting $\gamma = 3\rho L_{\nabla f_1}$ and $\beta \ge 6\rho L_{\nabla F} + (2\rho + \frac{2}{3\rho})L_{\nabla f_1} L_{f_2}^2$, we can obtain

$$\begin{aligned}&\left(-\frac{3\beta}{4} + 4\rho L_{\nabla F} + 2\gamma L_{f_2}^2\right)\|d^k\|^2 + \left(-\gamma + 2\rho L_{\nabla f_1}\right)\|f_2(x^k) - u_2^k\|^2 \\ &\quad + L_{\nabla f_1} L_{f_2}\|d^k\|\|f_2(x^k) - u_2^k\| \le -\frac{\rho L_{\nabla F}}{2}\|d^k\|^2 - \frac{\rho L_{\nabla f_1}}{2}\|f_2(x^k) - u_2^k\|^2\end{aligned}$$

Also, we have $(\beta/2)\|\tilde{y}^k - y^k\|^2 \le H_k(\tilde{y}^k) - H_k(y^k)$. Therefore, we can further bound $\mathbf{A}_k$ by

$$\begin{aligned}\mathbf{A}_k \le &-\frac{\rho L_{\nabla F}}{2}\|d^k\|^2 - \frac{\rho L_{\nabla f_1}}{2}\|G(x^k) - u^k\|^2 - \frac{\rho}{2L_{\nabla F}}\|\nabla F(x^k) - z^k\|^2 \\ &+ \left(2 + \frac{(8\rho + 1/\rho)L_{\nabla F} + 12\rho L_{\nabla f_1} L_{f_2}^2}{\beta}\right)\Big(H_k(\tilde{y}^k) - H_k(y^k)\Big).\end{aligned} \tag{B.37}$$

Telescoping (B.36) together with (B.37), we get

$$\rho\sum_{k=0}^{N}\tau_k\Big(L_{\nabla F}\|d^k\|^2 + L_{\nabla f_1}\|f_2(x^k) - u_2^k\|^2 + \frac{1}{L_{\nabla F}}\|\nabla F(x^k) - z^k\|^2\Big)$$

$$\leq 2W_0 + 2\sum_{k=0}^{N}\mathbf{R}_k + \left(4 + \frac{2(8\rho + 1/\rho)L_{\nabla F} + 24\rho L_{\nabla f_1}L_{f_2}^2}{\beta}\right)\sum_{k=0}^{N}\tau_k\left(H_k(\tilde{y}^k) - H_k(y^k)\right)$$

□

Proof of Theorem 4.2, part (a). The proof follows the same arguments in the proof of Theorem 4.1. Note that by Lemma B.7 and given values of α, γ in (4.12), we obtain

$$\rho\sum_{k=1}^{N}\tau_k\left[L_{\nabla F}\|d^k\|^2 + L_{\nabla f_1}\|f_2(x^k) - u_2^k\|^2 + \frac{1}{L_{\nabla F}}\|\nabla F(x^k) - z^k\|^2\right] \leq 2W_{\alpha,\gamma}(x^0, z^0, u^0)$$

$$+ 2\sum_{k=0}^{N}\mathbf{R}_k + \left(4 + \frac{2(8\rho + 1/\rho)L_{\nabla F} + 24\rho L_{\nabla f_1}L_{f_2}^2}{\beta}\right)\sum_{k=0}^{N}\tau_k\left(H_k(\tilde{y}^k) - H_k(y^k)\right).$$

Noting that

$$\mathbb{E}[\mathbf{R}_k|\mathscr{F}_k] = \tau_k^2\left[\frac{L_{\nabla F} + L_{\nabla\eta}}{2}D_{\mathcal{X}}^2 + \gamma\sigma_{G_2}^2 + (\alpha + 2L_\eta)\hat{\sigma}_{J_1}^2\hat{\sigma}_{J_2}^2\right] := \tau_k^2\hat{\sigma}^2,$$

and taking expectation of both sides, we can complete the proof with the same arguments in the proof of Theorem 4.1. The constants $\mathcal{C}_1$ and $\mathcal{C}_2$ turn out to be

$$\begin{aligned}\mathcal{C}_1 &= 4\left(\frac{\beta^2}{\rho L_{\nabla F}} + \frac{L_{\nabla F}}{\rho}\right)\Big\{W_{\alpha,\gamma}(x^0, z^0, u^0) + \hat{\sigma}^2 \\ &\qquad + 4D_{\mathcal{X}}^2(1+\delta)\left[2\beta + (8\rho + \frac{1}{\rho})L_{\nabla F} + 12\rho L_{\nabla f_1}L_{f_2}^2)\right]\Big\}, \\ \mathcal{C}_2 &= \frac{2}{\rho L_{\nabla f_1}}\Big\{W_{\alpha,\gamma}(x^0, z^0, u^0) + \hat{\sigma}^2 \\ &\qquad + 4D_{\mathcal{X}}^2(1+\delta)\left[2\beta + (8\rho + \frac{1}{\rho})L_{\nabla F} + 12\rho L_{\nabla f_1}L_{f_2}^2)\right]\Big\}.\end{aligned} \tag{B.38}$$

□

B.3.2. Proof of Theorem 4.2 for $T = 1$. To show the rate of convergence for Algorithm 8, we leverage the following merit function:

$$W_\alpha(x^k, z^k, u^k) = F(x^k) - F^\star - \eta(x^k, z^k) + \alpha\|\nabla F(x^k) - z^k\|^2, \tag{B.39}$$

where $\alpha > 0$, $\eta(\cdot,\cdot)$ is defined in (4.10).

LEMMA B.8. *Let $\{x^k, z^k\}_{k\geq 0}$ be the sequence generated by Algorithm 8 with $\beta_k \equiv \beta > 0$ and the merit function $W_\alpha(\cdot,\cdot)$ be defined in* (B.39) *with $\alpha = \frac{\beta}{4L^2_{\nabla F}}$. Under Assumptions 4.2 with $T = 1$, we have $\forall N \geq 0$*

$$\beta \sum_{k=0}^{N} \tau_k \left(\|d^k\|^2 + \frac{1}{2L^2_{\nabla F}} \|\nabla F(x^k) - z^k\|^2 \right)$$
$$\leq 4W_\alpha(x^0, u^0) + 4\sum_{k=0}^{N} \mathbf{R}_k + \left(12 + \frac{16L^2_{\nabla F}}{\beta^2}\right) \sum_{k=0}^{N} \tau_k \left(H_k(\tilde{y}^k) - H_k(y^k)\right)$$

where $d^k := y^k - x^k$, $H_k(\cdot), y^k$ are defined in (4.16), *$\Delta^{k+1} := \nabla F(x^k) - J_1^{k+1}$, and*

$$\begin{aligned} \mathbf{R}_k :=& \tau_k^2 \left[\frac{L_{\nabla F} + L_{\nabla\eta}}{2} D^2_{\mathcal{X}} + \alpha \|\Delta^{k+1}\|^2 \right] + \frac{L_\eta}{2} \|z^{k+1} - z^k\|^2 \\ &+ \tau_k \langle d^k, \Delta^{k+1} \rangle + \alpha r^{k+1}, \\ r^{k+1} :=& 2\tau_k \langle \Delta^{k+1}, (1 - \tau_k)[\nabla F(x^k) - z^k] + \nabla F(x^{k+1}) - \nabla F(x^k) \rangle. \end{aligned} \tag{B.40}$$

PROOF. The proof is a essentially a simplified version of the proof of Lemma B.7. Hence, we skip some arguments already presented earlier.

1. By the Lipschitzness of ∇F, we have

$$\begin{aligned} F(x^{k+1}) - F(x^k) \leq& \tau_k \langle \nabla F(x^k), d^k \rangle + \tau_k \|\nabla F(x^k) - z^k\| \|\tilde{y}^k - y^k\| \\ &+ \tau_k \beta \|d^k\| \|\tilde{y}^k - y^k\| + \tau_k \langle z^k + \beta d^k, \tilde{y}^k - y^k \rangle + \frac{L_{\nabla F} \tau_k^2 \|\tilde{d}^k\|^2}{2}. \end{aligned} \tag{B.41}$$

2. Also, by the lipschitzness of $\nabla\eta$ (Lemma B.2) and the optimality condition of in the definition of y^k, we have

$$\begin{aligned} \eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1}) \leq& -\beta\tau_k \|d^k\|^2 + \tau_k \big\langle z^k + \beta d^k, \tilde{y}^k - y^k \big\rangle \\ &- \tau_k \big\langle d^k, \nabla F(x^k) \big\rangle + \tau_k \big\langle d^k, \Delta^{k+1} \big\rangle + \frac{L_{\nabla\eta}}{2} \Big[\tau_k^2 \|\tilde{d}^k\|^2 + \|z^{k+1} - z^k\|^2 \Big]. \end{aligned} \tag{B.42}$$

3. By the update rule, we have

$$\begin{aligned}
\|\nabla F(x^{k+1}) - z^{k+1}\|^2 &= \|(1-\tau_k)[\nabla F(x^k) - z^k] + \nabla F(x^{k+1}) - \nabla F(x^k) + \tau_k \Delta^{k+1}\|^2 \\
&= \|(1-\tau_k)[\nabla F(x^k) - z^k] + \nabla F(x^{k+1}) - \nabla F(x^k)\|^2 + \tau_k^2 \|\Delta^{k+1}\|^2 + r^{k+1} \\
&\leq (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + \frac{1}{\tau_k}\|\nabla F(x^{k+1}) - \nabla F(x^k)\|^2 + \tau_k^2 \|\Delta^{k+1}\|^2 + r^{k+1} \qquad \text{(B.43)} \\
&\leq (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + \tau_k L_{\nabla F}^2 \|\tilde{d}^k\|^2 + \tau_k^2 \|\Delta^{k+1}\|^2 + r^{k+1} \\
&\leq (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + 2\tau_k L_{\nabla F}^2 (\|d^k\|^2 + \|\tilde{y}^k - y^k\|^2) + \tau_k^2 \|\Delta^{k+1}\|^2 + r^{k+1}
\end{aligned}$$

where $r^{k+1} := 2\tau_k \langle \Delta^k, (1-\tau_k)[\nabla F(x^k) - z^k] + \nabla F(x^{k+1}) - \nabla F(x^k)\rangle$.

4. By combing (B.41), (B.42) (B.43), rearranging the terms, and noting that $\langle z^k + \beta d^k, \tilde{y}^k - y^k\rangle = H_k(\tilde{y}^k) - H_k(y^k) - (\beta/2)\|\tilde{y}^k - y^k\|^2$ and $\|\tilde{d}^k\| \leq D_{\mathcal{X}}$, we obtain

$$W_{k+1} - W_k \leq \tau_k \mathbf{A}_k + \mathbf{R}_k \tag{B.44}$$

where $\mathbf{R}_k$ is defined in (B.40) and

$$\begin{aligned}
\mathbf{A}_k := &\left(-\beta + 2\alpha L_{\nabla F}^2\right)\|d^k\|^2 - \alpha\|\nabla F(x^k) - z^k\|^2 + \Big(\beta\|d^k\| + \|\nabla F(x^k) - z^k\|\Big)\|\tilde{y}^k - y^k\| \\
&+ \left(2\alpha L_{\nabla F}^2 - \beta\right)\|\tilde{y}^k - y^k\|^2 + 2\Big(H_k(\tilde{y}^k) - H_k(y^k)\Big).
\end{aligned}$$

We then provide a simplified upper bound for $\mathbf{A}_k$. By the Young's inequality, we have

$$\begin{gathered}
\beta\|d^k\|\|\tilde{y}^k - y^k\| \leq \frac{\beta}{4}\|d^k\|^2 + \beta\|\tilde{y}^k - y^k\|^2, \\
\|\nabla F(x^k) - z^k\|\|\tilde{y}^k - y^k\| \leq \frac{\alpha}{2}\|\nabla F(x^k) - z^k\|^2 + \frac{1}{2\alpha}\|\tilde{y}^k - y^k\|^2.
\end{gathered}$$

In addition, setting $\alpha = \frac{\beta}{4L_{\nabla F}^2}$ and noting $(\beta/2)\|\tilde{y}^k - y^k\|^2 \leq H_k(\tilde{y}^k) - H_k(y^k)$, we have

$$\mathbf{A}_k \leq -\frac{\beta}{4}\|d^k\|^2 - \frac{\beta}{8L_{\nabla F}^2}\|\nabla F(x^k) - z^k\|^2 + \left(3 + \frac{4L_{\nabla F}^2}{\beta^2}\right)\Big(H_k(\tilde{y}^k) - H_k(y^k)\Big) \tag{B.45}$$

Telescoping (B.44) together with (B.45), we get

$$\begin{aligned}
&\beta \sum_{k=0}^{N} \tau_k \Big(\|d^k\|^2 + \frac{1}{2L_{\nabla F}^2} \|\nabla F(x^k) - z^k\|^2 \Big) \\
&\le 4W_\alpha(x^0, u^0) + 4\sum_{k=0}^{N} \mathbf{R}_k + \left(12 + \frac{16L_{\nabla F}^2}{\beta^2}\right) \sum_{k=0}^{N} \tau_k \left(H_k(\tilde{y}^k) - H_k(y^k) \right)
\end{aligned}$$

□

Proof of Theorem 4.2, part (b). Given Lemma B.8, the proof follows the same arguments as in the proof of Theorem 4.1. The constant $\mathcal{C}_3$ turns out to be

$$\begin{aligned}
\mathcal{C}_3 = 8\left(\beta + \frac{2L_{\nabla F}^2}{\beta}\right) \Big\{ W_\alpha(x^0, u^0) + D_{\mathcal{X}}^2 \left[(1+\delta)\left(12\beta + \frac{16L_{\nabla F}^2}{\beta}\right) + \frac{L_{\nabla F} + L_{\nabla \eta}}{2} \right] \\
+ \alpha\sigma_{J_1}^2 + 2L_\eta \hat{\sigma}_{J_1}^2 \Big\}.
\end{aligned} \tag{B.46}$$

□

B.4. Proof of Theorem 4.3

We start with presenting the lemma below which leverages inequalities in Appendix B.4 to show a high-probability upper bound for terms involving in the previous analysis.

Lemma B.9. *Under the conditions of Lemma B.8 and Assumption 4.4, for any $\delta_1, \delta_2, \delta_3, a > 0$, we have*

(a) with probability at least $1-\delta_1$, $\sum_{k=0}^{N} \tau_k^2 \|\Delta^{k+1}\|^2 \le K^2 \log(2/\delta_1) \sum_{k=0}^{N} \tau_k^2$;

(b) with probability at least $1-\delta_2$,

$$\sum_{k=0}^{N} \tau_k^2 \sum_{i=0}^{k-1} \alpha_{i,k} \|\Delta^{i+1}\|^2 \le K^2 \log(2/\delta_2) \sum_{k=0}^{N} \tau_k^2,$$

where $\alpha_{i,k} > 0$ and $\sum_{i=0}^{k-1} \alpha_{i,k} = 1$;

(c) with probability at least $1-\delta_3$,

$$\begin{aligned}
&\sum_{k=0}^{N} \langle \Delta^{k+1}, 2\alpha\tau_k(1-\tau_k)[\nabla F(x^k) - z^k] + 2\alpha\tau_k(\nabla F(x^{k+1}) - \nabla F(x^k)) + \tau_k d^k \rangle \\
&\le 4a\log(1/\delta_3) + \frac{\beta^2 K^2}{aL_{\nabla F}^4} \sum_{k=0}^{N} \tau_k^2 (1-\tau_k) \left\| \nabla F(x^k) - z^k \right\|^2 + \frac{K^2}{a} \sum_{k=0}^{N} \tau_k^2 \left(4 + \frac{\beta^2\tau_k}{L_{\nabla F}^2}\right) \left\| d^k \right\|^2.
\end{aligned}$$

Proof of Lemma B.9. We first show (a). Using the law of total expectation, we have $\mathbb{E}\left[\exp\left(\frac{\|\tau_k \Delta^{k+1}\|^2}{\tau_k^2 K^2}\right)\right] \leq 2$, which implies that $\|\tau_k \Delta^{k+1}\|^2$ is $\tau_k^2 K^2$-sub-exponential. Thus, we have with probability at least $1 - \delta_1$,

$$\sum_{k=0}^{N} \tau_k^2 \|\Delta^{k+1}\|^2 \leq K^2 \log(2/\delta_1) \sum_{k=0}^{N} \tau_k^2. \tag{B.47}$$

We then show (b). Let $Z_k = \tau_k^2 \left\{\sum_{i=0}^{k-1} \alpha_{i,k} \left\|\Delta^{i+1}\right\|^2\right\} \forall k \geq 0$. Note that for all $k \geq 0$, $\|\Delta^{k+1}\|^2$ is K^2-sub-exponential, which further implies that the sub-exponential norm of Z_k ($k > 0$) satisfies $\|Z_k\|_{\psi_1} \leq \tau_k^2 K^2$. Therefore, we have for any $\delta_2 > 0$, with probability at least $1 - \delta_2$,

$$\sum_{k=0}^{N} Z_k \leq K^2 \log(2/\delta_2) \sum_{k=0}^{N} \tau_k^2. \tag{B.48}$$

To prove (c), we apply Lemma 1.1 and Lemma 1.3 with

$$\begin{aligned}
&X_i = \left\langle \Delta^{k+1}, 2\alpha\tau_k \left\{(1-\tau_k)[\nabla F(x^k) - z^k] + \nabla F(x^{k+1}) - \nabla F(x^k)\right\} + \tau_k d^k \right\rangle, \\
&K_i = \sqrt{c} K \left\| 2\alpha\tau_k \left\{(1-\tau_k)[\nabla F(x^k) - z^k] + \nabla F(x^{k+1}) - \nabla F(x^k)\right\} + \tau_k d^k \right\|, \\
&\quad b = 0, t = 4a \log(1/\delta_3).
\end{aligned}$$

Noting that $\alpha = \frac{\beta}{4L_{\nabla F}^2}$, we obtain that for all $a > 0$ with probability at least $1 - \delta_3$, $\sum_{i=0}^{N} X_i \leq 4a \log(1/\delta_3)$ and

$$\begin{aligned}
\sum_{i=0}^{N} X_i &\leq \frac{2cK^2}{a} \sum_{k=0}^{N} \left\| 2\alpha\tau_k \left\{(1-\tau_k)[\nabla F(x^k) - z^k] + \nabla F(x^{k+1}) - \nabla F(x^k)\right\} + \tau_k d^k \right\|^2 \\
&\leq \frac{4cK^2}{a} \sum_{k=0}^{N} \tau_k^2 \left\{ 4\alpha^2 \left\| (1-\tau_k)[\nabla F(x^k) - z^k] + \nabla F(x^{k+1}) - \nabla F(x^k) \right\|^2 + \left\| d^k \right\|^2 \right\} \\
&\leq \frac{4cK^2}{a} \sum_{k=0}^{N} \tau_k^2 \left\{ 4\alpha^2 (1-\tau_k) \left\| \nabla F(x^k) - z^k \right\|^2 + (1 + 4\alpha^2 L_{\nabla F}^2 \tau_k) \left\| d^k \right\|^2 \right\} \\
&= \frac{c\beta^2 K^2}{a L_{\nabla F}^4} \sum_{k=0}^{N} \tau_k^2 (1-\tau_k) \left\| \nabla F(x^k) - z^k \right\|^2 + \frac{cK^2}{a} \sum_{k=0}^{N} \tau_k^2 (4 + \frac{\beta^2 \tau_k}{L_{\nabla F}^2}) \left\| d^k \right\|^2,
\end{aligned}$$

where the third inequality comes from the convexity of $\|\cdot\|^2$ and the Lipschitzness of ∇F. □

Provided with the above lemma and Lemma B.8, we now present the complete proof of Theorem 4.3.

Proof of Theorem 4.3. Given the update rule of $\{z^k\}$ and the fact that $\tau_0 = 1$, we can obtain

$$z^k = \sum_{i=0}^{k-1} \alpha_{i,k} J_1^{i+1}, \quad \alpha_{i,k} = \frac{\tau_i}{\Gamma_{i+1}} \Gamma_k \ 1 \leq i \leq k, \quad \sum_{i=0}^{k-1} \alpha_{i,k} = 1 \ k \geq 1$$

where $\Gamma_k = \Gamma_1 \prod_{i=1}^{k-1}(1-\tau_i)$ and $\Gamma_1 = 1$. Thus,

$$\begin{aligned}
\left\|z^{k+1} - z^k\right\|^2 = \tau_k^2 \left\|J_1^{k+1} - z^k\right\|^2 &\leq 2\tau_k^2 \left\{ \left\|J_1^{k+1}\right\|^2 + \left\|\sum_{i=0}^{k-1} \alpha_{i,k} J_1^{i+1}\right\|^2 \right\} \\
&\leq 2\tau_k^2 \left\{ \left\|J_1^{k+1}\right\|^2 + \sum_{i=0}^{k-1} \alpha_{i,k} \left\|J_1^{i+1}\right\|^2 \right\} \\
&\leq 4\tau_k^2 \left\{ \left\|\Delta^{k+1}\right\|^2 + \left\|\nabla F(x^k)\right\|^2 + \sum_{i=0}^{k-1} \alpha_{i,k} \left[\left\|\Delta^{i+1}\right\|^2 + \|\nabla F(x^i)\|^2 \right] \right\} \\
&\leq 4\tau_k^2 \left\{ \left\|\Delta^{k+1}\right\|^2 + \sum_{i=0}^{k-1} \alpha_{i,k} \left\|\Delta^{i+1}\right\|^2 + 2L_F^2 \right\}
\end{aligned}$$

where the second inequality comes from the convexity of $\|\cdot\|^2$. Therefore, we have

$$\sum_{k=0}^{N} \|z^{k+1} - z^k\|^2 \leq 4\sum_{k=0}^{N} \tau_k^2 \|\Delta^{k+1}\|^2 + 4\sum_{k=0}^{N} \tau_k^2 \sum_{i=0}^{k-1} \alpha_{i,k} \|\Delta^{i+1}\|^2 + 8L_F^2 \sum_{k=0}^{N} \tau_k^2$$

Applying Lemma B.9 with $\delta_1 = \delta_2 = \delta_3 = \delta/3$ and $a = \frac{16c\beta K^2}{L_{\nabla F}^2}$ together with Lemma B.8, we have with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{k=0}^{N} \mathbf{R}_k =& \sum_{k=0}^{N} \langle \Delta^{k+1}, 2\alpha\tau_k(1-\tau_k)[\nabla F(x^k) - z^k] + 2\alpha\tau_k(\nabla F(x^{k+1}) - \nabla F(x^k)) + \tau_k d^k \rangle \\
&+ (\alpha + 2L_\eta) \sum_{k=0}^{N} \tau_k^2 \|\Delta^{k+1}\|^2 + 2L_\eta \sum_{k=0}^{N} \tau_k^2 \sum_{i=0}^{k-1} \alpha_{i,k} \|\Delta^{i+1}\|^2 \\
&+ \left[\frac{L_{\nabla F} + L_{\nabla \eta}}{2} D_{\mathcal{X}}^2 + 4L_\eta L_F^2 \right] \sum_{k=0}^{N} \tau_k^2 \\
\leq& \frac{64\beta K^2}{L_{\nabla F}^2} \log(3/\delta) + \frac{\beta}{16L_{\nabla F}^2} \sum_{k=0}^{N} \tau_k^2 (1-\tau_k) \left\|\nabla F(x^k) - z^k\right\|^2 + (\frac{L_{\nabla F}^2}{4\beta} + \frac{\beta}{16}) \sum_{k=0}^{N} \tau_k^2 \left\|d^k\right\|^2 \\
&+ \left[(\frac{\beta}{4L_{\nabla F}^2} + 4L_\eta) K^2 \log(6/\delta) + \frac{L_{\nabla F} + L_{\nabla \eta}}{2} D_{\mathcal{X}}^2 + 4L_\eta L_F^2 \right] \sum_{k=0}^{N} \tau_k^2
\end{aligned}$$

Thus, noting that $\|d^k\|^2 \leq D_{\mathcal{X}}^2 \ \forall k \geq 0$, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\beta \sum_{k=0}^{N} \tau_k \Bigg(\|d^k\|^2 + \frac{1}{4L_{\nabla F}^2} \|\nabla F(x^k) - z^k\|^2 \Bigg) \\
&\leq 4W_\alpha(x^0, u^0) + \left(12 + \frac{16L_{\nabla F}^2}{\beta^2}\right) \sum_{k=0}^{N} \tau_k \left(H_k(\tilde{y}^k) - H_k(y^k) \right) + \frac{256\beta K^2}{L_{\nabla F}^2} \log(3/\delta) \\
&+ \left[\left(\frac{\beta}{L_{\nabla F}^2} + 16L_\eta \right) K^2 \log(6/\delta) + \left(\frac{L_{\nabla F}^2}{\beta} + \frac{\beta}{4} + 2L_{\nabla F} + 2L_{\nabla \eta} \right) D_{\mathcal{X}}^2 + 16L_\eta L_F^2 \right] \sum_{k=0}^{N} \tau_k^2
\end{aligned}$$

Following the same arguments as in the proof of Theorem 4.1, we can complete the proof.

□

APPENDIX C

Appendix of Chapter 5

C.1. Proof for Theorem 5.1

In order to prove Theorem 5.1, we require the following result from [**NS17**] for the zeroth-order case.

LEMMA C.1. [**NS17**] *Let the function f has lipschitz continuous gradient with constant L. Consider the smoothed function $f_\nu(x) = \mathbb{E}_u[f(x+\nu u)]$ where $u \sim \mathcal{N}(0, \mathbf{I}_d)$. Then for any $x \in \mathbb{R}^d$,*

$$\mathbb{E}_u\left[\frac{f(x+\nu u)-f(x)}{\nu}u\right] = \nabla f_\nu(x) \tag{C.1}$$

$$\|\nabla f_\nu(x) - \nabla f(x)\| \le \frac{\nu}{2}L(d+3)^{\frac{3}{2}} \tag{C.2}$$

$$\frac{1}{\nu^2}\mathbb{E}_u[\{f(x+\nu u)-f(x)\}^2\|u\|^2] \le \frac{\nu^2}{2}L^2(d+6)^3 + 2(d+4)\|\nabla f(x)\|^2. \tag{C.3}$$

We now present the lemma below to bound the mean squared error for the zeroth-order gradient estimator.

LEMMA C.2. *Under Assumption 5.1, 5.2, 5.3, we have*

$$\mathbb{E}\left\|\bar{G}_\nu^t - \nabla f_\nu(x_{t-1})\right\|^2 \le \frac{4\rho L(d+4)(f(x_{t-1})-f(x^*))}{b_t} + \frac{\nu^2L^2(d+6)^3}{2b_t}, \tag{C.4}$$

$$\mathbb{E}\left\|\bar{G}_\nu^t - \nabla f(x_{t-1})\right\|^2 \le \frac{4\rho L(d+4)(f(x_{t-1})-f(x^*))}{b_t} + \nu^2L^2(d+6)^3. \tag{C.5}$$

PROOF. First note that by (C.1), we have

$$\mathbb{E}_{u,\xi}[\bar{G}_\nu^t] = \mathbb{E}_{u,\xi}[G_{t,j}] = \mathbb{E}_u\left[\frac{f(x_{t-1}+\nu u)-f(x_{t-1})}{\nu}u\right] = \nabla f_\nu(x_{t-1}),$$

Then by using (C.3) for F instead of f, under Assumption 5.2, 5.3, we can obtain

$$\begin{aligned}\mathbb{E}_{u,\xi}\left\|\bar{G}_\nu^t - \nabla f_\nu(x_{t-1})\right\|^2 &= \frac{1}{b_t}\mathbb{E}_{u,\xi}\|G_{t,j} - \nabla f_\nu(x_{t-1})\|^2\\ &\le \frac{1}{b_t}\mathbb{E}_{u,\xi}\|G_{t,j}\|^2\end{aligned}$$

$$
\begin{aligned}
&\le \frac{2(d+4)}{b_t}\mathbb{E}_\xi \left\|\nabla F(x_{t-1}, \xi_{t,j})\right\|^2 + \frac{\nu^2 L^2 (d+6)^3}{2b_t} \\
&\le \frac{4\rho L(d+4)(f(x_{t-1}) - f(x^*))}{b_t} + \frac{\nu^2 L^2 (d+6)^3}{2b_t}
\end{aligned}
$$

where the first inequality comes from the fact that the variance is less than the seocond moment.

To prove (C.5), we decompose the mean squared error into the bias and the variance by utilizing the results (C.4) and (C.2), i.e.,

$$
\begin{aligned}
\mathbb{E}\left\|\bar{G}_\nu^t - \nabla f(x_{t-1})\right\|^2 &= \mathbb{E}\left\|\bar{G}_\nu^t - \nabla f_\nu(x_{t-1})\right\|^2 + \left\|\nabla f_\nu(x_{t-1}) - \nabla f(x_{t-1})\right\|^2 \\
&\le \frac{4\rho L(d+4)(f(x_{t-1}) - f(x^*))}{b_t} + \frac{\nu^2 L^2 (d+6)^3}{2b_t} + \frac{\nu^2 L^2 (d+3)^3}{4} \\
&\le \frac{4\rho L(d+4)(f(x_{t-1}) - f(x^*))}{b_t} + \nu^2 L^2 (d+6)^3.
\end{aligned}
$$

□

We also need the following simple result in our proof.

Lemma C.3. *Assume that sequences $\{\phi_t\}_{t\ge 0} \ge 0$, $\{B_t\}_{t\ge 1}$, $\{\theta_t\}_{t\ge 1} \in [0,1]$ are given such that*

$$\phi_t \le (1-\theta_t)\phi_{t-1} + B_t. \tag{C.6}$$

Then, we have

$$\phi_T \le \Theta_T\left[\phi_0 + \sum_{t=1}^T \frac{B_t}{\Theta_t}\right],$$

where, for any $t \ge 2$,

$$\Theta_t = \Theta_1 \prod_{k=2}^t (1-\theta_k), \quad \textit{where} \;\; \Theta_1 = 1-\theta_1 \;\; \textit{if} \;\; \theta_1 < 1, \quad \Theta_1 = 1 \;\; \textit{if} \;\; \theta_1 = 1. \tag{C.7}$$

Proof. Dividing both sides of (C.6) by Θ_t, summing them up from $t = 1$ to $t = T$, noting non-negativity of ϕ_t and (C.7), we obtain the result. □

Proof for Theorem 5.1. For convenience, let g_t be the gradient estimator at t step. Thus, $g_t = \tilde{\nabla}_t$ for the first order method while in the zeroth order setting $g_t = \bar{G}_\nu^t$.

$$
\begin{aligned}
f(x_t) &\le f(x_{t-1}) + \langle \nabla f(x_{t-1}), x_t - x_{t-1}\rangle + \frac{L}{2}\left\|x_t - x_{t-1}\right\|^2 \\
&= f(x_{t-1}) + \gamma_t\langle \nabla f(x_{t-1}), d_t - x_{t-1}\rangle + \frac{L\gamma_t^2}{2}\left\|d_t - x_{t-1}\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq f(x_{t-1}) + \gamma_t \langle g_t, d_t - x_{t-1} \rangle + \gamma_t \langle \nabla f(x_{t-1}) - g_t, d_t - x_{t-1} \rangle + \frac{LD^2\gamma_t^2}{2} \\
&\leq f(x_{t-1}) + \gamma_t \langle g_t, x^* - x_{t-1} \rangle + \gamma_t \langle \nabla f(x_{t-1}) - g_t, d_t - x_{t-1} \rangle + \frac{LD^2\gamma_t^2}{2} \\
&= f(x_{t-1}) + \gamma_t \langle \nabla f(x_{t-1}), x^* - x_{t-1} \rangle + \gamma_t \langle \nabla f(x_{t-1}) - g_t, d_t - x^* \rangle + \frac{LD^2\gamma_t^2}{2} \\
&\leq f(x_{t-1}) + \gamma_t (f(x^*) - f(x_{t-1})) + \gamma_t \langle \nabla f(x_{t-1}) - g_t, d_t - x^* \rangle + \frac{LD^2\gamma_t^2}{2} \\
&\leq f(x_{t-1}) + \gamma_t (f(x^*) - f(x_{t-1})) + \frac{\gamma_t}{2\beta} \left\| \nabla f(x_{t-1}) - g_t \right\|^2 + \frac{D^2\gamma_t(L\gamma_t + \beta)}{2}.
\end{aligned}
$$

The last inequality comes from the Young's inequality: for any $\beta > 0$,

$$
\begin{aligned}
\langle \nabla f(x_{t-1}) - g_t, d_t - x^* \rangle &\leq \frac{1}{2\beta} \left\| \nabla f(x_{t-1}) - g_t \right\|^2 + \frac{\beta}{2} \left\| d_t - x^* \right\|^2 \\
&\leq \frac{1}{2\beta} \left\| \nabla f(x_{t-1}) - g_t \right\|^2 + \frac{D^2\beta}{2}.
\end{aligned}
$$

Denote $\phi_t = f(x_t) - f(x^*)$. Substracting $f(x^*)$ from both sides of the inequality and taking the conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_{t-1}]$, we have

$$
\text{(C.8)} \qquad \mathbb{E}[\phi_t|\mathcal{F}_{t-1}] \leq (1-\gamma_t)\phi_{t-1} + \frac{\gamma_t}{2\beta}\mathbb{E}[\left\| \nabla f(x_{t-1}) - g_t \right\|^2 |\mathcal{F}_{t-1}] + \frac{D^2\gamma_t(L\gamma_t + \beta)}{2}.
$$

For the first-order gradient estimator $g_t = \tilde{\nabla}_t$, we have the following bound for its variance under Assumption 5.4:

$$
\mathbb{E}[\left\| \nabla f(x_{t-1}) - \tilde{\nabla}_t \right\|^2 |\mathcal{F}_{t-1}] = \frac{1}{b_t}\mathbb{E}[\left\| \nabla f(x_{t-1}) - \nabla F(x_{t-1}, \xi_{t,j}) \right\|^2 |\mathcal{F}_{t-1}] \leq \frac{2\rho L \phi_{t-1}}{b_t}.
$$

Then by (C.8), we can obtain

$$
\mathbb{E}[\phi_t|\mathcal{F}_{t-1}] \leq (1-\gamma_t)\phi_{t-1} + \frac{\gamma_t \rho L}{\beta b_t}\phi_{t-1} + \frac{D^2\gamma_t(L\gamma_t + \beta)}{2}.
$$

Let $\gamma_t = \frac{4}{t+3}, \beta = \rho L \gamma_t = \frac{4\rho L}{t+3} > 0, b_t = \lceil (t+3)/2 \rceil$, then

$$
\text{(C.9)} \qquad \mathbb{E}[\phi_t|\mathcal{F}_{t-1}] \leq \left(1 - \frac{2}{t+3}\right)\phi_{t-1} + \frac{8(\rho+1)LD^2}{(t+3)^2}.
$$

Now, letting $\theta_t = \frac{2}{t+3}$, it is easy to check that $\Theta_t = \frac{6}{(t+2)(t+3)}$ due to (C.7). Hence, in the view of Lemma C.3, we have

$$\mathbb{E}[\phi_t] \leq \frac{6\phi_0}{(t+2)(t+3)} + \frac{8(\rho+1)LD^2}{t+3} \leq \frac{2[\phi_0 + 4(\rho+1)LD^2]}{t+3}.$$

The above inequality implies that to attain an ϵ-optimal point, the total number of interations T can be bounded by $\mathcal{O}(1/\epsilon)$. Hence, the number of the gradient calls $\sum_{t=1}^{T} b_t$ can be bounded by $\frac{T^2+7T}{4} = \mathcal{O}(T^2)$, and the number of calls to the linear minimization oracle immediately follows from this observation.

We now prove part (b). For the zeroth-order version, by (C.5) in Lemma C.2 and (C.8), we can obtain

$$\begin{aligned}\mathbb{E}[\phi_t|\mathcal{F}_{t-1}] &\leq (1-\gamma_t)\phi_{t-1} + \frac{\gamma_t}{2\beta}\mathbb{E}[\left\|\nabla f(x_{t-1}) - \bar{G}_\nu^t\right\|^2|\mathcal{F}_{t-1}] + \frac{D^2\gamma_t(L\gamma_t+\beta)}{2} \\ &\leq (1-\gamma_t)\phi_{t-1} + \frac{2\gamma_t\rho L(d+4)}{\beta b_t}\phi_{t-1} + \frac{\gamma_t\nu^2L^2(d+6)^3}{2\beta} + \frac{D^2\gamma_t(L\gamma_t+\beta)}{2}\end{aligned}$$

Let $\gamma_t = \frac{4}{t+3}, \beta = \gamma_t\rho L, b_t = (t+3)(d+4), \nu = D(T+3)^{-1}(d+6)^{-3/2} \leq D(t+3)^{-1}(d+6)^{-3/2}$, then we have

$$\mathbb{E}[\phi_t|\mathcal{F}_{t-1}] \leq \left(1 - \frac{2}{t+3}\right)\phi_{t-1} + \frac{8(\rho+\rho^{-1}+1)}{(t+3)^2LD^2}$$

Similarly, in the view of Lemma C.3, we obtain

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{2[f(x_0) - f(x^*)] + 8(\rho+\rho^{-1}+1)LD^2}{t+3}.$$

The above inequality implies that to attain an ϵ-optimal point, the total number of interations T can be bounded by $\mathcal{O}(1/\epsilon)$. Hence, the number of calls to the zeroth-order oracles $2\sum_{t=1}^{T} b_t$ can be bounded by $(d+4)(T^2+7T) = \mathcal{O}(dT^2)$, and the number of calls to the linear minimization oracle immediately follows from this observation. □

C.2. Proof of Theorem 5.2

Proof of Theorem 5.2. For convenience, let g_t be the gradient estimator at t step. Thus, $g_t = \tilde{\nabla}_t$ for the first order method while in the zeroth order setting $g_t = \bar{G}_\nu^t$. First note that by the updates in Algorithm 10, the convexity and the smoothness of f, we have

$$
\begin{aligned}
f(x_t) \leq & f(z_t) + \langle \nabla f(z_t), x_t - z_t \rangle + \frac{L}{2} \left\| x_t - z_t \right\|^2 \\
= & (1-\gamma_t)[f(z_t) + \langle \nabla f(z_t), x_{t-1} - z_t \rangle] + \gamma_t [f(z_t) + \langle \nabla f(z_t), y_t - z_t \rangle] + \frac{L\gamma_t^2}{2} \left\| y_t - y_{t-1} \right\|^2 \\
\leq & (1-\gamma_t) f(x_{t-1}) + \gamma_t [f(z_t) + \langle \nabla f(z_t), y_t - z_t \rangle] + \frac{L\gamma_t^2}{2} \left\| y_t - y_{t-1} \right\|^2 \\
= & (1-\gamma_t) f(x_{t-1}) + \gamma_t [f(z_t) + \langle \nabla f(z_t), y_t - z_t \rangle] + \frac{\beta_t \gamma_t}{2} \left\| y_t - y_{t-1} \right\|^2 \\
& - \frac{\gamma_t(\beta_t - L\gamma_t)}{2} \left\| y_t - y_{t-1} \right\|^2 . \qquad \text{(C.10)}
\end{aligned}
$$

And by (5.5), we have

$$
\langle g_t + \beta_t (y_t - y_{t-1}), y_t - x \rangle \leq \eta_t, \quad \forall x \in \Omega.
$$

Let $x = x^*$ in the above inequality. Then we have

$$
\begin{aligned}
\frac{1}{2} \left\| y_t - y_{t-1} \right\|^2 = & \frac{1}{2} \left\| y_{t-1} - x^* \right\|^2 - \langle y_{t-1} - y_t, y_t - x^* \rangle - \frac{1}{2} \left\| y_t - x^* \right\|^2 \\
\leq & \frac{1}{2} \left\| y_{t-1} - x^* \right\|^2 + \frac{1}{\beta_t} \langle g_t, x^* - y_t \rangle - \frac{1}{2} \left\| y_t - x^* \right\|^2 + \frac{\eta_t}{\beta_t}. \qquad \text{(C.11)}
\end{aligned}
$$

Denoting $\delta_t = g_t - \nabla f(z_t)$ and combining (C.10) and (C.11), we obtain

$$
\begin{aligned}
f(x_t) \leq & (1-\gamma_t) f(x_{t-1}) + \gamma_t f(x^*) + \gamma_t \langle \delta_t, x^* - y_t \rangle \\
& + \frac{\beta_t \gamma_t}{2} (\left\| y_{t-1} - x^* \right\|^2 - \left\| y_t - x^* \right\|^2) + \eta_t \gamma_t - \frac{\gamma_t}{2} (\beta_t - L\gamma_t) \left\| y_t - y_{t-1} \right\|^2 \\
= & (1-\gamma_t) f(x_{t-1}) + \gamma_t f(x^*) + \frac{\beta_t \gamma_t}{2} (\left\| y_{t-1} - x^* \right\|^2 - \left\| y_t - x^* \right\|^2) + \eta_t \gamma_t \\
& + \gamma_t \langle \delta_t, x^* - y_{t-1} \rangle + \gamma_t \langle \delta_t, y_{t-1} - y_t \rangle - \frac{\gamma_t}{2} (\beta_t - L\gamma_t) \left\| y_t - y_{t-1} \right\|^2 \\
\leq & (1-\gamma_t) f(x_{t-1}) + \gamma_t f(x^*) + \frac{\beta_t \gamma_t}{2} (\left\| y_{t-1} - x^* \right\|^2 - \left\| y_t - x^* \right\|^2) + \eta_t \gamma_t \\
& + \gamma_t \langle \delta_t, x^* - y_{t-1} \rangle + \frac{\gamma_t \left\| \delta_t \right\|^2}{2(\beta_t - L\gamma_t)},
\end{aligned}
$$

where the last inequality comes from the fact that

$$
\gamma_t \langle \delta_t, y_{t-1} - y_t \rangle \leq \frac{\gamma_t}{2(\beta_t - L\gamma_t)} \left\| \delta_t \right\|^2 + \frac{\gamma_t(\beta_t - L\gamma_t)}{2} \left\| y_t - y_{t-1} \right\|^2 .
$$

Substracting $f(x^*)$ from both sides of the above inequality, denoting $\phi_t = f(x_t) - f(x^*)$, $\theta_t = \gamma_t$, and in the view of Lemma C.3, we obtain

$$\phi_t \leq \Theta_t \left[\phi_0 + \sum_{k=1}^{t} \frac{B_k}{\Theta_k} \right], \tag{C.12}$$

where

$$B_t = \frac{\beta_t \gamma_t}{2} (\|y_{t-1} - x^*\|^2 - \|y_t - x^*\|^2) + \eta_t \gamma_t + \gamma_t \langle \delta_t, x^* - y_{t-1} \rangle + \frac{\gamma_t \|\delta_t\|^2}{2(\beta_t - L\gamma_t)}.$$

Choosing $\gamma_t = \theta_t = \frac{3}{t+2}$, we can easily check that $\Theta_t = \frac{6}{t(t+1)(t+2)}$ due to (C.7). Moreover, letting $\beta_t = \frac{4L}{t+2}, \eta_t = \frac{LD^2}{t(t+1)}$, we have $\sum_{k=1}^{t} \frac{\eta_k \gamma_k}{\Theta_k} \leq \frac{tLD^2}{2}$ and

$$\begin{aligned}
&\sum_{k=1}^{t} \frac{\beta_k \gamma_k}{\Theta_k} (\|y_{t-1} - x^*\|^2 - \|y_t - x^*\|^2) \\
&\leq \frac{\beta_1 \gamma_1}{\Theta_i} \|y_0 - x^*\|^2 + \sum_{k=2}^{t} \left(\frac{\beta_k \gamma_k}{\Theta_k} - \frac{\beta_{k-1} \gamma_{k-1}}{\Theta_{k-1}} \right) \|y_{t-1} - x^*\|^2 \\
&\leq \frac{\beta_1 \gamma_1}{\Theta_i} D^2 + \sum_{k=2}^{t} \left(\frac{\beta_k \gamma_k}{\Theta_k} - \frac{\beta_{k-1} \gamma_{k-1}}{\Theta_{k-1}} \right) D^2 = \frac{\beta_t \gamma_t D^2}{\Theta_t} = \frac{2LD^2 t(t+1)}{t+2},
\end{aligned}$$

where the last inequality comes from the fact $\frac{\beta_k \gamma_k}{\Theta_k} > \frac{\beta_{k-1} \gamma_{k-1}}{\Theta_{k-1}}$.

We now prove part (a). Let $g_t = \tilde{\nabla}_t$. Taking expectation for both sides of (C.12), and noting that $\mathbb{E}[\langle \delta_t, x^* - y_{t-1} \rangle] = 0$ and

$$\begin{aligned}
\mathbb{E}[\|\delta_t\|^2 | \mathcal{F}_{t-1}] &\leq \frac{2\rho L}{b_t} (f(z_t) - f(x^*)) \quad \triangleright \text{ by Assumption 5.4} \\
&\leq \frac{2\rho L}{b_t} \Big((1-\gamma_t)\phi_{t-1} + \gamma_t (f(y_{t-1}) - f(x^*)) \Big) \quad \triangleright z_t = (1-\gamma_t) x_{t-1} + \gamma_t y_{t-1} \\
&\leq \frac{2\rho L}{b_t} \left((1-\gamma_t)\phi_{t-1} + \gamma_t (\|\nabla f(x^*)\| D + \frac{LD^2}{2}) \right) \quad \triangleright \text{ by the smoothness} \\
&:= \frac{2\rho L}{b_t} \left((1-\gamma_t)\phi_{t-1} + \gamma_t \frac{KLD^2}{2} \right), \quad \triangleright K = \frac{\|\nabla f(x^*)\|}{LD} + 1
\end{aligned}$$

we can obtain

$$\mathbb{E}[\phi_t] \le \frac{6LD^2}{(t+2)^2} + \frac{3LD^2}{(t+1)(t+2)} + \frac{3}{t(t+1)(t+2)} \sum_{k=1}^{t} \frac{\rho k(k+1)\left((k-1)\mathbb{E}[\phi_{k-1}] + \frac{3KLD^2}{2}\right)}{b_k}$$

We now prove

$$\mathbb{E}[\phi_t] \le \frac{6LD^2}{(t+2)^2} + \frac{(12+3K)LD^2}{(t+1)(t+2)} \tag{C.13}$$

by induction. Set $b_k = \lceil 3\rho k(k+1) \rceil$. It is easy to check $\mathbb{E}[\phi_0] \le \frac{KLD^2}{2}$ by the smoothness of f which satisfies (C.13). If (C.13) holds for all $k \le t-1$, then with the above inequality we can obtain

$$\begin{aligned}
\mathbb{E}[\phi_t] &\le \frac{6LD^2}{(t+2)^2} + \frac{(3+\frac{3K}{2})LD^2}{(t+1)(t+2)} + \frac{1}{t(t+1)(t+2)} \sum_{k=1}^{t} (k-1)\mathbb{E}[\phi_{k-1}] \\
&\le \frac{6LD^2}{(t+2)^2} + \frac{(3+\frac{3K}{2})LD^2}{(t+1)(t+2)} + \frac{1}{t(t+1)(t+2)} \sum_{k=1}^{t} \left(\frac{6LD^2(k-1)}{(k+1)^2} + \frac{(12+3K)LD^2(k-1)}{k(k+1)} \right) \\
&\le \frac{6LD^2}{(t+2)^2} + \frac{(3+\frac{3K}{2})LD^2}{(t+1)(t+2)} + \frac{(18+3K)LD^2}{t(t+1)(t+2)} \sum_{k=1}^{t} \frac{1}{k+1} \\
&\le \frac{6LD^2}{(t+2)^2} + \frac{(12+3K)LD^2}{(t+1)(t+2)},
\end{aligned}$$

i.e., (C.13) holds for $k = t$. Therefore, to achieve an ϵ-optimal point, the number of outer iterations T can be bounded by $\mathcal{O}(1/\sqrt{\epsilon})$. Hence, the number of calls to the first order oracles can be bounded by

$$\sum_{t=1}^{T} b_t \le 3\rho \sum_{t=1}^{T} t(t+1) = \rho T(T+1)(T+2) = \mathcal{O}(T^3).$$

Due to the fact that the inner iterations indeed solves a convex constrained optimization problem by the classical Frank-Wolfe method with the exact line search, one can show that the number of inner iterations N_t performed at the t-th out iteration can be bounded by

$$N_t \le \left\lceil \frac{6\beta_t D^2}{\eta_t} \right\rceil = \mathcal{O}(t).$$

Thus, the number of calls to the linear minimization oracle can be bounded by

$$\sum_{t=1}^{T} N_t \le \mathcal{O}(T^2).$$

We now prove part (b). Let $g_t = \bar{G}_\nu^t$. Notice that $\bar{G}_\nu^t$ is a biased estimator of $\nabla f(z_t)$. We can obtain the following results by (C.2):

$$\begin{aligned}\mathbb{E}[\langle \delta_t, x^* - y_{t-1}\rangle] &= \mathbb{E}[\langle \nabla f_\nu(z_t) - \nabla f(z_t), x^* - y_{t-1}\rangle] + \mathbb{E}[\langle \bar{G}_\nu^t - \nabla f_\nu(z_t), x^* - y_{t-1}\rangle] \\ &= \mathbb{E}[\langle \nabla f_\nu(z_t) - \nabla f(z_t), x^* - y_{t-1}\rangle] \le \frac{\nu LD(d+3)^{3/2}}{2}.\end{aligned}$$

Besides, we can obtain a similar bound for $\mathbb{E}[\|\delta_t\|^2]$ by Lemma C.2.

$$\begin{aligned}\mathbb{E}[\|\delta_t\|^2 \,|\mathcal{F}_{t-1}] &\le \frac{4\rho L(d+4)(f(z_t) - f(x^*))}{b_t} + \nu^2 L^2 (d+6)^3 \\ &\le \frac{4\rho L(d+4)\big((1-\gamma_t)\phi_{t-1} + \frac{LD^2\gamma_t}{2}\big)}{b_t} + \nu^2 L^2(d+6)^3.\end{aligned}$$

where the last inequality is slightly different from the one for the first-order setting due to $\|\nabla f(x^*)\| = 0$ for convex cases under the moment-based WGC.

By (C.12), we have the following simplified inequality:

$$\begin{aligned}\mathbb{E}[\phi_t] \le &\frac{6LD^2}{(t+2)^2} + \frac{3LD^2}{(t+1)(t+2)} + \frac{\nu LD(d+3)^{3/2}}{2} + \frac{3\nu^2 L(d+6)^3}{2} \\ &+ \frac{6}{t(t+1)(t+2)} \sum_{k=1}^{t} \frac{\rho(d+4)k(k+1)\Big((k-1)\mathbb{E}[\phi_{k-1}] + \frac{3LD^2}{2}\Big)}{b_k}.\end{aligned}$$

Set $b_k = \lceil 6\rho k(k+1)(d+4)\rceil, \nu = \frac{D}{(T+2)^2(d+6)^{3/2}} \le \frac{D}{(t+2)^2(d+6)^{3/2}}$. Then we have

$$\mathbb{E}[\phi_t] \le \frac{8LD^2}{(t+2)^2} + \frac{12LD^2}{(t+1)(t+2)} + \frac{1}{t(t+1)(t+2)} \sum_{k=1}^{t} (k-1)\mathbb{E}[\phi_{k-1}].$$

Similar to the proof for part (a), we can finish the proof by induction and obtain the bounds for complexity. □

C.3. Zeroth-order SGD under Growth Conditions

In this section, we highlight that one can extend the results in [**VBS19**] only assuming access to stochastic zeroth-order oracle with corresponding variance-based growth conditions. Notice that both SGC and WGC are defined in the format of the relative shrinkage of $\mathbb{E}\|\nabla F(x,\xi)\|^2$. However, in the unconstrained setting, the corresponding variance-based versions are equivalent to the moment-based growth conditions (see Proposition 5.1 for WGC; for SGC, note that $\mathbb{E}\|\nabla F(x,\xi) - \nabla f(x)\|^2 = \mathbb{E}\|\nabla F(x,\xi)\|^2 - \|\nabla f(x)\|^2 = (\rho - 1)\|\nabla f(x)\|^2$).

We present the following result for the zeroth-order setting which directly follows the proofs in [**VBS19**]. We highlight that it is the zeroth-order version of Theorem 3 in [**VBS19**]. Similar results for other setups considered in [**VBS19**] can also be obtained for the zeroth-order setting.

Algorithm 12 Non-convex Zeroth-order SGD (ZO-SGD)

```
Input: $x_0 \in \Omega$, number of iterations $T$, $\eta$
for $t = 1, 2, \dots, T$ do
    Randomly pick $\xi_t$ and compute
        $x_t = x_{t-1} - \eta \frac{F(x_{t-1} + \nu u_t, \xi_t) - F(x_{t-1}, \xi_t)}{\nu} u_t := x_{t-1} - \eta G_t.$
    where $u_t$ is generated from $(0, \mathbf{I}_d)$.
end for
Output: $x_R$ where $R$ is uniformly distributed over $0, \dots, T-1$
```

Theorem C.1. *Consider solving the non-convex unconstrained L-smooth problem by Algorithm 12 with some appropriate constant step size η, if f satistifies SGC with constant ρ, then*

$$\mathbb{E}\|\nabla f(x_R)\|^2 \leq \mathcal{O}\left(\frac{1}{T}\right)$$

Proof. The zeroth-order SGD update is given by

$$x_t = x_{t-1} - \eta \frac{F(x_{t-1} + \nu u_t, \xi_t) - F(x_{t-1}, \xi_t)}{\nu} u_t := x_{t-1} - \eta G_t.$$

By the smoothness of f, we have

$$\begin{aligned} f(x_t) - f(x_{t-1}) &\leq \langle \nabla f(x_{t-1}, x_t - x_{t-1}\rangle + \frac{L}{2}\|x_t - x_{t-1}\|^2 \\ &= -\eta\langle \nabla f(x_{t-1}), G_t\rangle + \frac{L\eta^2}{2}\|G_t\|^2 \end{aligned}$$

Consider the term $\langle \nabla f(x_{t-1}), G_t\rangle$. Taking expectation with respect to ξ_t, u_t, we have

$$\begin{aligned} \mathbb{E}[\langle \nabla f(x_{t-1}), G_t\rangle] &= \langle \nabla f(x_{t-1}), \nabla f_\nu(x_{t-1})\rangle \\ &= \langle \nabla f(x_{t-1}), \nabla f(x_{t-1}) + \nabla f_\nu(x_{t-1}) - \nabla f(x_{t-1})\rangle \\ &\geq \|\nabla f(x_{t-1})\|^2 - \frac{\nu L(d+3)^{3/2}}{2}\|\nabla f(x_{t-1})\| \end{aligned}$$

Consider the term $\|G_t\|^2$. Taking expectation with respect to ξ_t, u_t, we have

$$\begin{aligned}\mathbb{E}\|G_t\|^2 &\leq \frac{\nu^2}{2}L^2(d+6)^3 + 2(d+4)\mathbb{E}\|\nabla F(x_{t-1},\xi_t)\|^2\\ &\leq \frac{\nu^2}{2}L^2(d+6)^3 + 2\rho(d+4)\|\nabla f(x_{t-1})\|^2\end{aligned}$$

Then, by the above inequalities, we can obtain

$$\begin{aligned}&\mathbb{E}[f(x_t) - f(x_{t-1})]\\ &\leq -\eta\|\nabla f(x_{t-1})\|^2 + \eta\frac{\nu L(d+3)^{3/2}}{2}\|\nabla f(x_{t-1})\| + \eta^2 L\rho(d+4)\|\nabla f(x_{t-1})\|^2 + \eta^2\frac{\nu^2}{4}L^3(d+6)^2\\ &\leq -\eta\|\nabla f(x_{t-1})\|^2 + \eta^2 L(d+3)\|\nabla f(x_{t-1})\|^2 + \frac{\nu^2 L(d+3)^2}{16}\\ &\qquad + \eta^2 L\rho(d+4)\|\nabla f(x_{t-1})\|^2 + \eta^2\frac{\nu^2}{4}L^3(d+6)^2\\ &\leq -\eta\|\nabla f(x_{t-1})\|^2 + \eta^2 L(\rho+1)(d+4)\|\nabla f(x_{t-1})\|^2 + \eta^2\frac{\nu^2}{4}L^3(d+6)^2 + \frac{\nu^2 L(d+3)^2}{16}.\end{aligned}$$

If $\eta = \frac{1}{2L(\rho+1)(d+4)}$, then we have

$$\begin{aligned}\mathbb{E}[f(x_t) - f(x_{t-1})] &\leq -\frac{\eta}{2}\|\nabla f(x_{t-1})\|^2 + \eta^2\frac{\nu^2}{4}L^3(d+6)^2 + \frac{\nu^2 L(d+3)^2}{16}\\ \Rightarrow \|\nabla f(x_{t-1})\|^2 &\leq \frac{2}{\eta}\mathbb{E}[f(x_{t-1}) - f(x_t)] + \eta\frac{\nu^2}{2}L^3(d+6)^2 + \frac{\nu^2 L(d+3)^2}{8\eta}\end{aligned}$$

Setting $\nu = \mathcal{O}(1/\sqrt{dT})$ and taking a telescoping sum of the above inequality, we can get the same $\mathcal{O}(1/T)$ rate for the non-convex setting. □

In the above proof, we did not pay careful attention to the exact constants of the tuning parameter, as our main point is to simply highlight it is possible to obtain a zeroth-order version of the results in [**VBS19**] under variance-based growth conditions and the logic of the proof is the same as [**VBS19**].

Bibliography

[AB13] P. Alquier and G. Biau, *Sparse single-index model.*, Journal of Machine Learning Research **14** (2013), no. 1.

[ABD$^+$18] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, *A reductions approach to fair classification*, International Conference on Machine Learning, 2018, pp. 60–69.

[ABTR21] Z. Akhtar, A. S. Bedi, S. T. Thomdapu, and K. Rajawat, *Projection-Free Algorithm for Stochastic Bi-level Optimization*, arXiv preprint arXiv:2110.11721 (2021).

[ACD$^+$19] Y. Arjevani, Y. Carmon, J. Duchi, D. Foster, N. Srebro, and B. Woodworth, *Lower bounds for non-convex stochastic optimization*, arXiv preprint arXiv:1912.02365 (2019).

[AF21] R. Astudillo and P. Frazier, *Bayesian optimization of function networks*, Advances in Neural Information Processing Systems **34** (2021).

[AFM17] Y. Atchadé, G. Fort, and E. Moulines, *On stochastic proximal gradient algorithms*, Journal of Machine Learning Research (2017).

[AYS19] S. Alghunaim, K. Yuan, and A. H. Sayed, *A linearly convergent proximal gradient algorithm for decentralized optimization*, Advances in Neural Information Processing Systems **32** (2019).

[BBM18] R. Bassily, M. Belkin, and S. Ma, *On exponential convergence of sgd in non-convex over-parametrized learning*, arXiv preprint arXiv:1811.02564 (2018).

[BG21] K. Balasubramanian and S. Ghadimi, *Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points*, Foundations of Computational Mathematics (2021), 1–42.

[BG22] ______, *Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points*, Foundations of Computational Mathematics **22** (2022), no. 1, 35–76.

[BGI$^+$17] J. Blanchet, D. Goldfarb, G. Iyengar, F. Li, and C. Zhou, *Unbiased simulation for optimizing stochastic function compositions*, arXiv preprint arXiv:1711.07564 (2017).

[BGN22] K. Balasubramanian, S. Ghadimi, and A. Nguyen, *Stochastic multilevel composition optimization algorithms with level-independent convergence rates*, SIAM Journal on Optimization **32** (2022), no. 2, 519–544.

[BHJ$^+$18] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, *Fairness in criminal justice risk assessments: The state of the art*, Sociological Methods & Research (2018).

[BL08] K. Bredies and D. A. Lorenz, *Linear convergence of iterative soft-thresholding*, Journal of Fourier Analysis and Applications **14** (2008), 813–837.

[BM08] S. Boyd and A. Mutapcic, *Stochastic subgradient methods*, Lecture Notes for EE364b, Stanford University (2008), 97.

[BS17] A. Beck and S. Shtern, *Linearly convergent away-step conditional gradient for non-strongly convex functions*, Mathematical Programming **164** (2017), no. 1-2, 1–27.

[BT00] D. P. Bertsekas and J. N. Tsitsiklis, *Gradient convergence in gradient methods with errors*, SIAM Journal on Optimization **10** (2000), no. 3, 627–642.

[BZK18] L. Berrada, A. Zisserman, and M. P. Kumar, *Deep frank-wolfe for neural network optimization*, arXiv preprint arXiv:1811.07591 (2018).

[CFKM20] W. Cong, R. Forsati, M. Kandemir, and M. Mahdavi, *Minimal variance sampling with provable guarantees for fast training of graph neural networks*, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1393–1403.

[CLK$^+$12] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing, *Smoothing proximal gradient method for general structured sparse regression*, (2012).

[CO19] A. Cutkosky and F. Orabona, *Momentum-based variance reduction in non-convex sgd*, Advances in neural information processing systems **32** (2019).

[CRS$^+$18a] K. Choromanski, M. Rowland, V. Sindhwani, R. Turner, and A. Weller, *Structured evolution with compact architectures for scalable policy optimization*, arXiv preprint arXiv:1804.02395 (2018).

[CRS$^+$18b] ———, *Structured evolution with compact architectures for scalable policy optimization*, Proceedings of the 35th International Conference on Machine Learning, PMLR, 2018.

[CSY21] 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** (2021), 4937–4948.

[CZS$^+$17] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, *ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models*, Proceedings of the 10th ACM workshop on artificial intelligence and security, 2017, pp. 15–26.

[DB19] A. Defazio and L. Bottou, *On the ineffectiveness of variance reduced optimization for deep learning*, Advances in Neural Information Processing Systems, 2019, pp. 1753–1763.

[DBLJ14] 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).

[DD19] D. Davis and D. Drusvyatskiy, *Stochastic model-based minimization of weakly convex functions*, SIAM Journal on Optimization **29** (2019), no. 1, 207–239.

[DF21] L. Dalcin and Y.-L. L. Fang, *mpi4py: Status update after 12 years of development*, Computing in Science & Engineering **23** (2021), no. 4, 47–54.

[DIKL18] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson, *Decoupled classifiers for group-fair and efficient machine learning*, Conference on Fairness, Accountability and Transparency, 2018, pp. 119–133.

[DJWW15] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, *Optimal rates for zero-order convex optimization: The power of two function evaluations*, IEEE Transactions on Information Theory **61** (2015), no. 5, 2788–2806.

[DLS16] P. Di Lorenzo and G. Scutari, *Next: In-network nonconvex optimization*, IEEE Transactions on Signal and Information Processing over Networks **2** (2016), no. 2, 120–136.

[DOBD$^+$18] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil, *Empirical risk minimization under fairness constraints*, Advances in Neural Information Processing Systems, 2018, pp. 2791–2801.

[DPR17] D. Dentcheva, S. Penev, and A. Ruszczyński, *Statistical estimation of composite risk functionals and risk optimization problems*, Annals of the Institute of Statistical Mathematics **69** (2017), no. 4, 737–760.

[DR70] V. Demyanov and A. Rubinov, *Approximate methods in optimization problems*, American Elsevier Publishing Co, 1970.

[DR18] J. Duchi and F. Ruan, *Stochastic methods for composite and weakly convex optimization problems*, SIAM Journal on Optimization **28** (2018), no. 4, 3229–3259.

[EN13] Y. Ermoliev and V. Norkin, *Sample average approximation method for compound stochastic optimization problems*, SIAM Journal on Optimization **23** (2013), no. 4, 2231–2263.

[Erm76] Y. Ermoliev, *Methods of stochastic programming*, Nauka, Moscow (1976).

[FGM17] R. M. Freund, P. Grigas, and R. Mazumder, *An extended frank-wolfe method with in-face directions, and its application to low-rank matrix completion*, SIAM Journal on optimization **27** (2017), no. 1, 319–346.

[FMO21] A. Fallah, A. Mokhtari, and A. Ozdaglar, *Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks*, Advances in Neural Information Processing Systems **34** (2021).

[FSS15] F. Facchinei, G. Scutari, and S. Sagratella, *Parallel selective algorithms for nonconvex big data optimization*, IEEE Transactions on Signal Processing **63** (2015), no. 7, 1874–1889.

[FW56] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval research logistics quarterly **3** (1956), no. 1-2, 95–110.

[GH13] D. Garber and E. Hazan, *A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization*, arXiv preprint arXiv:1301.4666 (2013).

[GH15] ———, *Faster rates for the frank-wolfe method over strongly-convex sets*, International Conference on Machine Learning, PMLR, 2015, pp. 541–549.

[Gha19] S. Ghadimi, *Conditional gradient type methods for composite nonlinear and stochastic optimization*, Mathematical Programming **173** (2019), no. 1-2, 431–464.

[GKS21] D. Garber, A. Kaplan, and S. Sabach, *Improved complexities of conditional gradient-type methods with applications to robust matrix recovery problems*, Mathematical Programming **186** (2021), no. 1, 185–208.

[GL13] S. Ghadimi and G. Lan, *Stochastic first-and zeroth-order methods for nonconvex stochastic programming*, SIAM Journal on Optimization **23** (2013), no. 4, 2341–2368.

[GLQ$^+$19] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik, *Sgd: General analysis and improved rates*, arXiv preprint arXiv:1901.09401 (2019).

[GRW20] S. Ghadimi, A. Ruszczynski, and M. Wang, *A single timescale stochastic approximation method for nested stochastic optimization*, SIAM Journal on Optimization **30** (2020), no. 1, 960–979.

[GS21] C. Geiersbach and T. Scarinci, *Stochastic proximal gradient methods for nonconvex problems in hilbert spaces*, Computational optimization and applications **78** (2021), 705–740.

[GSK18] D. Garber, S. Sabach, and A. Kaplan, *Fast generalized conditional gradient method with applications to matrix recovery problems*, arXiv preprint arXiv:1802.05581 (2018).

[GW21] D. Garber and N. Wolf, *Frank-Wolfe with a nearest extreme point oracle*, Conference on Learning Theory, PMLR, 2021, pp. 2103–2132.

[Hea82] D. W. Hearn, *The gap function of a convex program*, Operations Research Letters **1** (1982), no. 2, 67–71.

[HHZ17] M. Hong, D. Hajinezhad, and M.-M. Zhao, *Prox-pda: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks*, International Conference on Machine Learning, PMLR, 2017, pp. 1529–1538.

[HJN15] Z. Harchaoui, A. Juditsky, and A. Nemirovski, *Conditional gradient algorithms for norm-regularized smooth convex optimization*, Mathematical Programming **152** (2015), no. 1-2, 75–112.

[HK12] E. Hazan and S. Kale, *Projection-free online learning*, 29th International Conference on Machine Learning, ICML 2012, 2012, pp. 521–528.

[HK14] ______, *Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization*, The Journal of Machine Learning Research **15** (2014), no. 1, 2489–2512.

[HKMS19] H. Hassani, A. Karbasi, A. Mokhtari, and Z. Shen, *Stochastic conditional gradient++*, arXiv preprint arXiv:1902.06992 (2019).

[HL16] E. Hazan and H. Luo, *Variance-reduced and projection-free stochastic optimization*, International Conference on Machine Learning, 2016, pp. 1263–1271.

[HLPR19] N. J. Harvey, C. Liaw, Y. Plan, and S. Randhawa, *Tight analyses for non-smooth stochastic gradient descent*, Conference on Learning Theory, PMLR, 2019, pp. 1579–1613.

[HLVDMW17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[HM21] L. Hodgkinson and M. Mahoney, *Multiplicative noise and heavy tails in stochastic optimization*, International Conference on Machine Learning, PMLR, 2021, pp. 4262–4274.

[HMRT19] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, *Surprises in high-dimensional ridgeless least squares interpolation*, arXiv preprint arXiv:1903.08560 (2019).

[HPS16] M. Hardt, E. Price, and N. Srebro, *Equality of opportunity in supervised learning*, Advances in neural information processing systems, 2016, pp. 3315–3323.

[HZCH20] Y. Hu, S. Zhang, X. Chen, and N. He, *Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning*, Advances in Neural Information Processing Systems **33** (2020).

[HZSL13] K. Hou, Z. Zhou, A. M.-C. So, and Z.-Q. Luo, *On the linear convergence of the proximal gradient method for trace norm regularization*, Advances in Neural Information Processing Systems **26** (2013).

[Jag13] M. Jaggi, *Revisiting frank-wolfe: Projection-free sparse convex optimization.*, ICML (1), 2013, pp. 427–435.

[JEH16] K. Jaganathan, Y. C. Eldar, and B. Hassibi, *Phase retrieval: An overview of recent developments*, Optical Compressive Imaging (2016), 279–312.

[JNG$^+$19] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan, *A short note on concentration inequalities for random vectors with subgaussian norm*, arXiv preprint arXiv:1902.03736 (2019).

[JZ13] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, Advances in neural information processing systems **26** (2013).

[KLS21] A. Koloskova, T. Lin, and S. U. Stich, *An improved analysis of gradient tracking for decentralized machine learning*, Advances in Neural Information Processing Systems **34** (2021).

[KMR16] J. Kleinberg, S. Mullainathan, and M. Raghavan, *Inherent trade-offs in the fair determination of risk scores*, arXiv preprint arXiv:1609.05807 (2016).

[KNS16] H. Karimi, J. Nutini, and M. Schmidt, *Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition*, Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16, Springer, 2016, pp. 795–811.

[L$^+$15] Y. LeCun et al., *Lenet-5, convolutional neural networks*, URL: http://yann. lecun. com/exdb/lenet **20** (2015), no. 5, 14.

[Lan20] G. Lan, *First-order and stochastic optimization methods for machine learning*, vol. 1, Springer, 2020.

[LDS21] Y. Lu and C. De Sa, *Optimal complexity in decentralized training*, International Conference on Machine Learning, PMLR, 2021, pp. 7111–7123.

[LJJ15] S. Lacoste-Julien and M. Jaggi, *On the global linear convergence of frank-wolfe optimization variants*, Advances in Neural Information Processing Systems, 2015, pp. 496–504.

[LLT$^+$21] Y. Li, X. Liu, J. Tang, M. Yan, and K. Yuan, *Decentralized composite optimization with compression*, arXiv preprint arXiv:2108.04448 (2021).

[LM11] J. Liu and A. S. Morse, *Accelerated linear iterations for distributed averaging*, Annual Reviews in Control **35** (2011), no. 2, 160–165.

[LN13] S. Lee and A. Nedic, *Distributed random projection algorithm for convex optimization*, IEEE Journal of Selected Topics in Signal Processing **7** (2013), no. 2, 221–229.

[LO20] X. Li and F. Orabona, *A high probability analysis of adaptive sgd with momentum*, arXiv preprint arXiv:2007.14294 (2020).

[LP66] E. Levitin and B. Polyak, *Constrained minimization methods*, USSR Computational mathematics and mathematical physics **6** (1966), no. 5, 1–50.

[LPZZ17] G. Lan, S. Pokutta, Y. Zhou, and D. Zink, *Conditional accelerated lazy stochastic gradient descent*, International Conference on Machine Learning, 2017, pp. 1965–1974.

[LR18] T. Liang and A. Rakhlin, *Just interpolate: Kernel" ridgeless" regression can generalize*, arXiv preprint arXiv:1808.00387 (2018).

[LSY19] Z. Li, W. Shi, and M. Yan, *A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates*, IEEE Transactions on Signal Processing **67** (2019), no. 17, 4494–4506.

[LT10] Q. Ling and Z. Tian, *Decentralized sparse signal recovery for compressive sleeping wireless sensor networks*, IEEE Transactions on Signal Processing **58** (2010), no. 7, 3816–3827.

[LWK17] C. Louizos, M. Welling, and D. P. Kingma, *Learning sparse neural networks through l_0 regularization*, arXiv preprint arXiv:1712.01312 (2017).

[LZ16] G. Lan and Y. Zhou, *Conditional gradient sliding for convex optimization*, SIAM Journal on Optimization **26** (2016), no. 2, 1379–1409.

[LZSH19] S. Lu, X. Zhang, H. Sun, and M. Hong, *Gnsd: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization*, 2019 IEEE Data Science Workshop (DSW), IEEE, 2019, pp. 315–321.

[LZW22] Z. Lou, W. Zhu, and W. B. Wu, *Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent*, Journal of Machine Learning Research **23** (2022), 1–22.

[LZZ$^+$17] 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).

[MBB18] S. Ma, R. Bassily, and M. Belkin, *The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning*, International Conference on Machine Learning, 2018, pp. 3325–3334.

[MBG10] G. Mateos, J. A. Bazerque, and G. B. Giannakis, *Distributed sparse linear regression*, IEEE Transactions on Signal Processing **58** (2010), no. 10, 5262–5276.

[MBMXC22] G. Mancino-Ball, S. Miao, Y. Xu, and J. Chen, *Proximal stochastic recursive momentum methods for nonconvex composite decentralized optimization*, arXiv preprint arXiv:2211.11954 (2022).

[MDB21] L. Madden, E. Dall'Anese, and S. Becker, *High-probability convergence bounds for non-convex stochastic gradient descent*, arXiv preprint arXiv:2006.05610 (2021).

[MFGP17] K. Margellos, A. Falsone, S. Garatti, and M. Prandini, *Distributed constrained optimization and consensus in uncertain networks via proximal minimization*, IEEE Transactions on Automatic Control **63** (2017), no. 5, 1372–1387.

[MHK18a] A. Mokhtari, H. Hassani, and A. Karbasi, *Conditional gradient method for stochastic submodular maximization: Closing the gap*, International Conference on Artificial Intelligence and Statistics, PMLR, 2018, pp. 1886–1895.

[MHK18b] ———, *Stochastic conditional gradient methods: From convex minimization to submodular maximization*, arXiv preprint arXiv:1804.09554 (2018).

[MHK20] ———, *Stochastic conditional gradient methods: From convex minimization to submodular maximization*, Journal of machine learning research (2020).

[Mig94] A. Migdalas, *A regularization of the frank—wolfe method and unification of certain nonlinear programming methods*, Mathematical Programming **65** (1994), no. 1, 331–345.

[MRSY19] A. Montanari, F. Ruan, Y. Sohn, and J. Yan, *The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime*, arXiv preprint arXiv:1911.01544 (2019).

[MVL$^+$20] S. Y. Meng, S. Vaswani, I. Laradji, M. Schmidt, and S. Lacoste-Julien, *Fast and furious convergence: Stochastic second order methods under interpolation*, arXiv preprint arXiv:1910.04920 (2020).

[N$^+$18] Y. Nesterov et al., *Lectures on convex optimization*, vol. 137, Springer, 2018.

[Nit14] A. Nitanda, *Stochastic proximal gradient descent with acceleration techniques*, Advances in Neural Information Processing Systems **27** (2014).

[NOS17] A. Nedic, A. Olshevsky, and W. Shi, *Achieving geometric convergence for distributed optimization over time-varying graphs*, SIAM Journal on Optimization **27** (2017), no. 4, 2597–2633.

[NS17] Y. Nesterov and V. Spokoiny, *Random gradient-free minimization of convex functions*, Foundations of Computational Mathematics **17** (2017), no. 2, 527–566.

[NWS14] D. Needell, R. Ward, and N. Srebro, *Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm*, Advances in neural information processing systems, 2014, pp. 1017–1025.

[Ols17] A. Olshevsky, *Linear time average consensus and distributed optimization on fixed graphs*, SIAM Journal on Control and Optimization **55** (2017), no. 6, 3990–4014.

[Ora20] F. Orabona, *Almost sure convergence of sgd on smooth non-convex functions*, `https://parameterfree.com/2020/10/05/almost-sure-convergence-of-sgd-on-smooth-non-convex-functions/`, 2020, [Online; accessed 14-March-2023].

[PEK14] S. Patterson, Y. C. Eldar, and I. Keidar, *Distributed compressed sensing for static and time-varying networks*, IEEE Transactions on Signal Processing **62** (2014), no. 19, 4931–4946.

[PGM$^+$19] 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*, Advances in Neural Information Processing Systems 32 (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), Curran Associates, Inc., 2019, pp. 8024–8035.

[PN21] S. Pu and A. Nedić, *Distributed stochastic gradient tracking methods*, Mathematical Programming **187** (2021), no. 1, 409–457.

[PNPTD20] N. H. Pham, L. M. Nguyen, D. T. Phan, and Q. Tran-Dinh, *Proxsarah: An efficient algorithmic framework for stochastic composite nonconvex optimization*, The Journal of Machine Learning Research **21** (2020), no. 1, 4455–4502.

[Pol63] B. T. Polyak, *Gradient methods for minimizing functionals*, Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki **3** (1963), no. 4, 643–653.

[QGX$^+$21] Q. Qi, Z. Guo, Y. Xu, R. Jin, and T. Yang, *An online method for a class of distributionally robust optimization with non-convex objectives*, Advances in Neural Information Processing Systems **34** (2021).

[QHZ$^+$22] Z.-H. Qiu, Q. Hu, Y. Zhong, L. Zhang, and T. Yang, *Large-scale stochastic optimization of ndcg surrogates for deep learning with provable convergence*, arXiv preprint arXiv:2202.12183 (2022).

[QL17] G. Qu and N. Li, *Harnessing smoothness to accelerate distributed optimization*, IEEE Transactions on Control of Network Systems **5** (2017), no. 3, 1245–1260.

[QLX18] C. Qu, Y. Li, and H. Xu, *Non-convex conditional gradient sliding*, International Conference on Machine Learning, PMLR, 2018, pp. 4208–4217.

[QLX$^+$21] Q. Qi, Y. Luo, Z. Xu, S. Ji, and T. Yang, *Stochastic optimization of areas under precision-recall curves with provable convergence*, Advances in Neural Information Processing Systems **34** (2021).

[RBGM20] A. Roy, K. Balasubramanian, S. Ghadimi, and P. Mohapatra, *Escaping saddle-points faster under interpolation-like conditions*, Proceedings of the Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS) (2020).

[RDLS18] S. N. Ravi, T. Dinh, V. S. R. Lokhande, and V. Singh, *Constrained deep learning using conditional gradient and applications in computer vision*, arXiv preprint arXiv:1803.06453 (2018).

[RM51] H. Robbins and S. Monro, *A stochastic approximation method*, The Annals of Mathematical Statistics **22** (1951), no. 3, 400–407.

[RS06] A. Ruszczyński and A. Shapiro, *Optimization of convex risk functions*, Mathematics of operations research **31** (2006), no. 3, 433–452.

[RSPS16] S. J. Reddi, S. Sra, B. Póczos, and A. Smola, *Stochastic frank-wolfe methods for nonconvex optimization*, 2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2016, pp. 1244–1251.

[Rus87] A. Ruszczyński, *A linearization method for nonsmooth stochastic programming problems*, Mathematics of Operations Research **12** (1987), no. 1, 32–49.

[Rus08] ______, *A merit function approach to the subgradient method with averaging*, Optimisation Methods and Software **23** (2008), no. 1, 161–172.

[Rus21] A. Ruszczynski, *A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization*, SIAM Journal on Control and Optimization **59** (2021), no. 3, 2301–2320.

[RVV20] L. Rosasco, S. Villa, and B. C. Vũ, *Convergence of stochastic proximal gradient algorithm*, Applied Mathematics & Optimization **82** (2020), 891–917.

[RWC03] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric regression*, no. 12, Cambridge university press, 2003.

[Sch20] M. Schmidt, *Faster algorithms for deep learning? (presentation in vector institute: https://www.cs.ubc.ca/ schmidtm/documents/2020_ vector_ smallresidual.pdf)*, 2020.

[SHC$^+$17] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, *Evolution strategies as a scalable alternative to reinforcement learning*, arXiv preprint arXiv:1703.03864 (2017).

[SLA12] J. Snoek, H. Larochelle, and R. Adams, *Practical bayesian optimization of machine learning algorithms*, Advances in neural information processing systems, 2012, pp. 2951–2959.

[SLRB17] M. Schmidt, N. Le Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient*, Mathematical Programming **162** (2017), 83–112.

[SLWY15] W. Shi, Q. Ling, G. Wu, and W. Yin, *A proximal gradient algorithm for decentralized composite optimization*, IEEE Transactions on Signal Processing **63** (2015), no. 22, 6013–6023.

[SRB11] M. Schmidt, N. Roux, and F. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*, Advances in neural information processing systems **24** (2011).

[SS19] G. Scutari and Y. Sun, *Distributed nonconvex constrained optimization over time-varying digraphs*, Mathematical Programming **176** (2019), no. 1, 497–544.

[SSD22] Y. Sun, G. Scutari, and A. Daneshmand, *Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation*, SIAM Journal on Optimization **32** (2022), no. 2, 354–385.

[SSZ12] S. Shalev-Shwartz and T. Zhang, *Proximal stochastic dual coordinate ascent*, arXiv preprint arXiv:1211.2717 (2012).

[SSZ13] ______, *Stochastic dual coordinate ascent methods for regularized loss minimization.*, Journal of Machine Learning Research **14** (2013), no. 1.

[SSZ14] ______, *Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization*, International conference on machine learning, PMLR, 2014, pp. 64–72.

[SV09] T. Strohmer and R. Vershynin, *A randomized kaczmarz algorithm with exponential convergence*, Journal of Fourier Analysis and Applications **15** (2009), no. 2, 262.

[SYVS21] D. Sahu, J. Yao, M. Verma, and K. Shukla, *Convergence rate analysis of proximal gradient methods with applications to composite minimization problems*, Optimization **70** (2021), no. 1, 75–100.

[SZK19] A. K. Sahu, M. Zaheer, and S. Kar, *Towards gradient free and projection free stochastic optimization*, The 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 3468–3477.

[TLY+18] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, *d^2: Decentralized training over decentralized data*, International Conference on Machine Learning, PMLR, 2018, pp. 4848–4856.

[VBS19] S. Vaswani, F. Bach, and M. Schmidt, *Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron*, The 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 1195–1204.

[Ver18] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018.

[VML+19] S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien, *Painless stochastic gradient: Interpolation, line-search, and convergence rates*, Advances in Neural Information Processing Systems, 2019, pp. 3727–3740.

[WFL17] M. Wang, E. Fang, and H. Liu, *Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions*, Mathematical Programming **161** (2017), no. 1-2, 419–449.

[WGE17] G. Wang, G. B. Giannakis, and Y. C. Eldar, *Solving systems of random quadratic equations via truncated amplitude flow*, IEEE Transactions on Information Theory **64** (2017), no. 2, 773–794.

[WJZ+19] Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh, *Spiderboost and momentum: Faster variance reduction algorithms*, Advances in Neural Information Processing Systems **32** (2019).

[WL22] X. Wu and J. Lu, *A unifying approximate method of multipliers for distributed composite optimization*, IEEE Transactions on Automatic Control (2022).

[WLF16] M. Wang, J. Liu, and E. Fang, *Accelerating stochastic composition optimization*, Advances in Neural Information Processing Systems, 2016.

[WWW+16] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, *Learning structured sparsity in deep neural networks*, Advances in neural information processing systems **29** (2016).

[WYZY22] G. Wang, M. Yang, L. Zhang, and T. Yang, *Momentum accelerates the convergence of stochastic AUPRC maximization*, International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 3753–3771.

[WZC+21] Z. Wang, J. Zhang, T.-H. Chang, J. Li, and Z.-Q. Luo, *Distributed stochastic consensus optimization with momentum for nonconvex nonsmooth problems*, IEEE Transactions on Signal Processing **69** (2021), 4486–4501.

[XDKK21] R. Xin, S. Das, U. A. Khan, and S. Kar, *A stochastic proximal gradient framework for decentralized non-convex composite optimization: Topology-independent sample complexity and communication efficiency*, arXiv preprint arXiv:2110.01594 (2021).

[Xia09] L. Xiao, *Dual averaging method for regularized stochastic learning and online optimization*, Advances in Neural Information Processing Systems **22** (2009).

[XJY19] Y. Xu, R. Jin, and T. Yang, *Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems*, Advances in Neural Information Processing Systems **32** (2019).

[XKK21] R. Xin, U. A. Khan, and S. Kar, *An improved convergence analysis for decentralized online stochastic non-convex optimization*, IEEE Transactions on Signal Processing **69** (2021), 1842–1858.

[XTSS21] J. Xu, Y. Tian, Y. Sun, and G. Scutari, *Distributed algorithms for composite optimization: Unified framework and convergence analysis*, IEEE Transactions on Signal Processing **69** (2021), 3555–3570.

[XZSX15] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, *Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes*, 2015 54th IEEE Conference on Decision and Control (CDC), IEEE, 2015, pp. 2055–2060.

[YBL17] Z. Yang, K. Balasubramanian, and H. Liu, *High-dimensional non-gaussian single index models via thresholded score function estimation*, International conference on machine learning, PMLR, 2017, pp. 3851–3860.

[YSC19] A. Yurtsever, S. Sra, and V. Cevher, *Conditional gradient methods via stochastic path-integrated differential estimator*, Proceedings of the International Conference on Machine Learning-ICML 2019, 2019.

[YWF19] S. Yang, M. Wang, and E. Fang, *Multilevel stochastic gradient methods for nested composition optimization*, SIAM Journal on Optimization **29** (2019), no. 1, 616–659.

[YZLZ20] H. Ye, Z. Zhou, L. Luo, and T. Zhang, *Decentralized accelerated proximal gradient descent*, Advances in Neural Information Processing Systems **33** (2020), 18308–18317.

[ZBH$^+$16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning requires rethinking generalization*, arXiv preprint arXiv:1611.03530 (2016).

[ZCC$^+$18] D. Zhou, J. Chen, Y. Cao, Y. Tang, Z. Yang, and Q. Gu, *On the convergence of adaptive gradient methods for nonconvex optimization*, arXiv preprint arXiv:1808.05671 (2018).

[ZSM$^+$19] M. Zhang, Z. Shen, A. Mokhtari, H. Hassani, and A. Karbasi, *One sample stochastic frank-wolfe*, arXiv preprint arXiv:1910.04322 (2019).

[ZSM$^+$20] ———, *One-sample Stochastic Frank-Wolfe*, International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 4012–4023.

[ZX21] J. Zhang and L. Xiao, *Multilevel composite stochastic optimization via nested variance reduction*, SIAM Journal on Optimization **31** (2021), no. 2, 1131–1157.

[ZY18] J. Zeng and W. Yin, *On nonconvex decentralized gradient descent*, IEEE Transactions on signal processing **66** (2018), no. 11, 2834–2848.

[ZY19] J. Zhang and K. You, *Decentralized stochastic gradient tracking for non-convex empirical risk minimization*, arXiv preprint arXiv:1909.02712 (2019).