Title
Advances in Stochastic Optimization for Machine Learning

Author
Nguyen, Anthony

Publication Date
2021

Peer reviewed|Thesis/dissertation

Advances in Stochastic Optimization for Machine Learning

By

ANTHONY T. NGUYEN
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

APPLIED MATHEMATICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Krishnakumar Balasubramanian (Chair)

Shiqian Ma

Luis Rademacher

Committee in Charge

2021

To my parents for their sacrifices so I can have a better life.

Contents

Advances in Stochastic Optimization for Machine Learning

Abstract

We discuss two advances made in Stochastic Optimization where they arise out of a general problem, namely minimizing an objective function of the form $f(x) = \mathbb{E}_\xi[F(x,\xi)]$ for $x \in X \subseteq \mathbb{R}^n$, where $F(x,\xi)$ is a stochastic function with some random variable ξ.

The first project, in **Chapter** 2, deals with minimizing an objective function of the form $f_1 \circ \cdots \circ f_T(x)$ where $f_i(x) = \mathbb{E}_{\xi_i}[G_i(x,\xi_i)]$. In this setting, we assume that each component f_i is smooth, and in addition, we assume the access of a first-order oracle that outputs noisy estimates of the components and their derivatives. We introduce two algorithms that utilize moving average updates, and we prove that they converge to an ϵ-stationary point. The difference between these two algorithms is the first uses a mini-batch of samples in each iteration while the second uses linearized stochastic estimates of the function values. The sample complexities of the mini-batches and the stochastic linearized approaches for obtaining an ϵ-stationary point are $\mathcal{O}(\frac{1}{\epsilon^6})$ and $\mathcal{O}(\frac{1}{\epsilon^4})$, respectively.

The second project, in **Chapter** 3, discusses minimizing a convex function $f_0(x) = \mathbb{E}_{\xi_0}[F_0(x,\xi_0)]$ with functional inequality constraints $f_i(x) = \mathbb{E}_{\xi_i}[F_i(x,\xi_i)] \leqslant 0$ ($i \in \{1,\ldots,m\}$) using a zeroth-order oracle. We assume that we have access to noisy function value evaluations. The algorithm performs an extrapolation and numerically solves the dual optimization problem by performing a gradient ascent and descent at each iteration. Finally, the numerical solution is the weighted average of the iterates from the gradient descents. The number of calls to the oracle to find an ϵ-approximate optimal solution is $\mathcal{O}(\frac{(m+1)n}{\epsilon^2})$. Next, we present an algorithm in the non-convex setting based on [25]; utilizing our algorithm for the convex setting, the non-convex algorithm has sample complexity $\mathcal{O}(\frac{(m+1)n}{\epsilon^3})$.

Acknowledgments

This thesis would not be possible without the guidance of my advisor, Professor Krishna Balasubramanian. I was fortunate enough to work with Krishna on an interesting problem in Stochastic Optimization with many applications in machine learning. I would like to express my gratitude to him for introducing me to the field and providing constant attention and feedback to my research progress.

I would like to thank Professor Shiqian Ma and Professor Luis Rademacher for serving on my thesis committee and providing valuable feedback on my qualifying exam and dissertation. In addition, I would like to thank Professor Miles Lopes and Professor Debashis Paul for being on my qualifying exam committee. I would like to thank Miles Lopes again for suggesting that I contact Krishna when I was looking for a PhD advisor. I would like to thank Saeed Ghadimi for his help in my first PhD project. I would like to thank Stefan Wild for the work I have done on Zeroth Order Optimization at Argonne National Laboratory during Summer 2020.

I would like to thank the first floor administrative people: Tina, Victoria, and Sarah. I would like to thank my friends and colleagues at the Mathematics department for making my experience enjoyable. I would like to thank Sky for his friendship.

Finally, I want to thank my parents and family. Without their support, I would not be the person I am today. I am especially grateful to my mother who continues to show support, love, and care.

CHAPTER 1

Introduction

Optimization lies at the heart of machine learning. Most machine learning methods utilize optimization algorithms. When we train a machine learning model that makes predictions, so as a first step, we need to optimize its corresponding loss function - it tells us how much error is incurred from our training dataset. During training, one would optimize the model's parameters by gradient descent, an optimization algorithm that minimizes a convex, smooth function. In the machine learning community, there are many scientific papers on deep learning - learning based on a network of layered nodes. As part of the training process, backpropagation is used to compute the gradient of the neural network and is used to optimize the weights through mini-batch gradient descent.

So far, we have seen that Optimization plays a huge role in machine learning; in fact, it has a huge intersection in the training phase of a learning algorithm - minimizing a function. Now we turn our attention to a subfield of Optimization - Stochastic Optimization. This field refers to methods for minimizing or maximizing an objective function where randomness is present. One application of a stochastic optimization problem is least squares: minimizing $\|Ax-b\|_2^2$. To see this as a stochastic optimization problem, we can rewrite the least squares problem as $f_1 \circ f_2(x) = \|Ax - b\|_2^2$ where $f_1(x) = \|x\|_2^2$, $f_2(x) = Ax-b$, and these functions can be rewritten in the form $f_i(x) = \mathbb{E}_{\xi_i}[G_i(x, \xi_i)]$ for $i \in \{1, 2\}$ and random variables ξ_1, ξ_2.

In this dissertation, we examine two projects. The first project examines the T-level composition problem:

$$\min_{x \in X \subset \mathbb{R}^{d_T}} f_1 \circ f_2 \circ \cdots \circ f_T(x),$$

where $f_i(x) = \mathbb{E}_{\xi_i}[G_i(x, \xi_i)]$ with $f_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i-1}}$ for each $i \in \{1, 2, 3, \ldots, T\}$ $(d_0 = 1)$, and it proposes two stochastic algorithms to solve this problem. The second project examines the following objective function with functional inequality constraints:

$$\min_{x\in X\subset\mathbb{R}^n} f_0(x)$$
$$\text{s.t. } f_i(x)\leqslant 0, \quad i=1,\ldots,m,$$

where $i\in\{0,1,\ldots,m\}$, $f_i(x)=\mathbb{E}_{\xi_i}[F_i(x,\xi_i)]$, and it solves the problem in the convex and non-convex setting using zeroth-order information.

1.1. Soft introduction to Stochastic Optimization

We focus on a stochastic optimization problem of the form

$$\min_{x\in X\subseteq\mathbb{R}^n}\{f(x)=\mathbb{E}_\xi[F(x,\xi)]\} \tag{1.1}$$

since this closely resembles our two projects. We discuss solving this problem using a zero-order and a first-order oracle. Before doing so, we look at the following minimization problem

$$\min_{x\in X\subseteq\mathbb{R}^n} f(x) \tag{1.2}$$

where we do not know the form of f, in contrast to problem (1.1). We briefly discuss what zero-order and first-order Optimization means - both use zeroth-order and first-order oracles, respectively.

In first-order Optimization, a first-order oracle grants access to the function value $f(x)$ and its gradient $\nabla f(x)$ for each $x\in X\subseteq\mathbb{R}^n$. Therefore, we can use projected gradient descent method to find the minimizer. In the unconstrained case, our update is just gradient descent: $x_{k+1}=x_k-\lambda\nabla f(x_k)$ with $\lambda>0$.

For zeroth-order Optimization, we only have access to function values $f(x)$ for each $x\in X=\mathbb{R}^n$. We cannot use gradient information since we do not have access to such information, but we can approximate the gradient using Gaussian smoothing. According to [99], we form the Gaussian smoothing f_ν of f by the following formula:

$$f_\nu(x)=\mathbb{E}_u[f(x+\nu u)] \quad \text{where } u\sim\mathcal{N}(0,I_n) \quad \text{and} \quad \nu>0.$$

If we further assume that f is Lipschitz continuous, [99] says that

$$|f_\nu(x) - f(x)| \leqslant \nu L_f \sqrt{n} \quad \text{for } x \in X = \mathbb{R}^n,$$

where L_f is the Lipschitz constant of f - see [99] for other error bounds when f has a certain degree of smoothness. [99] showed that

$$\mathbb{E}_u \left[\frac{f(x+\nu u) - f(x)}{\nu} u \right] = \nabla f_\nu(x)$$

where $\nu > 0$ and

$$\|\nabla f_\nu(x) - \nabla f(x)\|_* \leqslant \frac{\nu}{2} L_{\nabla f}(n+3)^{3/2},$$

provided that f is continuously differentiable and has gradient lipschitz. Here, $\|\cdot\|_*$ denotes the dual norm. We refer interested readers to [99] for other types of error bound depending on the smoothness of f. We can use this gradient estimator and use gradient descent to minimize $f(x)$.

In the zeroth-order setting for (1.1), we have a zeroth-order oracle: for each $x \in X$, the stochastic oracle outputs $F(x, \xi)$ such that

$$\mathbb{E}[F(x, \xi)] = f(x).$$

Furthermore, an additional assumption in the stochastic oracle may include the following variance assumption:

$$\mathbb{E}[\|F(x, \xi) - f(x)\|_2^2] \leqslant \sigma_f^2. \tag{1.3}$$

As done previously, we need first-order information or some estimation of it, so we can come up with a method to solve this minimization problem. One approach is to use Gaussian smoothing on $F(x, \xi)$ for fixed ξ. Therefore, we have

$$\mathbb{E}_{u,\xi} \left[\frac{F(x+\nu u, \xi) - F(x, \xi)}{\nu} u \right] = \nabla f_\nu(x).$$

We add this biased derivative estimate to our zeroth-order oracle assumptions and add a variance assumption similar to (1.3). We can call this difference quotient expression $G(x, \xi, u)$ which is an unbiased estimator for $\nabla f_\nu(x)$ - see (3.1). From this, we can make the following variance assumption: $\mathbb{E}[\|G(x,\xi,u)-\nabla f_\nu(x)\|^2] \leq \sigma^2$ for some $\sigma \geq 0$. These quantities $F(x,\xi)$ and $G(x,\xi,u)$ would be our calls from the zeroth-order oracle, and we can perform a gradient descent method to solve the perturbed problem - minimizing $f_\nu(x)$ over $x \in X = \mathbb{R}^n$.

Returning to problem (1.1), we discuss the set up of the first-order method. Remark: In the first-order setting, we will use the notation $G(x,\xi)$ instead of $F(x,\xi)$ as our unbiased estimator of $f(x)$. Our first-order oracle will have the following setup: for each $x \in X \subseteq \mathbb{R}^n$, the stochastic oracle delivers a random variable and vectors $G(x,\xi)$ and $J(x,\xi)$ such that

$$\begin{aligned}
&\mathbb{E}[G(x,\xi)] = f(x),\\
&\mathbb{E}[J(x,\xi)] = \nabla f(x),\\
&\mathbb{E}[\|G(x,\xi) - f(x)\|^2] \leq \sigma_0^2,\\
&\mathbb{E}[\|J(x,\xi) - \nabla f(x)\|^2] \leq \sigma_1^2,
\end{aligned}$$

where σ_0, σ_1 are some non-negative constants, and we can perform a gradient descent using these random quantities to achieve an acceptable numerical solution.

Depending on whether our problem is convex or non-convex, the metric of convergence would measure how small $\mathbb{E}[f(x^k) - f(x^*)]$ is when f is convex and x^* minimizes f. In the non-convex setting, we examine the smallness of the quantity $\mathbb{E}[\|\nabla f(x^k)\|]$.

1.2. Introduction to Chapter 2: Optimizing nested functions

The research direction done in the first project - our contribution - started from [63]. In that research paper, the authors examined the following optimization problem:

$$\min_{x\in X} f_1 \circ f_2(x), \tag{1.4}$$

where $f_1(x) = \mathbb{E}_{\xi_1}[G_1(x,\xi_1)]$ and $f_2(x) = \mathbb{E}_{\xi_2}[G_2(x,\xi_2)]$ are Lipschitz smooth with G_1, G_2 being some stochastic functions, and X is a closed convex set. To solve this, an approach used in [63] utilizes a first-order oracle to obtain function and derivative estimates to our component functions f_1, f_2. Using these information, the authors use a gradient descent and make appropriate updates using moving averages; these updates include our approximate solutions, approximate gradients, and other approximations.

In the unconstrained setting and for simplicity, our convegence metric for this non-convex problem will measure in expectation how close our spatial iterates are to the local minimum and to the stational point - see [63] for more information.

Here comes our contribution: we numerically solve the optimization problem of the composition of $T \geqslant 3$ functions where our assumptions are similar to the two-level nested problem.

1.3. Introduction to T-level nested problem

We consider multi-level stochastic composition optimization problems of the form

$$\min_{x \in X} \Big\{ F(x) = f_1 \circ \cdots \circ f_T(x) \Big\}, \tag{1.5}$$

where $f_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i-1}}$ for $i = 1, \ldots, T$ ($d_0 = 1$) are continuously differentiable function and X is a closed convex set. We assume that the exact values and derivatives of f_i's are not available. In particular, we assume that $f_i(x) = \mathbb{E}_{\xi_i}[G_i(x, \xi_i)]$ for some random variables $\xi_i \in \mathbb{R}^{\tilde{d}_i}$.

Note that when $T = 1$, the problem reduces to the standard stochastic optimization problem which has been well-explored in the literature; see, for example [27, 62, 79, 106, 115], for a partial list. In this work, we consider stochastic first-order algorithms for solving eq. (1.5) when $T \geq 1$. Note that the gradient of the function $F(x)$ in eq. (1.5), has the form $\nabla F(x) = \nabla f_T(y_T) \nabla f_{T-1}(y_{T-1}) \cdots \nabla f_1(y_1)$, where $y_i = f_{i+1} \circ \cdots \circ f_T(x)$ for $1 \leq i < T$ and $y_T = x$. Our goal is to solve the above optimization problem, given access to noisy evaluations of ∇f_i's and f_i's. Precise assumptions on our stochastic first-order oracle considered will be stated later in section 2.1. Because of the nested nature of the gradient $\nabla F(x)$, obtaining an unbiased gradient estimator in the online setting, with controlled higher moments, becomes non-trivial.

Although problems of the form in eq. (1.5) have been considered since the work of [51], recently there has been a renewed interest on this problem due to applications arising in mathematical finance, nonparametric statistics, deep generative modeling and reinforcement learning. We refer the reader to [22, 26, 38, 52, 63, 100, 125, 126, 129, 131] for such applications and various algorithmic approaches for solving problem eq. (1.5). In particular [125] and [129] considered the case of

$T = 2$ and general T respectively, and analyzed stochastic gradient-type algorithms. Such an approach leads to level-dependent and sub-optimal convergence rates. However, large deviation and Central Limit Theorem results established in [52] and [38], respectively, show that in the sample-average or empirical risk minimization setting, the arg min of the problem in eq. (1.5) based on n samples, converges at a level-independent rate (i.e., the target accuracy is independent of T) to the true minimizer, under suitable regularity conditions. Hence, it is natural to ask the following question: *Is it possible to construct iterative online algorithms for solving problem eq.* (1.5) *with level-independent convergence rates?* Recently, for the case of $T = 2$, [63] proposed a single time-scale Nested Averaged Stochastic Approximation (NASA) algorithm with complexities matching the case of $T = 1$. This resolved the above question for $T = 2$. However, constructing similar algorithms for the case of general T had remained less investigated.

Main contributions. In this work, we propose two algorithms for solving problem (1.5) with level-independent convergence rates in the stochastic first-order oracle setting, under mild assumptions. Our complexity results are summarized in Table 1.1. The first algorithm is based on an extension of the NASA algorithm from [63] (proposed for the case of $T = 2$) to the general $T \geq 1$ setting, requiring a mini-batch of sample in each iteration. Although this algorithm has level-independent convergence rates, the sample complexity (i.e., the number of calls to stochastic first-order oracle) does not match that of standard stochastic gradient algorithm for $T = 1$ or the NASA algorithm for $T = 2$. The second algorithm is based on a modification to the NASA algorithm, motivated by the standard linearization technique [37, 45, 109, 110], mainly used for non-smooth problems. For any $T \geq 1$, we show that this algorithm has the same oracle complexity as that of the regular stochastic gradient algorithm for the case of $T = 1$, thereby providing a complete answer to the question above. We emphasize that unlike our first algorithm, this algorithm does not require a mini-batch of samples in any iteration and hence is more suitable to the online setting.

Comparisons to related works. A summary of our results, in comparison to the most related work of [129] is provided in Table 1.1. We remark that the approach and the results in [129] are provided only for the unconstrained setting. We also highlight the related work of [131] which considered problems of the form $\min_{x \in \mathbb{R}^{d_T}} \{F(x) + H(x)\}$, with $F(x)$ being a multi-level composite function as in eq. (1.5) and $H(x)$ being a convex and lower-semi-continuous function. Typically $H(x)$ could be considered as an indicator function of the constrained set X to relate the above problem to our setup in eq. (1.5). The algorithm proposed in [131] is a proximal variant of SPIDER variance reduction technique [53] and is a double-loop algorithm. Hence, it is predominantly applicable for finite-sum problems and is not so suitable for the general online problems that we focus on. Indeed, they assume 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. Furthermore, [131] assume a much stronger mean-square Lipschitz smoothness assumption on the individual functions f_i and their gradients, to obtain a complexity bound of $\mathcal{O}\left(T^6\rho^T/\epsilon^3\right)$, where ρ is a problem dependent constant factor. Furthermore, to obtain their result, they also need a mini-batch of samples, with batch sizes of the order $T^3\rho^T$, which makes their approach impractical to be used even for moderately large values of T. As mentioned above, our second algorithm does not have any such requirements, making it easy to be practically applicable for large values of T.

Furthermore, our Algorithm 2 is similar to the one proposed more recently in [110] for multi-level composition optimization. In his work, the author focuses on the nonsmooth case and provides asymptotic convergence of the proposed algorithm to a stationary point of the problem by analyzing a system of differential inclusions which requires the compactness of the feasible set X. The finite-time convergence analysis however, from our communication with the author, is not complete in the released manuscript. Hence we are not able to provide a detailed comparison of the sample complexities and assumptions on the oracle. We also remark that our choice of Lyapunov function in (2.15) is different from that used in [110], which makes an important part of our convergence analysis, distinct. This enables us, unlike [110], to relax the boundedness assumption of the feasible set thereby making our method applicable to the unconstrained applications as well.

1.3.1. Motivating Application. We now discuss a concrete motivating application for the T-level stochastic composition optimization problem we consider in this work. Let $x^* \in \mathbb{R}^d$ denote an unknown signal that we wish to recover. Suppose we are allowed to observe measurements of the form $y = a^\top x^* + \epsilon$, where $a \in N(0, I_d)$ is the random measurement vector and $\epsilon \sim N(0, 1)$ (for

Method	Convergence Rate	Oracle Complexity
[129]	$\mathcal{O}\left(N^{-4/(7+T)}\right)$	$\mathcal{O}\left(1/\epsilon^{(7+T)/2}\right)$
Algorithm 1	$\mathcal{O}\left(N^{-1/2}\right)$	$\mathcal{O}\left(1/\epsilon^{6}\right)$
Algorithm 2	$\mathcal{O}\left(N^{-1/2}\right)$	$\mathcal{O}\left(1/\epsilon^{4}\right)$

TABLE 1.1. Convergence rates and Oracle complexity results for finding an ϵ-pair $\bar{x}, \bar{z}$ of eq. (1.5); see Definition 2.1.1 for details. Convergence rate refers to the upper bound on $\mathbb{E}[V(x,z)]$ and oracle complexity refers to the number of calls to the stochastic first order oracle to obtain a ϵ-pair. Here, we only present the ϵ-related T dependencies. See Remark 2.1.1 and Remark 2.2.1 for more details.

simplicity) is the noise in the measurement. In this case, the following estimator,

$$\check{x} = \underset{x \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}(y - a^\top x)^2,$$

that minimizes the expected reconstruction error servers as good estimator of the true signal. This is indeed a single-level stochastic optimization problem. To actually get the minimizer, one could run the standard stochastic gradient algorithm for N iterations with a single sample $(y_i, a_i) \in \mathbb{R}^{d+1}$ in each iteration. Without further assumptions on x^*, we require $N \approx d$ to accurately estimate x^* [103, 108]. In compressed sensing [31, 42], the signal x^* is assumed to be k-sparse, i.e., it is assumed to consist of only k non-zero entries. Denote by $\|\cdot\|_0$, L_0 norm of a vector counting the number of non-zero coordinates of the vector. Then, under the sparsity assumption, for the stochastic gradient algorithm, to solve the following problem,

$$\bar{x} = \underset{x \in \mathbb{R}^d : \|x\|_0 \leq k}{\arg\min}\ \mathbb{E}(y - a^\top x)^2,$$

it is enough to require $N \approx k \log d$ (as opposed to $N \approx d$) samples for accurate reconstruction [3, 4]. Hence, when $k \ll d$, we get a huge improvement in terms of oracle complexity. Furthermore, real-world signals, like images, are empirically observed to satisfy the sparsity assumption stated above. Hence, the field of compressed sensing has revolutionized the field of signal processing [24, 49, 124].

Recently, motivated by the success of deep learning, [26] proposed a generative approach to compressed sensing. Here, it is assumed that there is a latent signal vector $z^* \in \mathbb{R}^k$, with $k \ll d$, such that for a given neural network $G : \mathbb{R}^k \to \mathbb{R}^d$, the true signal is given by $x^* = G(z^*)$. In other words, the true signal is assumed to lie in the range of a neural network, given the latent signal z^*. Similar to above, we are allowed to observe measurements of the form $y = a^\top G(z^*) + \epsilon$. In this case, the following estimator,

$$\bar{x} = \underset{z \in \mathbb{R}^k}{\arg\min}\ \mathbb{E}(y - a^\top G(z))^2,$$

was proposed in [26]; see also [70, 100, 128] for more details. Furthermore, the mapping G is assumed to be deep neural network with depth T'. That is, $G(z) = f_1 \circ f_2 \cdots \circ f_{T'}(z)$, where for $1 \le i \le T'$, the function $f_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$, with $d_{T'} = k$ and $d_1 = d$. Here, each component of the function $[f_i]_{j_i}$ for $1 \le i \le T'$ is given by

$$[f_i]_{j_i}(y) = \mathbb{E}_{p(g,b)}[\sigma(g^\top y - b)]$$

where $\sigma(s)$ is the activation function and $p(g, b) \in \mathbb{R}^{d+1}$ is a distribution over the weight and the bias at each layer. Typically the activation function is the ReLU function $\sigma(s) := \max\{0, s\}$ or the sigmoidal function $\sigma(s) := 1/(1 + e^{-s})$ and the distribution $p(g, b)$ is typically assumed to be Gaussian. Hence, the problem is a special case of the T-stage stochastic composite optimization problem outlined in (1.5). The statistical sample complexity of the above problem, for accurate reconstruction, requires the number of measurement to be of the order of k [26]. However, efficient algorithms for solving the above problem are less explored; see [70, 118] for some related works. Our proposed algorithms in this work, could potentially be used to solve the above problem efficiently – a thorough investigation is beyond the scope of the current project, however is interesting future work. It is worth emphasizing that, in the case of ReLU activation function, our smoothness assumptions are not immediately satisfied. However, it is possible to construct accurate and smooth approximations to ReLU functions, that satisfy our assumptions.

The rest of our project is organized as follows. In section 2.1, we present our first algorithm and analyze its convergence analysis for solving eq. (1.5) with any $T \ge 1$. In section 2.1, we present a modification of this algorithm and show that it can recover the best-known sample complexity for (single-level) smooth stochastic optimization. Some concluding remarks are also given in section 2.3.

1.4. Introduction to Chapter 3: Zeroth order Optimization with functional inequality constraints

Before we discuss our next contribution, we discuss the following optimization problem of minimizing an objective function with functional inequality constraints:

$$\begin{aligned} &\min_{x \in X \subseteq \mathbb{R}^n} f_0(x) \\ &\text{s.t. } f_i(x) \leqslant 0, \quad i = 1, \ldots, m; \end{aligned} \tag{1.6}$$

where f_i is smooth for each $i \in \{0, 1, \ldots, m\}$ and X is a convex compact set. In [25], this stochastic optimization problem is solved using a first-order oracle in the smooth convex and nonconvex setting; the authors optimize the dual problem of (1.6). Hence, the algorithm in the convex setting performs a gradient ascent and descent in each iteration. At the end of the algorithm, a weighted average of the iterates from the gradient descent is set as the numerical solution to the convex problem.

In terms of the convergence analysis metric, we examine the following quantities

$$\mathbb{E}[f_0(\bar{x}_T) - f_0^*] \quad \text{and} \quad \mathbb{E}[\|[\max(f_1(\bar{x}_T), 0), \ldots, \max(f_m(\bar{x}_T), 0)]^\top\|_2],$$

where T is the number of iterations of the algorithm, and the authors show these quantities are at most ϵ for a certain sample complexity $T := T_\epsilon$. This algorithm in [25] (known as the ConEx method) is used in the non-convex setting.

The idea in the non-convex setting is to augment the objective and contraint functions with strongly convex terms and use the ConEx method to numerically solve the augmented problem $K \geqslant 1$ times. Next, randomly pick $\hat{k} \in \{1, \ldots, K\}$ (where each has uniform chance of being picked), and return $x_{\hat{k}}$ as the numerical solution to the non-convex problem.

The convergence rate measures how much in expectation the numerical solution deviates from the KKT conditions - see [25]. The total number of calls to the first-order oracle to solve the non-convex setting is $\mathcal{O}(T_\epsilon/\epsilon)$. These two methods lead to our contribution: in the convex and non-convex setting, we solve these optimization problems using a zeroth-order oracle; the sample complexities are $\mathcal{O}((m+1)n/\epsilon^2)$ and $\mathcal{O}((m+1)n/\epsilon^3)$, respectively.

1.5. Introduction to optimization of an objective function with funtional constraints in Zeroth Order setting

We develop and analyze stochastic zeroth-order algorithms for solving the following non-linear optimization problem with functional constraints:

$$\begin{aligned} &\min_{x \in X} f_0(x) \\ &\text{s.t. } f_i(x) \leqslant 0, \quad i = 1, \ldots, m, \end{aligned} \tag{1.7}$$

where, for $i \in \{0, 1, \ldots, m\}$, $f_i(x) = \mathbb{E}_{\xi_i}[F_i(x, \xi_i)] : \mathbb{R}^n \to \mathbb{R}$, are continuous functions which are not necessarily convex, ξ_i is the noise vector associated with function f_i, and $X \subseteq \mathbb{R}^n$ is a convex compact set. In the stochastic zeroth-order setting, we neither observe the objective function f_0 nor the constraint functions f_i directly. We only have access to noisy function evaluations of them. Such zeroth-order optimization algorithms have been successfully applied to a diverse set of fields including culinary engineering [117], chemical engineering [69] and water-plant treatment [66, 91]. Within the field of statistical machine learning, such algorithms have proved to be useful for hyperparameter tuning [116] (see [65] for an overview of how *Google* uses such algorithms for hyperparameter tuning for their products), reinforcement learning [34, 56, 90, 112] and robotics [75, 76].

The study of stochastic zeroth-order optimization algorithms for unconstrained optimization problems goes back to the early works of [23, 73, 77, 97, 98, 105, 119, 121]. However, the study of zeroth-order algorithms and their oracle complexities for constrained problem as in (1.7) is limited (see Section 1.5.2 for details), despite the fact that several real-world machine learning problems fall under the setting of (1.7) (see Sections 1.5.1). This serves as our main motivation for developing stochastic zeroth-order optimization algorithms for solving (1.7), and analyzing their oracle complexity.

Our methodology is based on the recently proposed constrained extrapolation based primal-dual approach in [25] for the stochastic first-order setting. In this work, we extend this methodology to the stochastic zeroth-order setting based on Gaussian smoothing based zeroth-order stochastic gradient estimators. We characterize the precise way to set the *tuning parameters* of the algorithm so as to mitigate the issues caused by the *bias* in the stochastic zeroth-order gradient estimates.

Based on this, we demonstrate that for the case when the functions f_i, $i = 0, \ldots, m$, are convex, the number of calls to the stochastic zeroth-order oracle to achieve an appropriately defined ϵ-optimal solution for (1.7) is of order $\mathcal{O}((m+1)n/\epsilon^2)$. Furthermore, in the nonconvex setting, the number of calls to obtain an appropriately defined ϵ-optimal KKT solution of (1.7) is of order $\mathcal{O}((m+1)n/\epsilon^3)$. To our knowledge, these are the first non-asymptotic oracle complexity result for stochastic zeroth-order optimization with stochastic zeroth-order functional constraints. We illustrate the practical applicability of the developed methodology by testing its performance on benchmark simulation experiments for functionally constrained optimization problems, and a hyperparameter tuning problem which we discuss below.

1.5.1. Motivating Application. Our main motivation for studying constrained optimization problems in the zeroth-order setting is their applicability to hyperparameter tuning for machine learning algorithms. We refer the interested reader, for example, to [59, 71, 107, 116] for more details. Automating the process of selecting the optimal hyperparameters is crucial for making statistical machine learning methods widely applicable in practice.

In this work, we specifically concentrate on tuning the parameters of Hybrid or Hamiltonian Monte Carlo (HMC) sampling algorithm. HMC, proposed by [44], and popularized in the statistical machine learning community by [96] is a gradient-based sampling algorithm that works by discretizing the continuous time degenerate Langevin diffusion [82]. It has been used successfully as a state-of-the art sampler or a numerical integrator in the Bayesian statistical machine learning community [32, 33, 64, 72, 127]. However, in order to obtain successful performance in practice using HMC, several hyperparameters need to be tuned optimally.

Typically, the functional relationship between the hyperparameters that need to be tuned and the performance measure used is not available in an analytical form. We can only evaluate the performance of the sampler for various settings of the hyperparameter. Furthermore, in practice several constraints, for example, constraints on running times and constraints that enforce the generated samples to pass certain standard diagnostic tests [60, 61], are enforced in the hyperparameter tuning process. The functional relationship between such constraints and the hyperparameters is also not available analytically. This makes the problem of optimally setting the hyperparameters for HMC as a constrained zeroth-order optimization problem. In the context of HMC, [59, 71, 89] used Bayesian optimization techniques to set the hyperparameters.

1.5.2. Related works. The methodology developed for zeroth-order optimization in the operations research and statistics communities has a long and illustrious history to be summarized entirely. Similarly, in the machine learning community, Bayesian optimization techniques have been developed for optimizing functions with only noisy function evaluations. We refer the reader to [7, 13, 28, 35, 55, 78, 81, 86, 92, 93, 113, 120] for more details. In what follows, we focus on relevant literature from zeroth-order optimization and Bayesian optimization literature for *constrained optimization* problems.

When the constraint set is analytically available and only the objective function is not, [84] and [29] considered an augmented Lagrangian approach and an inexact restoration method respectively, and provided convergence analysis. Furthermore, [5, 14, 78] extended the popular mesh adaptive direct search to this setting. Projection-free methods based on Frank-Wolfe methods, have been considered in [19, 111] for the case when the constraint set is a convex subset of $\mathbb{R}^n$. Furthermore, [85] considered the case when the constraint set is a Riemannian submanifold embedded in $\mathbb{R}^n$ (and the function is defined only over the manifold).

When the objective function f and the constraint functions f_i, $i = 1, \ldots, m$ are both not available analytically, the methodology and the related analysis becomes relatively complicated. For this case, in the deterministic setting (i.e., we could obtain exact evaluations of the objective and the constraint functions at a given point), *filter methods* which reduce the objective function while trying to reduce constraint violations were proposed and analyzed in [10, 48, 104]. Barrier method in the zeroth-order setting was considered in [11, 12, 47, 54, 54, 67, 87, 88], with some works also developing line search approaches for setting the tuning parameters. Model based approaches were considered in the works of [16, 36, 66, 95, 123]. Furthermore, [15, 30] developed extensions of Nelder–Mead algorithm to the constrained setting. Several works in the statistical machine learning community also considered Bayesian optimization methods in the constrained setting, in both the noiseless and noisy seeting. We refer the reader, for example, to [1, 8, 17, 18, 50, 57, 58, 68, 71, 80, 83, 102]. On one hand, the above works demonstrate the interest in the optimization and machine learning communities for developing algorithms for constrained zeroth-order optimization problems. On the other hand, most of the above works are not designed to handle *stochastic* zeroth-order constrained optimization that we consider. Furthermore, a majority of the above works are methodological, and the few works that develop convergence analysis do so only in the asymptotic setting. To the best of our knowledge, there is no rigorous non-asymptotic analysis of the oracle complexity of zeroth-optimization when the constraints and the objective values are available only via *noisy* function evaluations.

CHAPTER 2

Stochastic Multi-level Composition Optimization Algorithms with Level-Independent Convergence Rates

2.1. Multi-level Nested Averaging Stochastic Gradient Method

In this section, we present our first algorithm for solving problem (1.5). As mentioned in section 1.3, the previously proposed stochastic gradient-type methods suffer in terms of the convergence rates when applied for solving this problem [129]. The main reason is the increased bias when estimating the stochastic gradient of F, for $T \geq 2$. Our proposed algorithm has a multi-level structure – in addition to estimating the gradient of F, we also estimate the values of inner functions f_i by a mini-batch moving average technique, extending the approach in [63] for any $T > 1$. This will enable us to provide an algorithm with improved convergence rates to the stationary points compared to the prior work [129]. Our approach is formally presented in Algorithm 1.

We now add a few remarks about Algorithm 1. First, note that at each iteration of this algorithm, we update the triple $(x^k, \{w^k\}_{i=1}^T, z^k)$, which are the convex combinations of the solutions to subproblem (2.1), the estimates of inner function values f_i, and the stochastic gradient of F at these points, respectively. It should be mentioned that we do not need to estimate the values of the outer function f_1. However, we include w_1^k in for the sake of completeness. Second, when $T = 2$ and $b_k = 1$, this algorithm reduces to the NASA algorithm presented in [63]. Indeed, Algorithm 1 is a direct generalization of the NASA method to the multi-level case $T \geq 3$. However, to prove convergence of Algorithm 1, we need to take a batch of samples in each iteration to reduce the noise associated with estimation of the inner function values, when $T > 2$. We now provide our convergence analysis for Algorithm 1. To do so, we define the following filtration,

$$\mathscr{F}_k := \sigma(\{x^0, \ldots, x^k, z^0, \ldots, z^k, w_1^0, \ldots, w_1^k, \ldots, w_T^0, \ldots, w_T^k, u^0, \ldots, u^k\}).$$

Algorithm 1 Multi-level Nested Averaging Stochastic Gradient Method

Input: Positive integer sequence $\{b_k\}_{k\geq 0}$ and initial points $x^0, z^0 \in X, \quad w_i^0 \in \mathbb{R}^{d_i} \quad 1 \leq i \leq T$,
for $k = 0, 1, 2, \dots,$ **do**

1. Compute

$$u^k = \arg\min_{y \in X} \left\{ \langle z^k, y - x^k \rangle + \frac{\beta_k}{2} \|y - x^k\|^2 \right\}, \tag{2.1}$$

stochastic gradients J_i^{k+1}, and function values $G_{i,j}^{k+1}$ at w_{i+1}^k for $i = \{1, \dots, T\}, j = \{1, \dots, b_k\}$ by denoting $w_{T+1}^k \equiv x^k$.

2. Set

$$x^{k+1} = (1 - \tau_k) x^k + \tau_k u^k, \tag{2.2}$$

$$z^{k+1} = (1 - \tau_k) z^k + \tau_k \prod_{i=1}^{T} J_{T+1-i}^{k+1}, \tag{2.3}$$

$$w_i^{k+1} = (1 - \tau_k) w_i^k + \tau_k \bar{G}_i^{k+1}, \qquad 1 \leq i \leq T, \tag{2.4}$$

where

$$\bar{G}_i^{k+1} = \frac{1}{b_k} \sum_{j=1}^{b_k} G_{i,j}^{k+1}. \tag{2.5}$$

Output:

Next, we state our main assumptions on the individual functions and the stochastic first-order oracle we use.

Assumption 1. *All functions $f_1, \dots, f_T$ and their derivatives are Lipschitz continuous with Lipschitz constants L_{f_i} and $L_{\nabla f_i}$, respectively.*

Assumption 2. *Denote $w_{T+1}^k \equiv x^k$. For each k, w_{i+1}^k being the input, the stochastic oracle outputs $G_i^{k+1} \in \mathbb{R}^{d_i}$ and $J_i^{k+1} \in \mathbb{R}^{d_i \times d_{i-1}}$ such that*

(1) $\mathbb{E}[J_i^{k+1} | \mathscr{F}_k] = [\nabla f_i(w_{i+1}^k)]^\top$, and $\mathbb{E}[G_i^{k+1} | \mathscr{F}_k] = f_i(w_{i+1}^k)$, for $1 \leq i \leq T$.

(2) $\mathbb{E}[\|G_i^{k+1} - f_i(w_{i+1}^k)\|^2 | \mathscr{F}_k] \leq \sigma_{G_i}^2$, and $\mathbb{E}[\|J_i^{k+1}\|^2 | \mathscr{F}_k] \leq \sigma_{J_i}^2$, for $1 \leq i \leq T$. Here $\|\cdot\|$ is any vector or matrix norm. For concreteness the reader could view them as the standard Euclidean norm (for vectors) and the operator norm (for matrices).

(3) Given $\mathscr{F}_k$, the outputs of the stochastic oracle at each level i, G_i^{k+1} and J_i^{k+1}, are independent.

(4) Given $\mathscr{F}_k$, 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 $\{J_i^{k+1}\}_{i=1,\dots,T}$.

Assumption 1 is a standard smoothness assumption made in the literature on nonlinear optimization. Similarly, Parts 1 and 2 in Assumption 2 are standard unbiasedness and bounded variance assumptions on the stochastic gradient, common in the literature. At this point, we re-emphasize that the assumptions made in [131] are stronger than our assumptions above, as they require mean-square smoothness of the individual random functions G_i and their gradients. Parts 3 and 4 are also essential to establish the converge results in the multi-level case; similar assumptions have been made, for example, in [129]. In the next couple of technical results, we provide some properties of composite functions that are required for our subsequent results.

Lemma 2.1.1. Define $F_i(x) = f_i \circ f_{i+1} \circ \cdots f_T(x)$. Under Assumption 1, the gradient of F_i is Lipschitz continuous with constant

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

Proof. We show the result by backward induction. Under Assumption 1, gradient of $F_T = f_T$ is Lipschitz continuous and so does that of F_{T-1} since for any $x, y \in X$, we have

$$\begin{aligned} \|\nabla F_{T-1}(x) - \nabla F_{T-1}(y)\| &= \|\nabla f_T(x) \nabla f_{T-1}(f_T(x)) - \nabla f_T(y) \nabla f_{T-1}(f_T(y))\| \\ &\leq \|\nabla f_T(x)\| \|\nabla f_{T-1}(f_T(x)) - \nabla f_{T-1}(f_T(y))\| \\ &+ \|\nabla f_{T-1}(f_T(y))\| \|\nabla f_T(x) - \nabla f_T(y)\| \\ &\leq (L_{f_T}^2 L_{\nabla f_{T-1}} + L_{f_{T-1}} L_{\nabla f_T}) \|x - y\|. \end{aligned}$$

Now, suppose that gradient of F_{i+1} is Lipschitz continuous for any $i \leq T-1$. Then, similar to the above relation, ∇F_i is Lipschitz continuous with constant

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

■

We remark that the above result has also been proved in [131], Lemma 5.2., with a slightly different proof.

Lemma 2.1.2. Define $F_i(x) = f_i \circ f_{i+1} \circ \cdots f_T(x)$ and $\nabla \bar{f}_i(x) = \nabla f_T(x) \nabla f_{T-1}(w_T) \cdots \nabla f_i(w_{i+1})$ for any $x \in X, w_j \in \mathbb{R}^{d_j}\quad j = i+1, \dots, T$. Then under assumption 1, we have

$$\|\nabla F_i(x) - \nabla \bar{f}_i(x)\| \leq \sum_{j=i}^{T-1} \frac{L_{\nabla f_j}}{L_{f_j}} L_{f_i} \cdots L_{f_T} \|F_{i+1}(x) - w_{j+1}\|.$$

Proof. We show the result by backward induction. The case $i = T$ is trivial.
When $i = T - 1$, under Assumption 1, we have

$$\begin{aligned}\|\nabla F_{T-1}(x) - \nabla f_T(x) \nabla f_{T-1}(w_T)\| &= \|\nabla f_T(x)[\nabla f_{T-1}(f_T(x)) - \nabla f_{T-1}(w_T)]\| \\ &\leq L_{\nabla f_{T-1}} L_{f_T} \|f_T(x) - w_T\|.\end{aligned}$$

Now assume that for any $i \leq T-2$,

$$\|\nabla F_{i+1}(x) - \nabla \bar{f}_{i+1}(x)\| \leq \sum_{j=i+1}^{T-1} \frac{L_{\nabla f_j}}{L_{f_j}} L_{f_{i+1}} \cdots L_{f_T} \|F_{j+1}(x) - w_{j+1}\|.$$

We then have

$$\begin{aligned}&\|\nabla F_i(x) - \nabla \bar{f}_i(x)\| = \|\nabla F_{i+1}(x) \nabla f_i(F_{i+1}(x)) - \nabla \bar{f}_i(x)\| \\ &\leq \|\nabla f_i(F_{i+1}(x))\| \|\nabla F_{i+1}(x) - \nabla \bar{f}_{i+1}(x)\| + \|\nabla \bar{f}_{i+1}(x)\| \|\nabla f_i(F_{i+1}(x)) - \nabla f_i(w_{i+1})\| \\ &\leq L_{f_i} \|\nabla F_{i+1}(x) - \nabla \bar{f}_{i+1}(x)\| + L_{\nabla f_i} L_{f_{i+1}} \cdots L_{f_T} \|F_{i+1}(x) - w_{i+1}\| \\ &\leq L_{f_i} \sum_{j=i+1}^{T-1} \frac{L_{\nabla f_j}}{L_{f_j}} L_{f_{i+1}} \cdots L_{f_T} \|F_{j+1}(x) - w_{j+1}\| + L_{\nabla f_i} L_{f_{i+1}} \cdots L_{f_T} \|F_{i+1}(x) - w_{i+1}\| \\ &= \sum_{j=i}^{T-1} \frac{L_{\nabla f_j}}{L_{f_j}} L_{f_i} \cdots L_{f_T} \|F_{j+1}(x) - w_{j+1}\|.\end{aligned}$$

■

Lemma 2.1.3. Under Assumption 1, for any $j \in \{1, \ldots, T-1\}$, we have

$$\|f_j \circ \cdots \circ f_T(w_{T+1}) - w_j\| \leq \|f_j(w_{j+1}) - w_j\| + \sum_{\ell=j+1}^{T} \left(\prod_{i=j}^{\ell-1} L_{f_i}\right) \|f_\ell(w_{\ell+1}) - w_\ell\|.$$

Proof. We show the results by backward induction. For $j = T-1$, we have

$$\begin{aligned}
\|f_{T-1} \circ f_T(w_{T+1}) - w_{T-1}\| &\leq \|f_{T-1} \circ f_T(w_{T+1}) - f_{T-1}(w_T)\| + \|f_{T-1}(w_T) - w_{T-1}\| \\
&\leq L_{f_{T-1}} \|f_T(w_{T+1}) - w_T\| + \|f_{T-1}(w_T) - w_{T-1}\|.
\end{aligned}$$

Now suppose the result holds for $j+1, j \in \{1, \ldots, T-2\}$. Then, we have

$$\begin{aligned}
\|f_j \circ f_{j+1} \circ \cdots f_T(w_{T+1}) - w_j\| &\leq \|f_j \circ \cdots f_T(w_{T+1}) - f_j(w_{j+1}) + f_j(w_{j+1}) - w_j\| \\
&\leq L_{f_j} \|f_{j+1} \circ \cdots \circ f_T(w_{T+1}) - w_{j+1}\| + \|f_j(w_{j+1}) - w_j\| \\
&\leq L_{f_j} \left[\|f_{j+1}(w_{j+2}) - w_{j+1}\| + \sum_{\ell=j+2}^{T} \left(\prod_{i=j+1}^{\ell-1} L_{f_i}\right) \|f_\ell(w_{\ell+1}) - w_\ell\| \right] \\
&+ \|f_j(w_{j+1}) - w_j\| \\
&= \|f_j(w_{j+1}) - w_j\| + \sum_{\ell=j+1}^{T} \left(\prod_{i=j}^{\ell-1} L_{f_i}\right) \|f_\ell(w_{\ell+1}) - w_\ell\|,
\end{aligned}$$

where the third inequality follows by induction hypothesis. ■

Lemma 2.1.4. Define

$$\begin{aligned}
&R_1 = L_{\nabla f_1} L_{f_2} \cdots L_{f_T}, \\
&R_j = L_{f_1} \ldots L_{f_{j-1}} L_{\nabla f_j} L_{f_{j+1}} \cdots L_{f_T} \quad 1 < j \leq T-1, \\
&C_2 = R_1, \\
&C_j = R_1 L_{f_2 \circ \cdots \circ f_{j-1}} + R_2 L_{f_3 \circ \cdots \circ f_{j-1}} + \cdots + R_{j-2} L_{f_{j-1}} + R_{j-1} \text{ with } 2 < j \leq T.
\end{aligned}$$

Assume that Assumption 1 holds. Then for $T \geq 3$,

$$\left\| \nabla F(x) - \nabla f_T(x) \prod_{i=2}^{T} \nabla f_{T+1-i}(w_{T+2-i}) \right\| \leq \sum_{j=2}^{T-1} C_j \|f_j(w_{j+1}) - w_j\| + C_T \|f_T(x) - w_T\|. \quad (2.6)$$

Proof. By lemma 2.1.2 and lemma 2.1.3, we have

$$\left\|\nabla F(x) - \nabla f_T(x) \prod_{i=2}^{T} \nabla f_{T+1-i}(w_{T+2-i})\right\| \leq \sum_{j=1}^{T-1} R_j \|f_{j+1} \circ \cdots \circ f_T(w_{T+1}) - w_{j+1}\|$$

$$= \sum_{j=1}^{T-2} R_j \|f_{j+1} \circ \cdots \circ f_T(w_{T+1}) - w_{j+1}\| + R_{T-1}\|f_T(w_{T+1}) - w_T\|$$

$$= \sum_{j=1}^{T-2} R_j \|f_{j+1}(w_{j+2}) - w_{j+1}\| + \sum_{j=1}^{T-2} R_j \sum_{\ell=j+2}^{T} \left(\prod_{i=j+1}^{\ell-1} L_{f_i}\right) \|f_\ell(w_{\ell+1}) - w_\ell\|$$

$$+ R_{T-1}\|f_T(w_{T+1}) - w_T\|.$$

The conclusion follows. To see this, term collecting $\|f_2(w_3) - w_2\|$, we have C_2.
For $2 < j \leq T$, term collecting $\|f_j(w_{j+1}) - w_j\|$, we have C_j. ∎

The following result also shows the Lipschitz continuity of the objective function of the subproblem (2.1). One can see [63] for a simple proof.

Lemma 2.1.5. Let $\eta(x, z)$ be defined as

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

Then the gradient of η w.r.t. (x, z) is Lipschitz continuous with the constant

$$L_{\nabla\eta} = 2\sqrt{(1+\beta)^2 + (1+\tfrac{1}{2\beta})^2}.$$

In the next result, we provide a recursion inequality for the error in estimating $f_i(w_{i+1})$ by w_i.

Lemma 2.1.6. Let $\{x^k\}_{k\geq 0}$ and $\{w_i^k\}_{k\geq 0}$ $1 \leq i \leq T$ be generated by algorithm 1. Denote

$$d^k = u^k - x^k, \qquad w_{T+1}^k \equiv x^k \quad \forall k \geq 0, \qquad A_{k,i} = f_i(w_{i+1}^{k+1}) - f_i(w_{i+1}^k) \quad 1 \leq i \leq T. \tag{2.7}$$

a) For any $i \in \{1, \ldots, T\}$,

$$\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 \leq (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \frac{1}{\tau_k}\|A_{k,i}\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + r_i^{k+1}, \tag{2.8}$$

$$\|w_i^{k+1} - w_i^k\|^2 \leq \tau_k^2\left[\|f_i(w_{i+1}^k) - w_i^k\|^2 + \|e_i^{k+1}\|^2 - 2\langle e_i^{k+1}, f_i(w_{i+1}^k) - w_i^k\rangle\right], \tag{2.9}$$

where

$$r_i^{k+1} = 2\tau_k \langle e_i^{k+1}, A_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k)\rangle, \qquad e_i^{k+1} = f_i(w_{i+1}^k) - \bar{G}_i^{k+1}. \tag{2.10}$$

b) If, in addition, f_i's are Lipschitz continuous, we have

$$\|f_T(x^{k+1}) - w_T^{k+1}\|^2 \le (1-\tau_k)\|f_T(x^k) - w_T^k\|^2 + L_{f_T}\tau_k\|d^k\|^2 + \tau_k^2\|e_T^{k+1}\|^2 + r_T^{k+1}, \tag{2.11}$$

$$\begin{aligned}\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 \le{}& (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + L_{f_i}^2\tau_k\left[\|f_{i+1}(w_{i+2}^k) - w_{i+1}^k\|^2 + \|e_{i+1}^{k+1}\|^2\right]\\ &+ \tau_k^2\|e_i^{k+1}\|^2 + \bar{r}_i^{k+1} \qquad\qquad 1 \le i \le T-1,\end{aligned} \tag{2.12}$$

where

$$\bar{r}_i^{k+1} = -2\tau_k L_{f_i}^2\langle e_{i+1}^{k+1}, f_{i+1}(w_{i+2}^k) - w_{i+1}^k\rangle + r_i^{k+1}. \tag{2.13}$$

Proof. Noting eq. (2.4), eq. (2.8), and eq. (2.10), we have

$$\begin{aligned}\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 &= \|A_{k,i} + f_i(w_{i+1}^k) - (1-\tau_k)w_i^k - \tau_k(f_i(w_{i+1}^k) - e_i^{k+1})\|^2\\ &= \|A_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k) + \tau_k e_i^{k+1}\|^2\\ &= \|A_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k)\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + r_i^{k+1}.\end{aligned}$$

Then, in the view of eq. (2.10), eq. (2.8) follows by noting that

$$\begin{aligned}\|A_{k,i} + (1-b\tau_k)(f_i(w_{i+1}^k) - w_i^k)\|^2 &= \|A_{k,i}\|^2 + (1-\tau_k)^2\|f_i(w_{i+1}^k) - w_i^k\|^2\\ &\quad + 2(1-\tau_k)\langle A_{k,i}, f_i(w_{i+1}^k) - w_i^k\rangle\\ &\le \|A_{k,i}\|^2 + (1-\tau_k)^2\|f_i(w_{i+1}^k) - w_i^k\|^2\\ &\quad + \left(\frac{1}{\tau_k} - 1\right)\|A_{k,i}\|^2 + (1-\tau_k)\tau_k\|f_i(w_{i+1}^k) - w_i^k\|^2\\ &= (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \frac{1}{\tau_k}\|A_{k,i}\|^2,\end{aligned} \tag{2.14}$$

due to Cauchy Schwartz and Young's inequalities. Also, eq. (2.9) directly follows from eq. (2.4) since

$$\begin{aligned}\|w_i^{k+1} - w_i^k\|^2 &= \|\tau_k(G_i^{k+1} - w_i^k)\|^2 = \tau_k^2\|f_i(w_{i+1}^k) - w_i^k - e_i^{k+1}\|^2 \\ &= \tau_k^2\left[\|f_i(w_{i+1}^k) - w_i^k\|^2 + \|e_i^{k+1}\|^2 - 2\langle e_i^{k+1}, f_i(w_{i+1}^k) - w_i^k\rangle\right].\end{aligned}$$

To show part b), note that by eq. (2.2), eq. (2.7), and Lipschitz continuity of f_i, we have

$$\|A_{k,T}\| \le L_{f_T}\|w_{T+1}^{k+1} - w_{T+1}^k\| = L_{f_T}\tau_k\|d^k\|, \qquad \|A_{k,i}\| \le L_{f_i}\|w_{i+1}^{k+1} - w_{i+1}^k\| \quad 1 \le i \le T-1.$$

The results then follows by noting eq. (2.8) and eq. (2.9).

■

We remark that the mini-batch sampling in (2.5) is only used to reduce the upper bound on the expectation of $\tau_k\|e_{i+1}^{k+1}\|^2$ in the right hand side of (2.12). Moreover, we do not need this inequality for $i = 1$ when establishing the convergence rate of Algorithm 1. Thus, when $T \le 2$, this algorithm convergences without using mini-batch of samples in each iteration, as shown in [63].

Denoting $w := (w_1, \ldots, w_T)$, we define the merit function

$$W(x, z, w) = F(x) - F^* - \eta(x, z) + \sum_{i=1}^{T-1}\gamma_i\|f_i(w_{i+1}) - w_i\|^2 + \gamma_T\|f_T(x) - w_T\|^2 \tag{2.15}$$

which will be used in our next result for establishing convergence analysis of Algorithm 1.

Lemma 2.1.7. Suppose that $\{x^k, z^k, u^k, w_1^k, \ldots, w_T^k\}_{k\ge 0}$ are generated by Algorithm 1 and Assumption 1 holds.

a) If

$$\gamma_0 := 0, \qquad \gamma_1, \lambda > 0, \qquad \beta_k \equiv \beta \ge \lambda + \gamma_T,$$

$$\gamma_j - \gamma_{j-1}L_{f_{j-1}}^2 - \lambda > 0, \qquad 4(\beta - \lambda - \gamma_T)(\gamma_j - \gamma_{j-1}L_{f_{j-1}}^2 - \lambda) \ge TC_j^2 \qquad j = 2, \ldots, T, \tag{2.16}$$

where C_j's are defined in Lemma 2.1.4, we have

$$\lambda\sum_{k=0}^{N-1}\tau_k\left[\|d^k\|^2 + \sum_{i=1}^{T-1}\|f_i(w_{i+1}^k) - w_i^k\|^2 + \|f_T(x^k) - w_T^k\|^2\right] \le W(x^0, z^0, w^0) + \sum_{k=0}^{N-1}R^{k+1}, \tag{2.17}$$

where

$$R^{k+1} := \tau_k^2 \sum_{i=1}^{T} \gamma_i \|e_i^{k+1}\|^2 + \tau_k \sum_{i=1}^{T-1} \gamma_i L_{f_i}^2 \|e_{i+1}^{k+1}\|^2 + \sum_{i=1}^{T-1} \gamma_i \bar{r}_i^{k+1} + \gamma_T r_T^{k+1} + \tau_k \langle d^k, \Delta^k \rangle,$$
$$+ \frac{(L_{\nabla F} + L_{\nabla \eta})\tau_k^2}{2} \|d^k\|^2 + \frac{L_{\nabla \eta}}{2} \|z^{k+1} - z^k\|^2, \tag{2.18}$$

$$\Delta^k := \nabla f_T(x^k) \prod_{i=2}^{T} \nabla f_{T+1-i}(w_{T+2-i}^k) - \prod_{i=1}^{T} J_{T-i+1}^{k+1}, \tag{2.19}$$

and $r_i^{k+1}, \bar{r}_i^{k+1}$ are defined in eq. (2.10) and eq. (2.13), respectively.

b) If parameters are chosen as

$$\gamma_0 = 0, \qquad \gamma_1 = 1, \qquad \gamma_j := 2^{j-1}(L_{f_1} \cdots L_{f_{j-1}})^2 \quad 2 \le j \le T,$$
$$\lambda = \frac{1}{2} \min_{1 \le i \le T} (\gamma_i - \gamma_{i-1} L_{f_{i-1}}^2), \qquad \beta \ge \lambda + \gamma_T + \frac{T \max_{2 \le i \le T} C_i^2}{4\lambda}. \tag{2.20}$$

Then, conditions in eq. (2.16) are satisfied.

Proof. First, note that by Lemma 2.1.1, we have

$$F(x^{k+1}) \le F(x^k) + \langle \nabla F(x^k), x^{k+1} - x^k \rangle + \frac{L_{\nabla F}}{2} \|x^{k+1} - x^k\|^2$$
$$= F(x^k) + \tau_k \langle \nabla F(x^k), d^k \rangle + \frac{L_{\nabla F} \tau_k^2}{2} \|d^k\|^2. \tag{2.21}$$

Second, note that by the optimality condition of eq. (2.1), we have

$$\langle z^k + \beta_k (u^k - x^k), x^k - u^k \rangle \ge 0, \qquad \langle z^k, d^k \rangle + \beta_k \|d^k\|^2 \le 0. \tag{2.22}$$

Then, noting eq. (2.2), eq. (2.3), and in the view of Lemma 2.1.5, we obtain

$$\begin{aligned}
\eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1}) &\le \langle z^k + \beta_k(u^k - x^k), x^{k+1} - x^k\rangle - \langle u^k - x^k, z^{k+1} - z^k\rangle \\
&+ \frac{L_{\nabla\eta}}{2}\left[\|x^{k+1} - x^k\|^2 + \|z^{k+1} - z^k\|^2\right] \\
&= \tau_k\langle 2z^k + \beta_k d^k, d^k\rangle - \tau_k\langle d^k, \prod_{i=1}^{T} J_{T-i+1}^{k+1}\rangle \qquad (2.23) \\
&+ \frac{L_{\nabla\eta}}{2}\left[\|x^{k+1} - x^k\|^2 + \|z^{k+1} - z^k\|^2\right] \\
&\le -\beta_k\tau_k\|d^k\|^2 - \tau_k\langle d^k, \prod_{i=1}^{T} J_{T-i+1}^{k+1}\rangle + \frac{L_{\nabla\eta}}{2}\left[\tau_k^2\|d^k\|^2 + \|z^{k+1} - z^k\|^2\right]. \qquad (2.24)
\end{aligned}$$

Third, noting Lemma 2.1.6.b), we have

$$\begin{aligned}
&\sum_{i=1}^{T-1}\gamma_i\left[\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 - \|f_i(w_{i+1}^k) - w_i^k\|^2\right] + \gamma_T\left[\|f_T(x^{k+1}) - w_T^{k+1}\|^2 - \|f_T(x^k) - w_T^k\|^2\right] \\
&\le \sum_{i=1}^{T-1}\gamma_i\left\{-\tau_k\left[\|f_i(w_{i+1}^k) - w_i^k\|^2 - L_{f_i}^2\|f_{i+1}(w_{i+2}^k) - w_{i+1}^k\|^2 - L_{f_i}^2\|e_{i+1}^{k+1}\|^2\right] + \tau_k^2\|e_i^{k+1}\|^2 + \bar{r_i}^{k+1}\right\} \\
&+ \gamma_T\left\{-\tau_k\left[\|f_T(x^k) - w_T^k\|^2 - L_{f_T}^2\|d^k\|^2\right] + \tau_k^2\|e_T^{k+1}\|^2 + r_T^{k+1}\right\} \\
&= -\tau_k\{\gamma_1\|f_1(w_2^k) - w_1^k\|^2 + \textstyle\sum_{j=2}^{T-1}[\gamma_j - \gamma_{j-1}L_{f_{j-1}}^2]\|f_j(w_{j+1}^k) - w_j^k\|^2 \qquad (2.25) \\
&+ [\gamma_T - \gamma_{T-1}L_{f_{T-1}}^2]\|f_T(x^k) - w_T^k\|^2\} + \tau_k\left[\sum_{i=1}^{T-1}\gamma_i L_{f_i}^2\|e_{i+1}^{k+1}\|^2 + \gamma_T\|d^k\|^2\right] \qquad (2.26) \\
&+ \tau_k^2\sum_{i=1}^{T}\gamma_i\|e_i^{k+1}\|^2 + \sum_{i=1}^{T-1}\gamma_i\bar{r_i}^{k+1} + \gamma_T r_T^{k+1}. \qquad (2.27)
\end{aligned}$$

Combining the above relation with eq. (2.23), eq. (2.21), noting definition of merit function in eq. (2.15), and in the view of lemma 2.1.4, we obtain

$$
\begin{aligned}
&W(x^{k+1}, z^{k+1}, w^{k+1}) - W(x^k, z^k, w^k)\\
&\le -\tau_k(\beta_k - \gamma_T)\|d^k\|^2 + \tau_k\|d^k\| \left[\sum_{j=2}^{T-1} C_j\|f_j(w_{j+1}) - w_j\| + C_T\|f_T(x) - w_T\|\right] + R^{k+1}\\
&- \tau_k\{\gamma_1\|f_1(w_2^k) - w_1^k\|^2\\
&+ \textstyle\sum_{j=2}^{T-1}[\gamma_j - \gamma_{j-1}L_{f_{j-1}}^2]\|f_j(w_{j+1}^k) - w_j^k\|^2 + [\gamma_T - \gamma_{T-1}L_{f_{T-1}}^2]\|f_T(x^k) - w_T^k\|^2\},
\end{aligned}
$$

where R^{k+1} is defined in eq. (2.18). Thus, if eq. (2.16) holds, we have

$$
\begin{aligned}
&W(x^{k+1}, z^{k+1}, w^{k+1}) - W(x^k, z^k, w^k)\\
&\le \lambda \sum_{k=0}^{N-1} \tau_k \left[\|d^k\|^2 + \sum_{i=1}^{T-1} \|f_i(w_{i+1}^k) - w_i^k\|^2 + \|f_T(x^k) - w_T^k\|^2\right] + R^{k+1}.
\end{aligned}
$$

Summing up the above inequalities and re-arranging the terms, we obtain eq. (2.17). It can be easily verified that condition eq. (2.16) is satisfied by the choice of parameters in eq. (2.20).

■

We introduce the following additional lemmas.

Lemma 2.1.8. Consider a sequence $\{\tau_k\}_{k\ge 0} \in (0,1]$, and define

$$
\Gamma_k = \Gamma_1 \prod_{i=1}^{k-1}(1-\tau_i) \qquad k \ge 2, \qquad \Gamma_1 = \begin{cases} 1 & \text{if } \tau_0 = 1,\\ 1-\tau_0 & \text{otherwise.}\end{cases} \tag{2.28}
$$

a) For any $k \ge 1$, we have

$$
\alpha_{i,k} = \frac{\tau_i}{\Gamma_{i+1}}\Gamma_k \quad 1 \le i \le k, \qquad \sum_{i=0}^{k-1}\alpha_{i,k} = \begin{cases} 1 & \text{if } \tau_0 = 1,\\ 1-\Gamma_k & \text{otherwise.}\end{cases}
$$

b) Suppose that $q_{k+1} \le (1-\tau_k)q_k + p_k \quad k \ge 0$ for sequences $\{q_k, p_k\}_{k\ge 0}$. Then, we have

$$q_k \le \Gamma_k \left[a q_0 + \sum_{i=0}^{k-1} \frac{p_i}{\Gamma_{i+1}} \right], \qquad a = \begin{cases} 0 & \text{if } \tau_0 = 1, \\ 1 & \text{otherwise.} \end{cases}$$

Proof. To show part a), note that

$$\sum_{i=0}^{k-1} \alpha_{i,k} = \Gamma_k \sum_{i=0}^{k-1} \frac{\tau_i}{\Gamma_{i+1}} = \frac{\tau_0 \Gamma_k}{\Gamma_1} + \sum_{i=1}^{k-1} \frac{\tau_i \Gamma_k}{\Gamma_{i+1}} = \frac{\tau_0 \Gamma_k}{\Gamma_1} + \Gamma_k \sum_{i=1}^{k-1} \left(\frac{1}{\Gamma_{i+1}} - \frac{1}{\Gamma_i} \right) = 1 - \frac{\Gamma_k}{\Gamma_1}(1-\tau_0).$$

To show part b), by dividing both sides of the inequality by Γ_{k+1} and noting eq. (2.28), we have

$$\frac{q_1}{\Gamma_1} \le \frac{(1-\tau_0)q_0 + p_0}{\Gamma_1}, \qquad \frac{q_{k+1}}{\Gamma_{k+1}} \le \frac{q_k}{\Gamma_k} + \frac{p_k}{\Gamma_{k+1}} \quad k \ge 1.$$

Summing up the above inequalities, we get the result. ∎

Proposition 2.1.1. *Suppose that Assumption 2 holds and (for simplicity) $\tau_0 = 1$, $\beta_k = \beta > 0$ for all k. Then, for any $k \ge 1$, we have*

$$\beta^2 \mathbb{E}[\|d^k\|^2 | \mathscr{F}_k] \le \mathbb{E}[\|z^k\|^2 | \mathscr{F}_k] \le \prod_{i=1}^{T} \sigma_{J_i}^2, \tag{2.29}$$

$$\mathbb{E}[\|z^{k+1} - z^k\|^2 | \mathscr{F}_k] \le 4\tau_k^2 \prod_{i=1}^{T} \sigma_{J_i}^2. \tag{2.30}$$

If, in addition, the batch size b_k in Algorithm 1 is set to

$$b_k = \left\lceil \frac{\max_{1\le i\le T} L_{f_i}^2}{\tau_k} \right\rceil \qquad k \ge 0, \tag{2.31}$$

we have

$$\mathbb{E}[R^{k+1} | \mathscr{F}_k] \le \tau_k^2 \left[\frac{1}{2} \left(\prod_{i=1}^{T} \sigma_{J_i}^2 \right) \left(\frac{L_{\nabla F} + (1+4\beta^2) L_{\nabla \eta}}{\beta^2} \right) + \sum_{i=1}^{T} \gamma_i \sigma_{G_i}^2 \right] := \tau_k^2 \sigma^2, \tag{2.32}$$

where R^{k+1} is defined in eq. (2.18)*.*

Proof. The first inequality in eq. (2.29) directly follows by eq. (2.22) and Cauchy-Schwarz inequality. Noting eq. (2.3), the fact that $\tau_0 = 1$, and in the view of Lemma 2.1.8, we obtain

$$z^k = \sum_{i=0}^{k-1} \alpha_{i,k} \left(\prod_{\ell=1}^{T} J_{T+1-l}^{i+1} \right).$$

By convexity of $\|\cdot\|^2$ and conditional independence, we conclude that

$$\begin{aligned}
\mathbb{E}[\|z^k\|^2|\mathscr{F}_k] &\le \sum_{i=0}^{k-1} \alpha_{i,k} \mathbb{E}\left[\left\| \prod_{\ell=1}^{T} J_\ell^{i+1} \right\|^2 \;\middle|\; \mathscr{F}_k \right] \\
&\le \sum_{i=0}^{k-1} \alpha_{i,k} \prod_{\ell=1}^{T} \mathbb{E}[\|J_\ell^{i+1}\|^2|\mathscr{F}_i] \le \sum_{i=0}^{k-1} \alpha_{i,k} \left(\prod_{\ell=1}^{T} \sigma_{J_\ell}^2 \right) = \prod_{\ell=1}^{T} \sigma_{J_\ell}^2.
\end{aligned}$$

Noting eq. (2.29), we have

$$\begin{aligned}
\mathbb{E}[\|z^{k+1} - z^k\|^2|\mathscr{F}_k] &\le \tau_k^2 \mathbb{E}\left[\left\| z^k - \prod_{\ell=1}^{T} J_\ell^{k+1} \right\|^2 \;\middle|\; \mathscr{F}_k \right] \\
&\le 2\tau_k^2 \left[\mathbb{E}[\|z^k\|^2|\mathscr{F}_k] + \mathbb{E}\left[\left\| \prod_{\ell=1}^{T} J_\ell^{k+1} \right\|^2 \;\middle|\; \mathscr{F}_k \right] \right] \\
&\le 2\tau_k^2 \left(\prod_{\ell=1}^{T} \sigma_{J_\ell}^2 + \prod_{\ell=1}^{T} \sigma_{J_\ell}^2 \right) \\
&= 4\tau_k^2 \left(\prod_{\ell=1}^{T} \sigma_{J_\ell}^2 \right).
\end{aligned}$$

Now, observe that by eq. (2.10), eq. (2.13), the choice of b_k in eq. (2.31), and under Assumption 2, we have

$$\mathbb{E}[\Delta^k|\mathscr{F}_k] = 0, \qquad \mathbb{E}[e_i^{k+1}|\mathscr{F}_k] = 0, \quad \text{which implies} \quad \mathbb{E}[r_i^{k+1}|\mathscr{F}_k] = \mathbb{E}[\bar{r}_i^{k+1}|\mathscr{F}_k] = 0,$$

$$\mathbb{E}[\|e_i^{k+1}\|^2|\mathscr{F}_k] = \mathbb{E}[\|\frac{1}{b_k} G_{i,j}^{k+1} - f_i(w_{i+1}^k)\|^2|\mathscr{F}_k] \le \frac{\sigma_{G_i}^2}{b_k} \le \min\left\{1, \frac{\tau_k}{\max_{1\le i\le T} L_{f_i}^2}\right\} \sigma_{G_i}^2.$$

Noting eq. (2.18), eq. (2.29), eq. (2.30), and the above observation, we obtain eq. (2.32). ■

Observe that Lemma 2.1.7 shows that the summation of $\|d^k\|$ and the errors in estimating the inner function values is bounded by summation of error terms R^k which is in the order of $\sum_{k=1}^N \tau_k^2$ as shown in Proposition 2.1.1. This is the main step in establishing the convergence of Algorithm 1.

Indeed, $\bar{x} \in X$ is a stationary point of eq. (1.5), if $u = \bar{x}$ and $\bar{z} = \nabla F(\bar{x})$, where

$$u = \arg\min_{y \in X} \left\{ \langle \bar{z}, y - \bar{x} \rangle + \frac{1}{2} \|y - \bar{x}\|^2 \right\}. \tag{2.33}$$

Thus, for a given pair of $(\bar{x}, \bar{z})$, we can define our termination criterion as follows.

Definition 2.1.1. A pair of $(\bar{x}, \bar{z})$ generated by Algorithm 1 is called an ϵ-stationary pair, if $\mathbb{E}[\sqrt{V(\bar{x}, \bar{z})}] \leq \epsilon$, where

$$V(x, z) = \|u - x\|^2 + \|z - \nabla F(x)\|^2, \tag{2.34}$$

and u is the solution to (2.33).

When $X = \mathbb{R}^{d_T}$, $V(x, z)$ provides an upper bound for the $\|\nabla F(x)\|^2$. One can see [63] for the relation between $V(\bar{x}, \bar{z})$ and other common gradient-based termination criteria such as gradient mapping. Furthermore, as shown in [63], we have

$$V(x^k, z^k) = \max(1, \beta_k^2) \|u^k - x^k\|^2 + \|z^k - \nabla F(x^k)\|^2, \tag{2.35}$$

where (x^k, u^k, z^k) are the solutions generated at iteration $k - 1$ of Algorithm 1. Noting this fact, we provide convergence rate of this algorithm by appropriately choosing β_k and τ_k in the next results.

Theorem 2.1.9. *Suppose that $\{x^k, z^k\}_{k \geq 0}$ are generated by Algorithm 1, Assumption 1 and Assumption 2 hold. Also assume that the parameters satisfy eq.* (2.20) *and step sizes $\{\tau_k\}$ are chosen such that*

$$\sum_{i=k+1}^{N} \tau_i \Gamma_i \leq c \Gamma_{k+1} \quad \forall k \geq 0 \text{ and } \forall N \geq 1, c \text{ is a positive constant.} \tag{2.36}$$

(a) *For every $N \geq 1$, we have*

$$\sum_{k=1}^{N} \tau_k \mathbb{E}[\|\nabla F(x^k) - z^k\|^2 | \mathscr{F}_k] \leq \mathcal{B}_1(\sigma^2, N), \tag{2.37}$$

where

$$\mathcal{B}_1(\sigma^2, N) = \frac{4cL^2(T-1)}{\lambda} \left[W(x^0, z^0, w^0) + \sigma^2 \sum_{k=0}^{N-1} \tau_k^2 \right] + c \prod_{\ell=1}^{T} \sigma_{J_\ell}^2 \sum_{k=0}^{N-1} \tau_k^2, \tag{2.38}$$

σ^2 is defined in eq. (2.32) and

$$L^2 = \max\left\{L^2_{\nabla F}, \max_{1\le i\le T} C_j^2\right\}. \tag{2.39}$$

(b) *As a consequence, we have*

$$\mathbb{E}[V(x^R, z^R)] \le \frac{1}{\sum_{k=1}^N \tau_k}\left\{\mathcal{B}_1(\sigma^2, N) + \frac{\max(1,\beta^2)}{\lambda}\left[W(x^0, z^0, w^0) + \sigma^2 \sum_{k=0}^N \tau_k^2\right]\right\}, \tag{2.40}$$

where the expectation is taken with respect to all random sequences generated by the method and an independent random integer number $R \in \{1, \dots, N\}$, whose probability distribution is given by

$$\mathbb{P}[R = k] = \frac{\tau_k}{\sum_{j=1}^N \tau_j}.$$

(c) *If, in addition, the stepsizes are set to*

$$\tau_0 = 1, \quad \tau_k = \frac{1}{\sqrt{N}} \quad \forall k = 1, \dots, N, \tag{2.41}$$

we have

$$\mathbb{E}[\|\nabla F(x^R) - z^R\|^2] \le \frac{1}{\sqrt{N}}\left[\frac{4L^2(T-1)\left[W(x^0, z^0, w^0) + 2\sigma^2\right]}{\lambda} + 2\prod_{\ell=1}^T \sigma^2_{J_\ell}\right] := \frac{\mathcal{B}_2(\sigma^2, N)}{\sqrt{N}}, \tag{2.42}$$

$$\mathbb{E}[V(x^R, z^R)] \le \frac{1}{\sqrt{N}}\left[\mathcal{B}_2(\sigma^2, N) + \frac{\max(1,\beta^2)}{\lambda}\left[W(x^0, z^0, w^0) + 2\sigma^2\right]\right], \tag{2.43}$$

$$\mathbb{E}[\|f_i(w^R_{i+1}) - w^R_i\|^2] \le \frac{1}{\lambda\sqrt{N}}\left[W(x^0, z^0, w^0) + 2\sigma^2\right] \qquad i = 1, \dots, T. \tag{2.44}$$

Proof. We first show part (a). Noting eq. (2.3), we have

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

where Δ^k is defined in eq. (2.18) and

$$\delta^k = \nabla F(x^k) - \nabla f_T(x^k)\prod_{i=2}^T \nabla f_{T+1-i}(w^k_{T+2-i}), \qquad \bar{\delta}^k = \frac{\nabla F(x^{k+1}) - \nabla F(x^k)}{\tau_k}.$$

Denoting $\bar{\Delta}_k = \langle \Delta^k, (1-\tau_k)(\nabla F(x^k) - z^k) + \tau_k(\delta^k + \bar{\delta}^k)\rangle$, we have

$$\begin{aligned}\|\nabla F(x^{k+1}) - z^{k+1}\|^2 &= \|(1-\tau_k)(\nabla F(x^k) - z^k) + \tau_k(\delta^k + \bar{\delta}^k)\|^2 + \tau_k^2\|\Delta^k\|^2 + 2\tau_k\bar{\Delta}_k \\ &\leq (1-\tau_k)\|\nabla F(x^k) - z^k\|^2 + 2\tau_k\left[\|\delta^k\|^2 + L_{\nabla F}^2\|d^k\|^2 + \bar{\Delta}_k\right] + \tau_k^2\|\Delta^k\|^2,\end{aligned}$$

where the inequality follows from convexity of $\|\cdot\|^2$ and Lipschitz continuity of gradient of F. Thus, in the view of Lemma 2.1.8, we obtain

$$\|\nabla F(x^k) - z^k\|^2 \leq 2\Gamma_k \sum_{i=0}^{k-1} \frac{\tau_i}{\Gamma_{i+1}} \left(\|\delta^i\|^2 + L_{\nabla F}\|d^i\|^2 + \bar{\Delta}_i + \frac{\tau_i}{2}\|\Delta^i\|^2\right),$$

which implies that

$$\begin{aligned}\sum_{k=1}^{N} \tau_k\|\nabla F(x^k) - z^k\|^2 &= 2\sum_{k=1}^{N}\tau_k\Gamma_k\sum_{i=0}^{k-1}\frac{\tau_i}{\Gamma_{i+1}}\left(\|\delta^i\|^2 + L_{\nabla F}^2\|d^i\|^2 + \bar{\Delta}_i + \frac{\tau_i}{2}\|\Delta^i\|^2\right) \\ &= 2\sum_{k=0}^{N-1}\frac{\tau_k}{\Gamma_{k+1}}\left(\sum_{i=k+1}^{N}\tau_i\Gamma_i\right)\left(\|\delta^k\|^2 + L_{\nabla F}^2\|d^k\|^2 + \bar{\Delta}_k + \frac{\tau_k}{2}\|\Delta^k\|^2\right) \\ &\leq 2c\sum_{k=0}^{N-1}\tau_k\left(\|\delta^k\|^2 + L_{\nabla F}^2\|d^k\|^2 + \bar{\Delta}_k + \frac{\tau_k}{2}\|\Delta^k\|^2\right), \qquad (2.45)\end{aligned}$$

where the last inequality follows from eq. (2.36).

Now, observe that under Assumption 2, we have

$$\mathbb{E}[\bar{\Delta}_k|\mathscr{F}_k] = 0, \qquad \mathbb{E}[\|\Delta_k\|^2|\mathscr{F}_k] \leq \mathbb{E}\left[\left\|\prod_{\ell=1}^{T} J_\ell^{k+1}\right\|^2 \middle| \mathscr{F}_k\right] \leq \prod_{\ell=1}^{T}\sigma_{J_\ell}^2.$$

Moreover, by Lemma 2.1.4 and the fact that $(\sum_{i=1}^n a_i)^2 \leq n\sum_{i=1}^n a_i^2$ for nonnegative a_i's, we have

$$\begin{aligned}\|\delta_k\|^2 &= \left\|\nabla F(x) - \nabla f_T(x)\prod_{i=2}^{T}\nabla f_{T+1-i}(w_{T+2-i})\right\|^2 \\ &\leq 2(T-1)\sum_{j=2}^{T-1} C_j^2\|f_j(w_{j+1}) - w_j\|^2 + 2C_T^2\|f_T(x) - w_T\|^2.\end{aligned}$$

Combining the above observations with eq. (2.46) and in the view of eq. (2.39), we obtain

$$\sum_{k=1}^{N} \tau_k \mathbb{E}[\|\nabla F(x^k) - z^k\|^2 | \mathscr{F}_k] \le 4cL(T-1) \sum_{k=0}^{N-1} \tau_k \left(\sum_{j=2}^{T-1} \|f_j(w_{j+1}) - w_j\|^2 + \|f_T(x) - w_T\|^2 + \|d^k\|^2 \right)$$
$$+ c \prod_{\ell=1}^{T} \sigma_{J_\ell}^2 \sum_{k=0}^{N-1} \tau_k^2. \tag{2.46}$$

Then, eq. (2.37) follows from the above inequality, eq. (2.17), and eq. (2.32).

Part (b) then follows from part (a), eq. (2.35), eq. (2.17), and noting that

$$\mathbb{E}[V(x^R, z^R)] = \frac{\sum_{k=1}^{N} \tau_k V(x^k, z^k)}{\sum_{j=1}^{N} \tau_j}.$$

Part (c) also follows by noting that choice of τ_k in eq. (2.41) implies that

$$\sum_{k=1}^{N} \tau_k \ge \sqrt{N}, \quad \sum_{k=0}^{N} \tau_k^2 = 2, \quad \Gamma_k = \left(1 - \frac{1}{\sqrt{N}}\right)^{k-1},$$
$$\sum_{i=k+1}^{N} \tau_i \Gamma_i = \left(1 - \frac{1}{\sqrt{N}}\right)^{k} \frac{1}{\sqrt{N}} \sum_{i=0}^{N-k-1} \left(1 - \frac{1}{\sqrt{N}}\right)^{i} \le \left(1 - \frac{1}{\sqrt{N}}\right)^{k},$$

ensuring condition eq. (2.36) with $c = 1$.

■

Remark 2.1.1. *The result in (2.43) implies that to find an ϵ-stationary point of (1.5) (see, definition 2.1.1), Algorithm 1 requires $\mathcal{O}(\rho^T T^4/\epsilon^4)$ number of iterations, where ρ is a constant depending on the problem parameters (i.e., Lipschitz constants and noise variances). Thus, the total number of used samples is bounded by*

$$\sum_{k=1}^{T} b_k = \mathcal{O}\left(\frac{\rho^T T^6}{\epsilon^6}\right)$$

due to (2.31) and (2.41). This bound is much better than $\mathcal{O}\left(1/\epsilon^{(7+T)/2}\right)$ obtained in [129] when $T > 4$[1] *. In particular, it exhibits the level-independent behavior as discussed in Section 1.3. Note that, we obtain constants of order ρ^T, for example, when $\sigma_{J_i}^2$ in eq.* (2.32) *are all of equal. We*

[1]Following the presentation in [129], we only present the ϵ-related T dependence for their result.

emphasize that [129] and [131] also have such constant factors that depend exponentially on T, in their proofs and the final results.

Remark 2.1.2. *The bound in (2.44) also implies that the errors in estimating the inner function values decrease at the same rate that we converge to the stationary point of the problem. This is essential to obtain a rate of convergence similar to that of single-level problems. Moreover, (2.42) shows that the stochastic estimate z^k also converges at the same rate to the gradient of the objective function at the stationary point where x^k converges to.*

Although our results for Algorithm 1 show improved convergence rates compared to [129], it is still worse than $\mathcal{O}\left(1/\epsilon^4\right)$ obtained in [63] for the case of $T = 2$. Furthermore, the batch sizes b_k is of order ρ^T for some constant ρ which makes it impractical. In the next section, we show that both of these issues could be fixed by a properly modified variant of Algorithm 1.

2.2. Multi-level Nested Linearized Averaging Stochastic Gradient Method

In this section, we present a linearized variant of Algorithm 1 which can achieve the state-of-art rate of convergence for problem (1.5) for any $T \geq 1$. Indeed, when $T > 2$, we have accumulated errors in estimating the inner function values. Hence, in Algorithm 1 we use mini-batch sampling in (2.4) to reduce the noise associated with the stochastic function values. However, this increases the sample complexity of the algorithm. To resolve this issue, instead of using the point estimates of f_i's, we use their stochastic linear approximations in (2.47). With this modification, a refined convergence analysis enables us to obtain a sample complexity of $\mathcal{O}(1/\epsilon^4)$ with Algorithm 2, for any $T \geq 1$ without using any mini-batches. Here, we remark that similar linearization techniques have been proposed as early as [109] in other contexts. Furthermore, it was also used in [37, 45] and [110] recently for the two-level and multi-level cases respectively.

Algorithm 2 Multi-level Nested Linearized Averaging Stochastic Gradient Method

Set $b_k = 1$ in Algorithm 1 and replace (2.4) with

$$w_i^{k+1} = (1-\tau_k)w_i^k + \tau_k G_i^{k+1} + J_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k), \qquad 1 \leq i \leq T. \tag{2.47}$$

To establish the rate of convergence of Algorithm 2, we need to make the following additional assumption on the fourth-moments of the outputs of the stochastic oracle, similar to [129].

Assumption 3. *Denote $w_{T+1}^k \equiv x^k$. Instantiate the conditions in Assumption 2. In addition to that, the stochastic oracle satisfies, for $1 \leq i \leq T$,*

(1) $\mathbb{E}[\|J_i^{k+1}\|^4|\mathscr{F}_k] \leq \kappa_{J_i}^4$, $\mathbb{E}[\|J_i^{k+1} - \nabla f_i(w_{i+1}^k)\|^2|\mathscr{F}_k] \leq \varrho_{J_i}^2$, $\mathbb{E}[\|J_i^{k+1} - \nabla f_i(w_{i+1}^k)\|^4|\mathscr{F}_k] \leq \varkappa_{J_i}^4$,

(2) $\mathbb{E}[\|G_i^{k+1} - f_i(w_{i+1}^k)\|^4|\mathscr{F}_k] \leq \kappa_{G_i}^4$.

The above assumptions are trivially satisfied when the ξ_is are drawn from any light-tailed distributions (for example, sub-Gaussian). Relaxing the bounded fourth-moment assumptions to the bounded second-moment assumption, as in section 2.1 seems extremely challenging without strong assumptions on the objective function and the constraint set X. The next result provides the recursion on the errors in estimating the inner function values.

Lemma 2.2.1. Let $\{x^k\}_{k\geq 0}$ and $\{w_i^k\}_{k\geq 0}$ $1 \leq i \leq T$ be generated by Algorithm 2. Define, for $1 \leq i \leq T$,

$$e_i^{k+1} := f_i(w_{i+1}^k) - G_i^{k+1},\ \hat{e}_i^{\,k+1} := \nabla f_i(w_{i+1}^k) - J_i^{k+1}, \tag{2.48}$$

$$A_{k,i} := f_i(w_{i+1}^{k+1}) - f_i(w_{i+1}^k) - \nabla f_i(w_{i+1}^k)(w_{i+1}^{k+1} - w_{i+1}^k). \tag{2.49}$$

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

$$\begin{aligned}\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 \leq (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \frac{L_{\nabla f_i}^2}{4\tau_k}\|w_{i+1}^{k+1} - w_{i+1}^k\|^4 + \tau_k^2\|e_i^{k+1}\|^2 + \dot{r}_i^{k+1} \\ + \|\hat{e}_i^{\,k+1}\|^2\|w_{i+1}^{k+1} - w_{i+1}^k\|^2,\end{aligned} \tag{2.50}$$

where,

$$\begin{aligned}\dot{r}_i^{k+1} &:= 2\tau_k\langle e_i^{k+1}, A_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k) + \hat{e}_i^{\,k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle \\ &+ 2\langle \hat{e}_i^{\,k+1}(w_{i+1}^{k+1} - w_{i+1}^k), A_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k)\rangle.\end{aligned} \tag{2.51}$$

b) Furthermore, we have for $1 \leq i \leq T$,

$$
\begin{aligned}
\|w_i^{k+1} - w_i^k\|^2 \leq \tau_k^2 \left[2\|f_i(w_{i+1}^k) - w_i^k\|^2 + \|e_i^{k+1}\|^2 + \frac{2}{\tau_k^2}\|J_i^{k+1}\|^2\|w_{i+1}^{k+1} - w_{i+1}^k\|^2 \right] + 2\ddot{r}_i^{k+1}, \\
\ddot{r}_i^{k+1} := \tau_k\langle -e_i^{k+1}, \tau_k(f_i(w_{i+1}^k) - w_i^k) + J_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle, \\
\|w_i^{k+1} - w_i^k\|^4 \leq \tau_k^4 \left[6\|f_i(w_{i+1}^k) - w_i^k\|^4 + 35\|e_i^{k+1}\|^4 + \frac{40}{\tau_k^4}\|J_i^{k+1}\|^4\|w_{i+1}^{k+1} - w_{i+1}^k\|^4 \right] \\
+ 4\ddot{r}_i^{k+1} \left[2\tau_k^2\|f_i(w_{i+1}^k) - w_i^k\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + 2\|J_i^{k+1}\|^2\|w_{i+1}^{k+1} - w_{i+1}^k\|^2 \right].
\end{aligned}
$$

Proof. We first prove part a). When $1 \leq i < T$, by definition of $A_{k,i}, \hat{e}_i^{k+1}, G_i^{k+1}, w_i^{k+1}$, and $\dot{r}_i^{k+1}$, we have

$$
\begin{aligned}
&\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 \\
=&\|A_{k,i} + f_i(w_{i+1}^k) + \nabla f_i(w_{i+1}^k)(w_{i+1}^{k+1} - w_{i+1}^k) - (1-\tau_k)w_i^k - \tau_k G_i^{k+1} - J_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2 \\
=&\|A_{k,i} + \widehat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k) + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k) + \tau_k e_i^{k+1}\|^2 \\
=&\|\widehat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2 + \|A_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k)\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + r_i^{k+1} \\
\leq&\|A_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k)\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + \dot{r}_i^{k+1} + \|\widehat{e}_i^{k+1}\|^2\|w_{i+1}^{k+1} - w_{i+1}^k\|^2.
\end{aligned}
$$

Combining the above inequality with (2.14) and noting that under Assumption 1,

$$
\|A_{k,i}\| \leq \frac{1}{2}\min\left\{4L_{f_i}\|w_{i+1}^{k+1} - w_{i+1}^k\|, L_{\nabla f_i}\|w_{i+1}^{k+1} - w_{i+1}^k\|^2\right\}, \tag{2.52}
$$

we obtain eq. (2.50).

We now prove part b). Note that by the definition of eq. (2.4) and eq. (2.48), Cauchy-Schwartz and Young's inequality, we have for $1 \le i \le T$,

$$\begin{aligned}
\|w_i^{k+1} - w_i^k\|^2 &= \|\tau_k(G_i^{k+1} - w_i^k) + J_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2 \\
&= \tau_k^2\|G_i^{k+1} - w_i^k\|^2 + \|J_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2 + 2\tau_k\langle G_i^{k+1} - w_i^k, J_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle \\
&\le \tau_k^2\|G_i^{k+1} - w_i^k\|^2 + 2\|J_i^{k+1}\|^2\|w_{i+1}^{k+1} - w_{i+1}^k\|^2 + \tau_k^2\|f_i(w_{i+1}^k) - w_i^k\|^2 \\
&+ 2\tau_k\langle -e_i^{k+1}, J_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle \\
&= 2\tau_k^2\|f_i(w_{i+1}^k) - w_i^k\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + 2\|J_i^{k+1}\|^2\|w_{i+1}^{k+1} - w_{i+1}^k\|^2 \\
&+ 2\tau_k\langle -e_i^{k+1}, \tau_k(f_i(w_{i+1}^k) - w_i^k) + J_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle.
\end{aligned}$$

Computing the squared of both sides of the above inequality and noting that

$$\langle a, b + c\rangle^2 \le \|a\|^2\|b + c\|^2 \le 2\|a\|^4 + \|b\|^4 + \|c\|^4,$$

we obtain the last result.

■

We now require the following intermediate results to proceed.

Lemma 2.2.2. For two vectors x, y of equal dimension and any $\delta > 0$, we have

$$\|x + y\|^2 \le (1 + \delta)\|x\|^2 + \left(1 + \frac{1}{\delta}\right)\|y\|^2, \tag{2.53}$$

$$\|x + y\|^4 \le (1 + \delta)^3\|x\|^4 + \left(1 + \frac{1}{\delta}\right)^3\|y\|^4. \tag{2.54}$$

Proof. By Cauchy Schwartz inequality, Young's inequality, and the fact that

$$2\langle x, y\rangle = 2\left\langle \sqrt{\delta}x, \frac{y}{\sqrt{\delta}}\right\rangle \le \delta\|x\|^2 + \frac{\|y\|^2}{\delta},$$

eq. (2.53) follows. Next, by *eq.* (2.53) and Young's inequality, we have

$$\begin{aligned}
\|x+y\|^4 &\le (1+\delta)^2\|x\|^4 + \left(1+\frac{1}{\delta}\right)^2\|y\|^4 + 2(1+\delta)\left(1+\frac{1}{\delta}\right)\|x\|^2\|y\|^2 \\
&\le (1+\delta)^2\|x\|^4 + \left(1+\frac{1}{\delta}\right)^2\|y\|^4 + (1+\delta)^2\delta\|x\|^4 + \left(1+\frac{1}{\delta}\right)^2\frac{1}{\delta}\|y\|^4 \\
&= (1+\delta)^3\|x\|^4 + \left(1+\frac{1}{\delta}\right)^3\|y\|^4.
\end{aligned}$$

■

Lemma 2.2.3. Let α_i, p_i, q_i, be sequences such that $\alpha_i = p_i + \alpha_{i+1}q_i$ for $1 \le i \le T$. Then, for $1 \le i < T$, we have

$$\alpha_i = p_i + \sum_{j=i+1}^{T} p_j\left(\prod_{\ell=i}^{j-1} q_\ell\right) + \alpha_{T+1}\left(\prod_{\ell=i}^{T} q_\ell\right).$$

Proof. Base case for $i = T-1$, we have

$$\alpha_{T-1} = p_{T-1} + \alpha_T q_{T-1} = p_{T-1} + q_{T-1}p_T + q_{T-1}q_T\alpha_{T+1}.$$

Assume for all $1 < i+1 \le T-1$, the result holds. We show it holds for the ith case. By induction hypothesis,

$$\alpha_{i+1} = p_{i+1} + \sum_{j=i+2}^{T} p_j\left(\prod_{\ell=i+1}^{j-1} q_\ell\right) + \alpha_{T+1}\left(\prod_{\ell=i+1}^{T} q_\ell\right).$$

Then

$$\alpha_i = p_i + q_i\left[p_{i+1} + \sum_{j=i+2}^{T} p_j\left(\prod_{\ell=i+1}^{j-1} q_\ell\right) + \alpha_{T+1}\left(\prod_{\ell=i+1}^{T} q_\ell\right)\right] = p_i + \sum_{j=i+1}^{T} p_j\left(\prod_{\ell=i}^{j-1} q_\ell\right) + \alpha_{T+1}\left(\prod_{\ell=i}^{T} q_\ell\right).$$

This proves the inductive step.

■

In the next result, we show how the moments of $\|w_i^{k+1} - w_i^k\|$ decrease in the corresponding order of τ_k. This is a crucial step on bounding the errors in estimating the inner function values.

Lemma 2.2.4. Under Assumption 1 and Assumption 3, for $1 \le i \le T$, and with the choice of $\tau_0 = 1$ (for simplicity), we have

$$\mathbb{E}[\|w_i^{k+1} - w_i^k\|^2 | \mathscr{F}_k] \le \tilde{c}_i \ \tau_k^2, \tag{2.55}$$

$$\mathbb{E}[\|w_i^{k+1} - w_i^k\|^4 | \mathscr{F}_k] \le c_i \ \tau_k^4, \tag{2.56}$$

where

$$\tilde{c}_i = \begin{cases} 18 \left[\sigma_{G_i}^2 + \left(\sum_{j=i+1}^{T-1} \sigma_{G_j}^2 + \sigma_{G_T}^2\right) \Upsilon\right] + \left(\prod_{i=1}^{T} \sigma_{J_i}^2\right) \beta^{-2} \ \Upsilon & \text{for } 1 \le i < T-1, \\ 32 \ \sigma_{G_{T-1}}^2 + 18 \ \sigma_{G_T}^2 \ \Phi + \left(\prod_{i=1}^{T} \sigma_{J_i}^2\right) \beta^{-2} \ \Psi & \text{for } i = T-1, \\ 5 \ \sigma_{G_T}^2 + \left(\prod_{i=1}^{T} \sigma_{J_i}^2\right) \beta^{-2} \ [16 L_{f_T}^2 + 4\varrho_{J_T}^2 + 2\sigma_{J_T}^2] & \text{for } i = T. \end{cases}$$

$$c_i = \begin{cases} 3107 \ \kappa_{G_i}^4 + \Theta \ (\sum_{j=i+1}^{T} 3107 \ \kappa_{G_j}^4 + \sigma_d) & \text{for } 1 \le i < T-1, \\ 3107 \ \kappa_{G_{T-1}}^4 + 3107 \ \kappa_{G_T}^4 \ \Xi + \sigma_d \ \Omega & \text{for } i = T-1, \\ 3107 \ \kappa_{G_T}^4 + \sigma_d \ [2^8 \cdot 3L_{f_T}^4 + 2^8 \cdot 3\varkappa_{J_T}^4 + 2^4 \cdot 3\kappa_{J_T}^4] & \text{for } i = T. \end{cases}$$

with

$$\Upsilon := \prod_{\ell=i}^{j-1} 18L_{f_\ell}^2 + 8\varrho_{J_\ell}^2 + 4\sigma_{J_\ell}^2, \qquad \Theta := \prod_{\ell=i}^{T-1} 2^8 \cdot 3L_{f_\ell}^4 + 2^8 \cdot 3\varkappa_{J_\ell}^4 + 2^4 \cdot 3\sigma_{J_\ell}^4,$$

$$\Phi := 18L_{f_{T-1}}^2 + 8\varrho_{J_{T-1}}^2 + 4\sigma_{J_{T-1}}^2, \qquad \Xi := 2^8 \cdot 3L_{f_{T-1}}^4 + 2^7 \cdot 3\varkappa_{J_{T-1}}^4 + 2^4 \cdot 3\sigma_{J_{T-1}}^4,$$

$$\Psi := \prod_{\ell=T-1}^{T} 18L_{f_\ell}^2 + 8\varrho_{J_\ell}^2 + 4\sigma_{J_\ell}^2, \qquad \Omega := \prod_{\ell=T-1}^{T} 2^8 \cdot 3L_{f_\ell}^4 + 2^7 \cdot 3\varkappa_{J_\ell}^4 + 12\sigma_{J_\ell}^4.$$

Before proceeding, we remark the order of $\tilde{c}_i$ and c_i could be $\mathcal{O}(C^T)$ for some universal constant $C > 1$. We did not try to optimize the constants appearing in the definition of $\tilde{c}_i$ and c_i, as our main focus in this work is on the convergence rates.

Proof of lemma 2.2.4. First, we start with some notations. Recall the definitions of $A_{k,i}, e_i^{k+1}, \hat{e}_i^{k+1}$ and define for $1 \le i \le T$,

$$D_{k,i} := A_{k,i} + \tau_k e_i^{k+1} + \hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k). \tag{2.57}$$

Then, we have for $i \leq i \leq T$,

$$f_i(w_{i+1}^{k+1}) - w_i^{k+1} = (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k) + D_{k,i}. \tag{2.58}$$

We now prove eq. (2.55). By equation eq. (2.58), Lemma 2.2.2 using $\delta = \tau_k$, we obtain

$$\begin{aligned}
\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 &\leq (1-\tau_k^2)(1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \frac{(1+\tau_k)}{\tau_k}\|D_{k,i}\|^2 \\
&\leq (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \frac{2}{\tau_k}\|D_{k,i}\|^2. \qquad (2.59)
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\|D_{k,i}\|^2 &= \|A_{k,i}\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + \|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2 + 2\tilde{r}_{k,i}, \qquad (2.60) \\
r'_{k,i} &= \langle A_{k,i}, \tau_k e_i^{k+1} + \hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle + \tau_k\langle e_i^{k+1}, \hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle
\end{aligned}$$

which together with the fact that $\mathbb{E}[\tilde{r}_{k,i}|\mathscr{F}_k] = 0$ under Assumption 2, we have imply that

$$\begin{aligned}
\mathbb{E}[\|D_{k,i}\|^2|\mathscr{F}_k] &= \mathbb{E}[\|A_{k,i}\|^2|\mathscr{F}_k] + \tau_k^2\mathbb{E}[\|e_i^{k+1}\|^2|\mathscr{F}_k] + \mathbb{E}[\|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2|\mathscr{F}_k] \\
&\leq \tau_k^2\mathbb{E}[\|e_i^{k+1}\|^2|\mathscr{F}_k] + \left(4L_{f_i}^2 + \mathbb{E}[\|\hat{e}_i^{k+1}\|^2|\mathscr{F}_k]\right)\mathbb{E}[\|w_{i+1}^{k+1} - w_{i+1}^k\|^2|\mathscr{F}_k], \qquad (2.61)
\end{aligned}$$

where the second inequality follows from (2.52). Hence, noting the result from Proposition 2.2.1.a), $w_{T+1}^k = x^k$, and under Assumption 3, we have

$$\mathbb{E}[\|D_{k,T}\|^2|\mathscr{F}_k] \leq \tau_k^2\left[\sigma_{G_T}^2 + \left(4L_{f_T}^2 + \varrho_{J_T}^2\right)\left(\prod_{i=1}^T \sigma_{J_i}^2\right)\beta^{-2}\right].$$

Using (2.59) with $i = T$, the above inequality, and Lemma 2.1.8 with the choice of $\tau_0 = 1$, we have

$$\mathbb{E}[\|f_T(x^k) - w_T^k\|^2|\mathscr{F}_k] \leq 2\left[\sigma_{G_T}^2 + \left(4L_{f_T}^2 + \varrho_{J_T}^2\right)\left(\prod_{i=1}^T \sigma_{J_i}^2\right)\beta^{-2}\right]. \tag{2.62}$$

Moreover, under Assumption 3 and Lemma 2.2.1.b), we have

$$\mathbb{E}[\|w_{i+1}^{k+1} - w_i^k\|^2|\mathscr{F}_k] \leq \tau_k^2\mathbb{E}\left[2\|f_i(w_{i+1}^k) - w_i^k\|^2 + \|e_i^{k+1}\|^2 + \frac{2}{\tau_k^2}\|J_i^{k+1}\|^2\|w_{i+1}^{k+1} - w_{i+1}^k\|^2\middle|\mathscr{F}_k\right], \tag{2.63}$$

implying that

$$\mathbb{E}[\|w_T^{k+1} - w_T^k\|^2|\mathscr{F}_k] \leq \tau_k^2 \left[5\sigma_{G_T}^2 + 2(8L_{f_T}^2 + 2\varrho_{J_T}^2 + \sigma_{J_T}^2)\left(\prod_{i=1}^{T}\sigma_{J_i}^2\right)\beta^{-2}\right]. \tag{2.64}$$

This completes the proof of eq. (2.55) 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 $\mathbb{E}[\|w_{i+1}^{k+1} - w_{i+1}^k\|^2|\mathscr{F}_k] \leq \tilde{c}_{i+1}\tau_k^2$ for some $1 \leq i < T$. Hence, by eq. (2.60) and under Assumption 3, we have

$$\mathbb{E}[\|D_{k,i}\|^2|\mathscr{F}_k] \leq \tau_k^2[\sigma_{G_i}^2 + (4L_{f_i}^2 + \varrho_{J_i}^2)\tilde{c}_{i+1}],$$

which together with Lemma 2.1.8, imply that

$$\mathbb{E}[\|f_i(w_{i+1}^k) - w_i^k\|^2|\mathscr{F}_k] \leq 2[\sigma_{G_i}^2 + (4L_{f_i}^2 + \varrho_{J_i}^2)\tilde{c}_{i+1}].$$

Thus, by eq. (2.63), we obtain

$$\mathbb{E}[\|w_i^{k+1} - w_i^k\|^2|\mathscr{F}_k] \leq \tau_k^2[5\sigma_{G_i}^2 + 2(4L_{f_i}^2 + \varrho_{J_i}^2 + 2\sigma_{J_i}^2)\tilde{c}_{i+1}],$$

where after using Lemma 2.2.3, $\tilde{c}_i$ for $1 \leq i \leq T-2$, is as defined in the statement of Lemma 2.2.4. Hence, we obtain the claim in eq. (2.55) by induction.

We now start proving eq. (2.56). We start with $i = T$. By equation eq. (2.58), Lemma 2.2.2 and setting $\delta = \tau_k$ we get

$$\begin{aligned}\|f_T(x^{k+1}) - w_T^{k+1}\|^4 &\leq (1-\tau_k^2)^3(1-\tau_k)\|f_T(x^k) - w_T^k\|^4 + \frac{(1+\tau_k)^3}{\tau_k^3}\|D_{k,T}\|^4 \\ &\leq (1-\tau_k)\|f_T(x^k) - w_T^k\|^4 + \frac{8}{\tau_k^3}\|D_{k,T}\|^4.\end{aligned}$$

Now, by eq. (2.60), we have

$$\begin{aligned}
\|D_{k,i}\|^4 &= \|A_{k,i}\|^4 + \tau_k^4\|e_i^{k+1}\|^4 + \|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^4 + 4r_{k,i}'^2 + 2\tau_k^2\|e_i^{k+1}\|^2\|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2 \\
&\quad + 2\|A_{k,i}\|^2\left(\tau_k^2\|e_i^{k+1}\|^2 + \|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2\right) \\
&\quad + 4r_{k,i}'\left(\|A_{k,i}\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + \|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2\right), \\
r_{k,i}'^2 &\le 2\|A_{k,i}\|^2\left(\tau_k^2\|e_i^{k+1}\|^2 + \|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2 + 2\tau_k\langle e_i^{k+1}, \hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle\right) \\
&\quad + 2\tau_k^2\|e_i^{k+1}\|^2\|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2.
\end{aligned}$$

implying that

$$\begin{aligned}
\|D_{k,i}\|^4 &\le \|A_{k,i}\|^4 + \tau_k^4\|e_i^{k+1}\|^4 + \|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^4 + 4\tau_k^2\|e_i^{k+1}\|^2\|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2 \\
&\quad + 4\|A_{k,i}\|^2\left(\tau_k^2\|e_i^{k+1}\|^2 + \|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2\right) + 4r_{k,i}'', \qquad (2.65) \\
r_{k,i}'' &= r_{k,i}'\left(\|A_{k,i}\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + \|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2\right) + \tau_k\|A_{k,i}\|^2\langle e_i^{k+1}, \hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle.
\end{aligned}$$

By definition of d^k and Assumption 3, we obtain $\|A_{k,T}\| \le 2L_{f_T}\tau_k\|d^k\|$. By this inequality and by applying Lemma 2.2.2 with $\delta = 1$, we have

$$\begin{aligned}
\|D_{k,T}\|^4 &\le 8\left[\|A_{k,T}\|^4 + \tau_k^4\|e_T^{k+1} + \hat{e}_T^{k+1}d^k\|^4\right] \\
&\le 8\tau_k^4[16L_{f_T}^4\|d^k\|^4 + \|e_T^{k+1} + \hat{e}_T^{k+1}d^k\|^4] \\
&\le 64\tau_k^4[2L_{f_T}^4\|d^k\|^4 + \|e_T^{k+1}\|^4 + \|d^k\|^4\|\hat{e}_T^{k+1}\|^4].
\end{aligned}$$

By Assumption 3 and Proposition 2.2.1, we have

$$\mathbb{E}[\|D_{k,T}\|^4|\mathscr{F}_k] \le 64\tau_k^4[2L_{f_T}^4\sigma_d + \kappa_{G_T}^4 + \varkappa_{J_T}^4\sigma_d].$$

Hence, by Lemma 2.1.8, we obtain

$$\mathbb{E}[\|f_T(x^k) - w_T^k\|^4|\mathscr{F}_k] \le 8^3[2L_{f_T}^4\sigma_d + \kappa_{G_T}^4 + \varkappa_{J_T}^4\sigma_d].$$

Now, by Assumption 3 and Lemma 2.2.1, we

$$\mathbb{E}[\|w_T^{k+1} - w_T^k\|^4 | \mathscr{F}_k] \le \tau_k^4 [3072\{2L_{f_T}^4 \sigma_d + \kappa_{G_T}^4 + \varkappa_{J_T}^4 \sigma_d\} + 40\sigma_d \sigma_{J_T}^4 + 35 \cdot \kappa_{G_T}^4].$$

This completes the proof of eq. (2.56) when $i = T$. We now use induction to complete the proof. By the above result, the base case of $i = T$ holds. Assume that $\mathbb{E}[\|w_{i+1}^{k+1} - w_{i+1}^k\|^4 | \mathscr{F}_k] \le c_{i+1}\tau_k^4$, for some $1 \le i < T$. Then, note that by using eq. (2.58), we have

$$\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^4 \le (1 - \tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^4 + \left(\frac{1+\tau_k}{\tau_k}\right)^3 \|D_{k,i}\|^4.$$

Since f_i is Lipschitz under Assumption 3, $\|A_{k,i}\| \le 2L_{f_i}\|w_{i+1}^{k+1} - w_{i+1}^k\|$. Using this fact and Lemma 2.2.2 with $\delta = 1$, in eq. (2.65), we obtain

$$\mathbb{E}[\|D_{k,i}\|^4 | \mathscr{F}_k] \le 64\tau_k^4 [2L_{f_i}^4 c_{i+1} + \kappa_{G_i}^4 + \varkappa_{J_i}^4 c_{i+1}].$$

Using the above inequality, Lemma 2.1.8, and our setting $\tau_0 = 1$, we obtain

$$\mathbb{E}[\|f_i(w_{i+1}^k) - w_i^k\|^4 | \mathscr{F}_k] \le 8^3 [2L_{f_i}^4 c_{i+1} + \kappa_{G_i}^4 + \varkappa_{J_i}^4 c_{i+1}].$$

By Assumption 3 and Lemma 2.2.1, we obtain

$$\mathbb{E}[\|w_i^{k+1} - w_i^k\|^4 | \mathscr{F}_k] \le \tau_k^4 [3072[2L_{f_i}^4 c_{i+1} + \kappa_{G_i}^4 + \varkappa_{J_i}^4 c_{i+1}] + \kappa_{G_i}^4 + 4\kappa_{J_i}^4 c_{i+1}],$$

where after using Lemma 2.2.3, c_i for $1 \le i \le T-2$, is as defined in the statement of Lemma 2.2.4. Hence, we obtain the claim in eq. (2.56) by induction.

■

The next result is the counterpart of Lemma 2.1.7 for Algorithm 2.

Lemma 2.2.5. Recall the definition of the merit function in eq. (2.15). Define $w^k := (w_1^k, \ldots, w_T^k)$ for $k \ge 0$. Let $\{x^k, z^k, u^k, w_1^k, \ldots, w_T^k\}_{k \ge 0}$ be the sequence generated by Algorithm 2. Suppose for $1 \le i \le T$, we have

$$\max_{2 \le j \le T} C_j^2 \le \frac{(\beta_k - \lambda)}{T}(\gamma_i b - \lambda) \tag{2.66}$$

where C_j's are defined in Lemma 2.1.4. Then, under Assumption 1 and Assumption 3, we have

$$\lambda \sum_{k=0}^{N-1} \tau_k \left[\|d^k\|^2 + \sum_{i=1}^{T-1} \|f_i(w_{i+1}^k) - w_i^k\|^2 + \|f_T(x^k) - w_T^k\|^2 \right] \le W(x^0, z^0, w^0) + \sum_{k=0}^{N-1} \hat{R}^{k+1}, \quad (2.67)$$

where, for any $k \ge 0$,

$$\hat{R}^{k+1} := \left(\sum_{i=1}^{T} \gamma_i \hat{r}_i^{k+1} \right) + \frac{\tau_k^2}{2} \left[(L_{\nabla F} + L_{\nabla \eta} + 2C_T L_{f_T}) \|d^k\|^2 \right] + \tau_k \langle d^k, \Delta_k \rangle + \frac{L_{\nabla \eta}}{2} \|z^{k+1} - z^k\|^2,$$

$$\hat{r}_i^{k+1} = \frac{L_{\nabla f_i}^2}{4\tau_k} \|w_{i+1}^{k+1} - w_{i+1}^k\|^4 + \|\hat{e}_i^{k+1}\|^2 \|w_{i+1}^{k+1} - w_{i+1}^k\|^2 + \tau_k^2 \|e_i^{k+1}\|^2 + \dot{r}_i^{k+1},$$

and Δ_k and $\dot{r}_i^{k+1}$ are, respectively, defined in (2.19) and (2.51). Furthermore, notice that eq. (2.66) is satisfied, when we pick

$$\gamma_i = 1, \qquad \lambda = 1/2, \qquad \beta_k \equiv \beta \ge \frac{1}{2} + 2T \max_{2 \le j \le T} C_j^2. \quad (2.68)$$

Proof. Noting Lemma 2.2.1 and definition of $\hat{r}_i^{k+1}$, we have

$$\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 - \|f_i(w_{i+1}^k) - w_i^k\|^2 \le -\tau_k \|f_i(w_{i+1}^k) - w_i^k\|^2 + \hat{r}_i^{k+1},$$

$$\|f_T(x^{k+1}) - w_T^{k+1}\|^2 - \|f_T(x^k) - w_T^k\|^2 \le -\tau_k \|f_T(x^k) - w_T^k\|^2 + \hat{r}_T^{k+1}.$$

Combining the above inequalities with (2.21), (2.23), and noting definition of the merit function in (2.15), we obtain

$$
\begin{aligned}
&W(x^{k+1}, z^{k+1}, w^{k+1}) - W(x^k, z^k, w^k) \\
\leq &- \beta_k \tau_k \|d^k\|^2 + \sum_{j=2}^{T-1} \tau_k C_j \|d^k\| \|f_j(w_{j+1}^k) - w_j^k\| + \tau_k C_T \|d^k\| \|f_T(x^k) - w_T^k\| \\
+ &\sum_{i=1}^{T-1} -\gamma_i \tau_k \|f_i(w_{i+1}^k) - w_i^k\|^2 - \gamma_T \tau_k \|f_T(x^k) - w_T^k\|^2 + R^{k+1} \\
\leq &- \beta_k \tau_k \|d^k\|^2 + \sum_{j=1}^{T-1} \tau_k \sqrt{\left(\frac{\beta_k - \lambda}{T}\right)(\gamma_j - \lambda)} \|d^k\| \|f_j(w_{j+1}^k) - w_j^k\| \\
+ &\tau_k \sqrt{\left(\frac{\beta_k - \lambda}{T}\right)(\gamma_T - \lambda)} \|d^k\| \|f_T(x^k) - w_T^k\| \\
+ &\sum_{i=1}^{T-1} -\gamma_i \tau_k \|f_i(w_{i+1}^k) - w_i^k\|^2 - \gamma_T \tau_k \|f_T(x^k) - w_T^k\|^2 + R^{k+1} \\
\leq &- \lambda \tau_k [\|d^k\|^2 + \sum_{i=1}^{T-1} \|f_j(w_{j+1}^k) - w_j^k\|^2 + \|f_T(x^k) - w_T^k\|^2] + R^{k+1},
\end{aligned}
$$

where the second to the last inequality follows by condition eq. (2.66) and last follows by Young's inequality. Thus, by summing up the above inequalities and re-arranging the terms, we obtain (2.67). also It is easy to see that eq. (2.66) holds, by picking the parameters as in eq. (2.68). ■

In the next result, we show the error terms in the right hand side of (2.67) is bounded in the order of $\sum_{k=1}^N \tau_k^2$ in expectation.

Proposition 2.2.1. *Suppose $\beta_k = \beta > 0$ for all k and $\tau_0 = 1$. We then have*

$$
\begin{aligned}
&\beta^4 \mathbb{E}[\|d^k\|^4 | \mathscr{F}_k] \leq \mathbb{E}[\|z^k\|^4 | \mathscr{F}_k] \leq \prod_{i=1}^T \kappa_{J_i}^4 := \beta^4 \sigma_d \quad \forall k \geq 1, \\
&\mathbb{E}[\hat{R}^{k+1} | \mathscr{F}_k] \leq \hat{\sigma}^2 \tau_k^2,
\end{aligned}
$$

where

$$\hat{\sigma}^2 := \sum_{i=1}^{T-1} \gamma_i \left(\frac{L^2_{\nabla f_i} c_{i+1}}{4} + \varrho^2_{J_i} \tilde{c}_{i+1} + \sigma^2_{G_i} \right) + \frac{\gamma_T L^2_{\nabla f_T} \sigma_d}{4} + 2L_{\nabla \eta} \left(\prod_{i=1}^{T} \sigma^2_{J_i} \right)$$
$$+ \frac{1}{2} \left[2\gamma_T \sigma^2_{J_T} + \frac{1}{\beta_k^2} \left(\prod_{i=1}^{T} \sigma^2_{J_i} \right) \left\{ 2\gamma_T \varrho^2_{J_T} + L_{\nabla F} + L_{\nabla \eta} + 2C_T L_{f_T} \right\} \right]. \tag{2.69}$$

Proof. Noting the convexity of $\|\cdot\|^4$, the first inequality follows similarly to that of Proposition 2.1.1 and hence we omit the details. Noting $\mathbb{E}[\Delta_k|\mathscr{F}_k] = 0$, definition of R^{k+1}, $\mathbb{E}[\dot{r}_i^{k+1}|\mathscr{F}_k] = 0$ for $1 \le i \le T$, Lemma 2.2.5, Lemma 2.2.4 and Assumption 3, we obtain σ^2 as in eq. (2.69).

■

We remark that the c_{i+1} in the right hand side of (2.69) indeed appears as $\tau_k c_{i+1}$ and so τ_k reduces the affect of larger constants in the definition of c_{i+1}. However, for simplicity we just removed the τ_k in the definition of $\hat{\sigma}^2$. We are now ready to state the convergence rates via the following theorem.

Theorem 2.2.6. *Suppose that $\{x^k, z^k\}_{k\ge 0}$ are generated by Algorithm 2, and Assumption 1 and Assumption 3 hold. Also assume the parameters satisfy eq.* (2.68) *and the step sizes $\{\tau_k\}$ satisfy (2.36).*

(a) *The results in parts a) and b) of eq.* (2.37) *still hold by replacing σ^2 by $\hat{\sigma}^2$.*

b) *If the stepsizes are set to (2.41), the results of part c) of eq.* (2.37) *also hold with replacing σ^2 by $\hat{\sigma}^2$.*

Proof. The proof follows from the same arguments in the proof of eq. (2.37) by noticing (2.67), and Proposition 2.2.1, hence, we skip the details.

■

Remark 2.2.1. *Note that Algorithm 2 does not use a mini-batch of samples in any iteration. Thus, (2.43) (in which σ^2 is replaced by $\hat{\sigma}^2$) implies that the total sample complexity of Algorithm 2 for finding an ϵ-stationary point of eq.* (1.5)*, is bounded by $\mathcal{O}(c^T T^6/\epsilon^4)$ which is better in the order of magnitude than the complexity bound of Algorithm 1. Furthermore, this bound matches the complexity bound obtained in [63] for the two-level composite problem which in turn is in the same order for single-level smooth stochastic optimization.*

2.3. Concluding remarks

In this project, we proposed two algorithms, with level-independent convergence rates, for stochastic multi-level composition optimization problems under the availability of a certain stochastic first-order oracle. We show that under a bounded second moment assumption on the outputs of the stochastic oracle, our first proposed algorithm, by using a mini-batch of samples in each iteration, achieves a sample complexity of $\mathcal{O}(1/\epsilon^6)$ for finding an ϵ-stationary point of the multi-level composite problem. By modifying this algorithm and making a bounded fourth moment assumption, we show that we can improve the sample complexity to $\mathcal{O}(1/\epsilon^4)$ which seems to be unimprovable even for single-level stochastic optimization problems, without further assumptions [9, 43]. For future work, it is interesting to establish CLT and normal approximation results for the online algorithms we presented in this work for stochastic multi-level composition optimization problems, similar to the results in [6, 39, 103, 108, 130] for the standard stochastic gradient algorithm when $T = 1$.

CHAPTER 3

Stochastic Zeroth-order Functional Constrained Optimization

Notations: Let $\mathbf{0}$ denote the vector of elements 0 and $[m] \coloneqq \{1, \ldots, m\}$. Let $f(x) \coloneqq [f_1(x), \ldots, f_m(x)]^T$; then, the constraints in (1.7) be expressed as $f(x) \leqslant \mathbf{0}$. We use $\xi \coloneqq [\xi_1, \cdots, \xi_m]$ to denote the random vectors in the constraints. Furthermore, $\|\cdot\|$ denotes a general norm and $\|\cdot\|_*$ denotes its dual norm defined as $\|z\|_* \coloneqq \sup\{z^T x : \|x\| \leq 1\}$. Furthermore, $[x]_+ \coloneqq \max\{x, 0\}$ for any $x \in \mathbb{R}$. For any vector $x \in \mathbb{R}^k$, we define $[x]_+$ as element-wise application of the operator $[\cdot]_+$.

3.1. Preliminaries

We first describe the precise assumptions to be made on the *stochastic zeroth-order oracle* in this work.

Assumption 4. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. For $i \in \{0, \ldots, m\}$ and for any $x \in \mathbb{R}^n$, the zeroth-order oracle outputs an estimator $F_i(x, \xi_i)$ of $f_i(x)$ such that $\mathbb{E}[F_i(x, \xi_i)] = f_i(x)$, $\mathbb{E}[F_i(x, \xi_i)^2] \leq \sigma_{f_i}^2$, $\mathbb{E}[\nabla F_i(x, \xi_i)] = \nabla f_i(x)$, $\mathbb{E}[\|\nabla F_i(x, \xi_i) - \nabla f_i(x)\|_*^2] \leqslant \sigma_i^2$, where $\|\cdot\|_*$ denotes the dual norm.*

The assumption above assumes that we have accesses to a stochastic zeroth-order oracle which provides unbiased function evaluations with bounded variance. It is worth noting that in the above assumption, we do not necessarily assume the noise ξ_i is additive. Furthermore, we allow for different noise models for the objective function and the m constraint functions, which is a significantly general model compared to several existing works [40]. Our gradient estimator is then constructed by leveraging the Gaussian smoothing technique [98, 99]. For $\nu_i \in (0, \infty)$ we introduce the smoothed function $f_{i,\nu_i}(x) = \mathbb{E}_{u_i}[f_i(x + \nu_i u_i)]$ where $u_i \sim N(0, I_n)$ and independent across i. We can estimate the gradient of this smoothed function using function evaluations of f_i. Specifically, we define the stochastic zeroth-order gradient of $f_{i,\nu_i}(x)$ as

$$G_{i,\nu_i}(x, \xi_i, u_i) = \frac{F_i(x + \nu_i u_i, \xi_i) - F_i(x, \xi_i)}{\nu_i} u_i, \tag{3.1}$$

which is an unbiased estimator of $\nabla f_{i,\nu_i}(x)$, i.e., we have $\mathbb{E}_{u,\xi_i}[G_{i,\nu_i}(x, \xi_i, u)] = \nabla f_{i,\nu_i}(x)$. However, it is well-known that $G_{i,\nu_i}(x, \xi_i, u_i)$ is a biased estimator of $\nabla f_i(x)$. An interpretation of the gradient estimator in (3.1) as a consequence of Gaussian Stein's identity, popular in the statistics literature [122], was provided in [20].

The gradient estimator in (3.1) is referred to as the two-point estimator in the literature. The reason is that, for a given random vector ξ_i, it is assumed that the stochastic function in (3.1) could be evaluated at two points, $F(x + \nu_i u_i, \xi_i)$ and $F(x, \xi_i)$. Such an assumption is satisfied in several statistics, machine learning and simulation based optimization and sampling problems; see for example [2, 41, 46, 62, 94, 99, 120]. Yet another estimator in the literature is the one-point estimator, which assumes that for each ξ_i, we observe only one noisy function evaluation $F(x + \nu_i u_i, \xi_i)$. It is well-known that the one-point setting is more challenging than the two-point setting [114]. From a theoretical point of view, the use of two-point evaluation based gradient estimator is primarily motivated by the sub-optimality (in terms of oracle complexity) of one-point feedback based stochastic zeroth-order optimization methods either in terms of the approximation accuracy or dimension dependency. For the rest of this work, we focus on the two-point setting and leave the question of obtaining results in the one-point setting as future work. We now describe our assumptions on the objective and constraint functions.

Assumption 5. *Function F_i has Lipschitz continuous gradient with constant L_i, almost surely for any ξ_i, i.e., $\|\nabla F_i(y, \xi_i) - \nabla F_i(x, \xi_i)\|_* \leqslant L_i \|y - x\|$, which consequently implies that $|F_i(y, \xi_i) - F_i(x, \xi_i) - \langle \nabla F_i(x, \xi_i), y - x\rangle| \leqslant \frac{L_i}{2}\|y - x\|^2$ for $i \in \{0, 1, \ldots, m\}$.*

Assumption 6. *Function F_i is Lipschitz continuous with constant M_i, almost surely for any ξ_i, i.e., $|F_i(y, \xi_i) - F_i(x, \xi_i)| \leq M_i \|y - x\|$, for $i \in \{0, 1, \ldots, m\}$.*

The above smoothness assumptions are standard in the literature on stochastic zeroth-order optimization and are made in several works [20, 62, 99] for obtaining oracle complexity results. It is easy to see that Assumption 5 implies that for $i \in \{0, \ldots, m\}$, f_i has Lipschitz continuous gradient with constant L_i since $\|\nabla f_i(y) - \nabla f_i(x)\|_* \leqslant \mathbb{E}[\|\nabla F(y,\xi) - \nabla F(x,\xi)\|_*] \leqslant L_i\|y-x\|$, due to Jensen's inequality for the dual norm. By similar reasoning and Assumption 6, we also see that f_i is Lipschitz continuous with constant M_i. Due to Assumptions 5 and 6, we also have the following:

$$\begin{aligned} \|f(x_1) - f(x_2)\|_2 &\leqslant M_f\|x_1 - x_2\|, \\ \|f(x_1) - f(x_2) - \nabla f(x_2)^T(x_1 - x_2)\|_2 &\leqslant \frac{L_f}{2}\|x_1 - x_2\|^2, \\ \|\nabla f(x_2)^T(x_1 - x_2)\|_2 &\leqslant M_f\|x_1 - x_2\|, \end{aligned} \tag{3.2}$$

for all $x_1, x_2 \in \mathbb{R}^n$, where $\nabla f(\cdot) := [\nabla f_1(\cdot), \ldots, \nabla f_m(\cdot)] \in \mathbb{R}^{n\times m}$ and constants M_f and L_f are defined as

$$M_f := \sqrt{\textstyle\sum_{i=1}^m M_i^2} \text{ and } L_f := \sqrt{\textstyle\sum_{i=1}^m L_i^2}. \tag{3.3}$$

We now state the definition of the **prox**-function and the **prox**-operator. The class of algorithms based on **prox**-operators are called as proximal algorithms. Such algorithms have turned out to be particularly useful for efficiently solving various machine learning problems in the recent past. We refer the interested reader to [21, 101] for more details.

Definition 3.1.1. Let $\omega : X \to \mathbb{R}$ be continuously differentiable, L_ω-Lipschitz gradient smooth, and 1-strongly convex with respect to $\|\cdot\|$ function. We define the **prox**-function associated with $\omega(\cdot)$, $\forall x, y \in \mathbb{R}^n$, as

$$W(y,x) := \omega(y) - \omega(x) - \langle \nabla\omega(x), y - x\rangle. \tag{3.4}$$

Based on the smoothness and strong convexity of $\omega(x)$, we have the following relation, $\forall x, y \in \mathbb{R}^n$:

$$W(y,x) \leqslant \frac{L_\omega}{2}\|x-y\|^2 \leqslant L_\omega W(x,y). \tag{3.5}$$

For any $v \in \mathbb{R}^n$, we define the following **prox**-operator

$$\mathbf{prox}(v, \tilde{x}, \eta) := \operatorname*{arg\,min}_{x\in X}\{\langle v, x\rangle + \eta W(x, \tilde{x})\}. \tag{3.6}$$

The function W is also called as Bregman divergence in the literature. A canonical example of W is that of the Euclidean distance function $\|x-y\|^2$ which is useful when $X=\mathbb{R}^n$. We will see in Section 3.2.1 that our algorithm is based on the above **prox**-operator.

Finally, we have the following results which will prove to be useful for subsequent calculations. Let $u := [u_1, \cdots, u_m]$ and $D_X := \sup_{x,y} \sqrt{W(x,y)}$ be the diameter of the set X. We start with the following well-known result on the stochastic zeroth-order gradient estimator in (3.1).

THEOREM 3.1.1 ([99]). *For a Gaussian random vector $u \sim N(0, I_n)$ we have*

$$\mathbb{E}[\|u\|^k] \leqslant (n+k)^{k/2} \tag{3.7}$$

for any $k \geqslant 2$. Moreover, the following statements hold for any function ψ whose gradient is Lipschitz continuous with constant L

a) The gradient of $\psi_\nu(x) := \mathbb{E}_u[\psi(x+\nu u)]$ is Lipschitz continuous with constant L_ν such that $L_\nu \leqslant L$.

b) For any $x \in \mathbb{R}^n$, we have

$$|\psi_\nu(x) - \psi(x)| \leqslant \frac{\nu^2}{2} Ln, \tag{3.8}$$

$$\|\nabla\psi_\nu(x) - \nabla\psi(x)\| \leqslant \frac{\nu}{2} L(n+3)^{3/2}. \tag{3.9}$$

c) For any $x \in \mathbb{R}^n$, we have

$$\frac{1}{\nu^2}\mathbb{E}_u[\{\psi(x+\nu u) - \psi(x)\}^2\|u\|^2] \leqslant \frac{\nu^2}{2} L^2(n+6)^3 + 2(n+4)\|\nabla\psi(x)\|^2. \tag{3.10}$$

Lemma 3.1.2. Let $\nu := [\nu_1, \cdots, \nu_m]$, $F_\nu(x,\xi,u) := [F_1(x+\nu_1 u_1, \xi_1), \ldots, F_m(x+\nu_m u_m, \xi_m)]^T$ and $f_\nu(x) := [f_{1,\nu_1}(x), \ldots, f_{m,\nu_m}(x)]^T$. Under Assumption 6, we have

$$\mathbb{E}_{u,\xi}[\|F_\nu(x,\xi,u) - f_\nu(x)\|^2] \leqslant \sigma_{f,\nu}^2, \tag{3.11}$$

where $\sigma_{f,\nu}^2 := (\sum_{i=1}^m 4(n+2)M_i^2\nu_i^2 + L_i^2\nu_i^4 n^2) + 2\sigma_f^2$, where $\sigma_f^2 = \sum_{i=1}^m \sigma_{f_i}^2$.

PROOF OF LEMMA 3.1.2. Note that

$$\|F_\nu(x,\xi,u) - f_\nu(x)\|^2 = \textstyle\sum_{i=1}^m (f_{i,\nu_i}(x) - F_i(x+\nu_i u, \xi))^2.$$

By Young's inequality, we have

$$
\begin{aligned}
|F_i(x+\nu_i u,\xi) - f_{i,\nu_i}(x)|^2 &= |[F_i(x+\nu_i u,\xi) - F_i(x,\xi)] + [F_i(x,\xi) - f_i(x)] + [f_i(x) - f_{i,\nu_i}(x)]|^2 \\
&\leqslant 4|F_i(x+\nu_i u,\xi) - F_i(x,\xi)|^2 + 4|f_i(x) - f_{i,\nu_i}(x)|^2 + 2|F_i(x,\xi) - f_i(x)|^2 \\
&\leqslant 4M_i^2\nu_i^2\|u\|^2 + 4\left(\frac{\nu_i^2}{2}L_i n\right)^2 + 2|F_i(x,\xi) - f_i(x)|^2.
\end{aligned}
$$

Now, by Assumption 6 and Theorem 3.1.1, we have

$$
\mathbb{E}|f_{i,\nu_i}(x) - F_i(x+\nu_i u,\xi)|^2 \leq 4M_i^2\nu_i^2(n+2) + 2\sigma_{f,i}^2 + L_i^2\nu_i^4 n^2.
$$

Consequently, we obtain

$$
\mathbb{E}\|F_\nu(x,\xi,u) - f_\nu(x)\|^2 \leqslant (\textstyle\sum_{i=1}^m 4M_i^2\nu_i^2(n+2) + L_i^2\nu_i^4 n^2) + 2\sigma_f^2 =: \sigma_{f,\nu}^2.
$$

∎

Lemma 3.1.3. Under Assumptions 4 and 5, we have

$$
\mathbb{E}_{u,\xi}[\|G_{i,\nu_i}(x,\xi,u) - \nabla f_{i,\nu_i}(x)\|^2] \leqslant \sigma_{i,\nu_i}^2 \tag{3.12}
$$

where $\sigma_{i,\nu_i}^2 \coloneqq \nu_i^2 L_i^2(n+6)^3 + 10(n+4)[\sigma_i^2 + \tilde{B}_i^2]$, with $\tilde{B}_i := \frac{\nu_i}{2}L_i(n+3)^{3/2} + L_i D_X + M_i$.

Proof of Lemma 3.1.3. First note that by Theorem 3.1.1, we have

$$
\begin{aligned}
\frac{1}{\nu_i^2}\mathbb{E}_u[\{F_i(x+\nu_i u,\xi) - F_i(x,\xi)\}^2\|u\|^2] &\leqslant \frac{\nu_i^2}{2}L_i^2(n+6)^3 + 2(n+4)\|\nabla F_i(x,\xi)\|^2 \\
&\leqslant \frac{\nu_i^2}{2}L_i^2(n+6)^3 + 4(n+4)[\|\nabla F_i(x,\xi) - \nabla f_i(x)\|^2 \\
&\quad + \|\nabla f_i(x)\|^2].
\end{aligned} \tag{3.13}
$$

Next note that

$$
\begin{aligned}
\|\nabla f_{i,\nu_i}(x)\| &\leqslant \|\nabla f_{i,\nu_i}(x) - \nabla f_i(x)\| + \|\nabla f_i(x)\| \\
&\leqslant \frac{\nu_i}{2}L_i(n+3)^{3/2} + L_i D_X + \|\nabla f_i(x^*)\| \\
&\leqslant \frac{\nu_i}{2}L_i(n+3)^{3/2} + L_i D_X + M_i =: \tilde{B}_i,
\end{aligned}
$$

where M_i is from Assumption 6. Taking the expectation with respect to ξ on both sides of (3.13), we have

$$\mathbb{E}[||G_{i,\nu_i}(x,\xi,u)\|^2] \leqslant \frac{\nu_i^2}{2} L_i^2(n+6)^3 + 4(n+4)[\sigma_i^2 + \tilde{B}_i^2].$$

From the above inequalities, using Assumptions 5 and 6, Theorem 3.1.1, and Young's inequality, we have

$$\begin{aligned}\mathbb{E}[\|G_{i,\nu_i}(x,\xi,u) - \nabla f_{i,\nu_i}(x)\|^2] &\leqslant 2\mathbb{E}[\|G_{i,\nu_i}(x,\xi,u)\|^2] + 2\|\nabla f_{i,\nu_i}(x)\|^2 \\ &\leqslant \nu_i^2 L_i^2(n+6)^3 + 8(n+4)[\sigma_i^2 + \tilde{B}_i^2] + 2\tilde{B}_i^2 \\ &\leqslant \nu_i^2 L_i^2(n+6)^3 + 10(n+4)[\sigma_i^2 + \tilde{B}_i^2],\end{aligned}$$

which completes the proof. ■

3.2. Stochastic Zeroth-order Constraint Extrapolation Method

In this section, we present our algorithm for solving the stochastic zeroth-order functional constrained optimization problem (1.7). In order to extend the method in [25] to the zeroth-order setting, we make several modifications to their framework that we illustrate below, and use the Gaussian smoothing based gradient estimates to handle the unavailability of gradients. The main challenge to overcome for our theoretical analysis is setting the choice of tuning parameters to mitigate the bias present in the stochastic zeroth-order stochastic gradient estimates. We emphasize that this becomes a non-trivial problem due to the fact that both the objective and the constraint functions are only accessible through noisy function evaluations.

3.2.1. Algorithmic Methodology. The constraint extrapolation framework of [25] is a novel primal-dual method that proceeds by (i) considering the Lagrangian formulation of (1.7), (ii) constructing linear approximations for the constraint functions, and (iii) constructing an *extrapolation operation* which enables acceleration. Such an approach has the advantage that: (i) it does not require the projection of Lagrangian multipliers onto a possibly unknown bounded set (which is required by several other primal-dual methods), (ii) it is a single-loop algorithm with a built-in acceleration step. [25] showed that such an approach helps achieve better rate of convergence than existing methods for solving (1.7) in the stochastic first-order setting.

The Lagrangian of (1.7) is given by

$$\min_{x\in X}\max_{y\geqslant \mathbf{0}}\{\mathcal{L}(x,y) := f_0(x) + \textstyle\sum_{i=1}^m \; y_i f_i(x)\}. \tag{3.14}$$

In other words, (x^*, y^*) is a *saddle point* of the Lagrange function $\mathcal{L}(x,y)$ such that

$$\mathcal{L}(x^*, y) \leqslant \mathcal{L}(x^*, y^*) \leqslant \mathcal{L}(x, y^*), \tag{3.15}$$

for all $x \in X, y \geqslant \mathbf{0}$, whenever the optimal dual, y^*, exists. Throughout this work, we assume the existence of y^* satisfying (3.15). In order to handle the zeroth-order setting, we also define Lagrangian with the smoothed functions as

$$\mathcal{L}_\nu(x,y) := f_{0,\nu_0}(x) + \textstyle\sum_{i=1}^m \; y_i f_{i,\nu_i}(x). \tag{3.16}$$

Now, we describe the linearization in the context of the iterates directly as it will be easier to understand in the stochastic setting that we are in. Let $x^{(t)}$ be the sequence of the algorithm (to be discussed later). The linearization of $f(\cdot)$ at the point $x^{(t)}$, with respect to the point $x^{(t-1)}$, is given by

$$\ell_f(x^{(t)}) := f_\nu(x^{(t-1)}) + \nabla f_\nu(x^{(t-1)})^T(x^{(t)} - x^{(t-1)}),$$

where similar to ∇f, we define $\nabla f_\nu(x^{(t-1)}) := [\nabla f_{1,\nu_1}(x^{(t-1)}), \ldots, \nabla f_{m,\nu_m}(x^{(t-1)})]$. For the implementation, we use the version of linearization with the Gaussian smoothing based stochastic zeroth-order gradients. In particular, we define

$$\begin{aligned}\ell_F(x^{(t)}) := F_\nu(&x^{(t-1)}, \bar\xi^{(t-1)}, \bar u^{(t-1)})\\ &+ G_\nu(x^{(t-1)}, \overline{\xi}^{(t-1)}, \overline{u}^{(t-1)})^T(x^{(t)} - x^{(t-1)}),\end{aligned}$$

where $G_\nu(x^{(t-1)}, \overline{\xi}^{(t-1)}, \overline{u}^{(t-1)}) \in \mathbb{R}^{n\times m}$ is given by

$$[G_{1,\nu_1}(x^{(t-1)}, \overline{\xi}_1^{(t-1)}, \overline{u}_1^{(t-1)}), \ldots, G_{m,\nu_m}(x^{(t-1)}, \overline{\xi}_m^{(t-1)}, \overline{u}_m^{(t-1)})].$$

Algorithm 3 Stochastic Zeroth-Order Constraint Extrapolation Method (`SZO-ConEx`)

Input: $\nu > \mathbf{0}, (x^{(0)}, y^{(0)}), \{\gamma_t, \tau_t, \eta_t, \theta_t\}_{t \geqslant 0}, T$.

1: Set $(x^{(-1)}, y^{(-1)}) \leftarrow (x^{(0)}, y^{(0)})$,
$F_\nu(x^{(-1)}, \overline{\xi}^{(-1)}, \overline{u}^{(-1)}) \leftarrow F_\nu(x^{(0)}, \overline{\xi}^{(0)}, \overline{u}^{(0)})$,
$\ell_F(x^{(-1)}) \leftarrow \ell_F(x^{(0)})$.

2: **for** $t = 0, \ldots, T-1$ **do**

3: $\quad s^{(t)} \leftarrow (1+\theta_t)\ell_F(x^{(t)}) - \theta_t \ell_F(x^{(t-1)})$.

4: $\quad y^{(t+1)} \leftarrow [y^{(t)} + \frac{1}{\tau_t} s^{(t)}]_+$.

5: $\quad x^{(t+1)} \leftarrow \mathbf{prox}\Big(G_{0,\nu_0}(x^{(t)}, \xi_0^{(t)}, u_0^{(t)})$
$+ \sum_{i=1}^m G_{i,\nu_i}(x^{(t)}, \xi_i^{(t)}, u_i^{(t)}) y_i^{(t+1)}, x^{(t)}, \eta_t \Big)$.

6: **return** $\bar{x}_T = (\sum_{t=0}^{T-1} \gamma_t)^{-1} \sum_{t=0}^{T-1} \gamma_t x^{(t+1)}$.

Here, by $\overline{\xi}^{(t-1)}, \overline{u}^{(t-1)}$ we mean an independent (of $\xi^{(t-1)}, u^{(t-1)}$, respectively) realization of random objects ξ, u, respectively.

Based on this, the overall procedure, termed as `SZO-ConEx` is provided in Algorithm 3. Step 3, which is based on the linearization discussed above, forms the main methodological innovation over existing primal-dual method. Step 4 and Step 5 respectively correspond to the gradient ascent step and the proximal gradient descent step to solve the saddle point problem in the Lagrangian formulation. At a high-level, the algorithm could be interpreted as using the constraint extrapolation method of [25] for solving (3.16), as the gradients used in Algorithm 3 are essentially unbiased estimators of the smoothed functions $f_{\nu,i}$ (for $i = 0, \ldots, m$). However, as the smoothing parameters ν_i (for $i = 0, \ldots, m$) tend to zero, $\mathcal{L}_\nu(x, y)$ converges to $\mathcal{L}(x, y)$ defined in (3.14). On the other hand, the parameters ν_i are in the denominator of the stochastic zeroth-order gradient estimators (see (3.1)). Hence, we cannot let them tend to zero at any arbitrary rate. Picking the ν_i to balance this tension forms the crux of our analysis. This also makes our analysis significantly more challenging and different from the stochastic first-order analysis of [25].

3.2.2. Convex Setting. We now provide our theoretical results for the case when the functions f_i, for $i = 0, \ldots, m$, are convex. We start by describing the measure of optimality we consider, for solving (1.7).

DEFINITION 3.2.1. A point $\bar{x}$ is an ϵ-approximately optimal solution in expectation, for (1.7), if it satisfies $\mathbb{E}[f_0(\bar{x}) - f_0^*] \leqslant \epsilon$ and $\mathbb{E}[\|[f(\bar{x})]_+\|_2] \leqslant \epsilon$, where f_0^* is the optimal value of (1.7) and the expectation is with respect to the randomness arising due to ξ_i and u_i across all iterations.

The first part of the above definition corresponds to the standard optimality condition for the convex problem. The next part corresponds to constraint violation. Our main result is described in Theorem 3.2.8. We define $M_X \coloneqq \sup_{x \in X} \|x\|$. Furthermore, we define $\sigma_\nu \coloneqq [\sigma_{1,\nu_1}, \cdots, \sigma_{m,\nu_m}]$, where σ_{i,ν_i}, for $i = 0, \ldots, m$ are as defined in Lemma 3.1.3, $\sigma_{X,f} \coloneqq (\sigma_{f,\nu}^2 + D_X^2 \|\sigma_\nu\|_2^2)^{1/2}$ (where $\sigma_{f,\nu}^2$ is as defined in Lemma 3.1.2).

Next, in order to obtain the oracle complexity of Algorithm 3, we define a primal-dual gap function for the equivalent saddle point problem (3.14). In particular, given a pair of feasible solution $z = (x, y)$ and $\bar{z} = (\bar{x}, \bar{y})$ of (3.14), we define the primal-dual gap function $Q(z, \bar{z})$ as

$$Q(z, \bar{z}) \coloneqq \mathcal{L}(x, \bar{y}) - \mathcal{L}(\bar{x}, y). \tag{3.17}$$

For the remainder of the project, we denote $Q_\nu(z, \bar{z}) = \mathcal{L}_\nu(x, \bar{y}) - \mathcal{L}_\nu(\bar{x}, y)$. Now we establish the error between these two functions.

Lemma 3.2.1. Under Assumptions 4, 5 and 6, we have

$$|Q(z, \bar{z}) - Q_\nu(z, \bar{z})| \leqslant \nu_0^2 L_0 n + M_X n (\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}, \tag{3.18}$$

where $M_X = \sup_{x \in X} \|x\|$.

PROOF. First, we claim that the following is true:

$$\|f(x) - f_\nu(x)\| = \frac{n}{2} (\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}. \tag{3.19}$$

To see that, note that since the components f_i of f have continuous Lipschitz gradient and using Theorem 3.1.1, we have

$$\begin{aligned}
\|f(x) - f_\nu(x)\| &= (\textstyle\sum_{i=1}^m (f_i(x) - f_{i,\nu_i}(x))^2)^{1/2} \\
&\leqslant \left(\sum_{i=1}^m \left(\frac{\nu_i^2 L_i n}{2} \right)^2 \right)^{1/2} \\
&= \left(\sum_{i=1}^m \frac{\nu_i^4}{4} L_i^2 n^2 \right)^{1/2} \\
&= \frac{n}{2} (\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}.
\end{aligned}$$

Utilizing this relation, using Theorem 3.1.1 and Cauchy-Schwartz inequality, we have

$$\begin{aligned}
|Q(z, \bar{z}) - Q_v(z, \bar{z})| &= |\mathcal{L}(x, \bar{y}) - \mathcal{L}(\bar{x}, y) - \mathcal{L}_\nu(x, \bar{y}) + \mathcal{L}_\nu(\bar{x}, y)| \\
&= |f_0(x) + \bar{y}^T f(x) - f_0(\bar{x}) - y^T f(\bar{x}) - f_{0,\nu_0}(x) - \bar{y}^T f_\nu(x) + f_{0,\nu_0}(\bar{x}) + y^T f_\nu(\bar{x})| \\
&\leqslant |f_0(x) - f_{0,\nu_0}(x)| + |f_0(\bar{x}) - f_{0,\nu_0}(\bar{x})| + |\bar{y}^T [f(x) - f_\nu(x)]| + |y^T [f(\bar{x}) - f_\nu(\bar{x})]| \\
&\leqslant |f_0(x) - f_{0,\nu_0}(x)| + |f_0(\bar{x}) - f_{0,\nu_0}(\bar{x})| + \|\bar{y}\| \|f(x) - f_\nu(x)\| + \|y\| \|f(\bar{x}) - f_\nu(\bar{x})\| \\
&\leqslant |f_0(x) - f_{0,\nu_0}(x)| + |f_0(\bar{x}) - f_{0,\nu_0}(\bar{x})| + M_X [\|f(x) - f_\nu(x)\| + \|f(\bar{x}) - f_\nu(\bar{x})\|] \\
&\leqslant \nu_0^2 L_0 n + M_X [n (\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}].
\end{aligned}$$

This conclues the proof. ∎

We also state the following results from [25], which is require in the proofs later.

Lemma 3.2.2 ([25]). Assume that $g : S \to \mathbb{R}$ satisfies

$$g(y) \geqslant g(x) + \langle g'(x), y - x \rangle + \mu W(y, x), \quad \forall x, y \in S \tag{3.20}$$

for some $\mu \geqslant 0$, where S is convex set in $\mathbb{R}^n$. If $\bar{x} = \arg\min_{x \in S} \{g(x) + W(x, \tilde{x})\}$, then $g(\bar{x}) + W(\bar{x}, \tilde{x}) + (\mu + 1) W(x, \bar{x}) \leqslant g(x) + W(x, \tilde{x})$, $\forall x \in S$.

Lemma 3.2.3 ([25])**.** Let $\rho_0, \ldots, \rho_j$ be a sequence of elements in $\mathbb{R}^n$ and let S be a convex set in $\mathbb{R}^n$. Define the sequence $v_t, t = 0, 1, \ldots,$ as follows: $v_0 \in S$ and

$$v_{t+1} = \arg\min_{x \in S} \langle \rho_t, x \rangle + \frac{1}{2}\|x - v_t\|_2^2.$$

Then for any $x \in S$ and $t \geqslant 0$, the following inequalities hold

$$\langle \rho_t, v_t - x \rangle \leqslant \frac{1}{2}\|x - v_t\|_2^2 - \frac{1}{2}\|x - v_{t+1}\|_2^2 + \frac{1}{2}\|\rho_t\|_2^2, \tag{3.21}$$

$$\textstyle\sum_{t=0}^{j} \langle \rho_t, v_t - x \rangle \leqslant \frac{1}{2}\|x - v_0\|_2^2 + \frac{1}{2}\displaystyle\sum_{t=0}^{j} \|\rho_t\|_2^2. \tag{3.22}$$

Lemma 3.2.4. Let $\{a_t\}_{t \geqslant 0}$ be a nonnegative sequence, $m_1, m_2 \geqslant 0$ be constants such that $a_0 \leqslant m_1$ and the following relation holds for all $t \geqslant 1$:

$$a_t \leqslant m_1 + m_2 \textstyle\sum_{k=0}^{t-1} a_k.$$

Then we have $a_t \leqslant m_1(1 + m_2)^t$.

Proof. We prove this lemma by induction. When $t = 0$, we have $a_0 \leq m_1$ by hypothesis. Assume for all $t \geq 0$, $a_t \leq m_1(1 + m_2)^t$. By induction hypothesis on a_k for all $k \in \{0, \ldots, t\}$ and hypothesis, we have

$$\begin{aligned}
a_{t+1} &\leqslant m_1 + m_2 \textstyle\sum_{k=0}^{t} a_k \\
&\leqslant m_1 + m_2 \textstyle\sum_{k=0}^{t} m_1(1 + m_2)^k \\
&\leqslant m_1 \left[1 + m_2 \textstyle\sum_{k=0}^{t}(1 + m_2)^k\right] \\
&\leqslant m_1 \left[1 + m_2 \frac{(1 + m_2)^{t+1} - 1}{m_2}\right] = m_1(1 + m_2)^{t+1}.
\end{aligned}$$

Hence, we conclude the proof. ∎

Lemma 3.2.5. Suppose Assumptions 4, 5 and 6 are satisfied. Let $B \geqslant 0$ be a constant and assume that $\{\gamma_t, \eta_t, \tau_t, \theta_t\}$ is a non-negative sequence satisfying

$$\gamma_t \theta_t = \gamma_{t-1}, \qquad \gamma_t \tau_t \leqslant \gamma_{t-1} \tau_{t-1}, \qquad \tau_t \eta_t \leqslant \gamma_{t-1} \eta_{t-1}, \tag{3.23}$$

and

$$(2M_f)^2\frac{\theta_t}{\theta_{t-1}} \leqslant \frac{\tau_t(\eta_{t-2}-L_0-BL_f)}{12}, \quad \theta_t(M_f)^2 \leqslant \frac{\tau_t(\eta_{t-1}-L_0-BL_f)}{12},$$

$$(2M_f)^2\frac{1}{\theta_{T-1}} \leqslant \frac{\tau_{T-1}(\eta_{T-2}-L_0-BL_f)}{12}, \quad M_f^2 \leqslant \frac{\tau_{T-1}(\eta_{T-1}-L_0-BL_f)}{12}, \tag{3.24}$$

where M_f, L_f are defined in (3.3). Then, for all $T \geqslant 1$ and $z \in \{(x,y) : x \in X, y \geqslant \mathbf{0}\}$, we have

$$\begin{aligned}
&\sum_{t=0}^{T-1}\gamma_t Q_\nu(z^{(t+1)}, z) + \sum_{t=0}^{T-1}\gamma_t[\langle \delta_t^G, x^{(t)} - x\rangle - \langle \delta_{t+1}^F, y^{(t+1)} - y\rangle] \\
&\leqslant \gamma_0\eta_0 W(x, x^{(0)}) - \gamma_{T-1}\eta_{T-1}W(x, x^{(T)}) + \frac{\gamma_0\tau_0}{2}\|y - y^{(0)}\|_2^2 - \frac{\gamma_{T-1}\tau_{T-1}}{12}\|y - y^{(T)}\|_2^2 \\
&+ \ \textstyle\sum_{t=0}^{T-1}\frac{2\gamma_t}{\eta_t - L_0 - BL_f}\left[\|\delta_t^G\|_*^2 + \left(\frac{L_f D_X}{2}[\|y\|_2 - B]_+\right)^2\right] \\
&+ \ \textstyle\sum_{t=1}^{T-1}\frac{3\gamma_t\theta_t^2}{2\tau_t}\|q_t - \bar{q}_t\|_2^2 + \frac{3\gamma_{T-1}}{2\tau_{T-1}}\|q_T - \bar{q}_T\|_2^2.
\end{aligned} \tag{3.25}$$

Here $q_t := \ell_F(x^{(t)}) - \ell_F(x^{(t-1)}), \bar{q}_t := \ell_f(x^{(t)}) - \ell_f(x^{(t-1)})$, $\delta_t^F := \ell_F(x^{(t)}) - \ell_f(x^{(t)})$ and $\delta_t^G := G_{0,\nu_0}(x^{(t)}, \xi_i^{(t)}, u_i^{(t)}) + \sum_{i\in[m]} G_{i,\nu_i}(x^{(t)}, \xi_i^{(t)}, u_i^{(t)})y_i^{(t+1)} - f'_{0,\nu_0}(x^{(t)}) - \sum_{i=1}^m f'_{i,\nu_i}(x^{(t)})y_i^{(t+1)}$.

Proof. Note that $y^{(t+1)} = \arg\min_{y\geqslant\mathbf{0}}\langle -s^{(t)}, y\rangle + \frac{\tau_t}{2}\|y - y^{(t)}\|_2^2$. Hence, using Lemma 3.2.2 with $y \mapsto \langle -s^{(t)}, y\rangle$ and $\mu = 0$, we have for all $y \geqslant \mathbf{0}$,

$$-\langle s^{(t)}, y^{(t+1)} - y\rangle \leqslant \frac{\tau_t}{2}[\|y - y^{(t)}\|_2^2 - \|y^{(t+1)} - y^{(t)}\|_2^2 - \|y - y^{(t+1)}\|_2^2]. \tag{3.26}$$

Let us denote $v_t := f'_{0,\nu_0}(x^{(t)}) + \sum_{i\in[m]} f'_{i,\nu_i}(x^{(t)})y_i^{(t+1)}$
and $V_t := G_{0,\nu_0}(x^{(t)}, \xi_i^{(t)}, u_i^{(t)}) + \sum_{i\in[m]} G_{i,\nu_i}(x^{(t)}, \xi_i^{(t)}, u_i^{(t)})y_i^{(t+1)}$.

Then using Lemma 3.2.2 with $x \mapsto \langle V_t, x\rangle$ and the optimality of $x^{(t+1)}$, we have for all $x \in X$,

$$\langle V_t, x^{(t+1)} - x\rangle \leqslant \eta_t[W(x, x^{(t)}) - W(x^{(t+1)}, x^{(t)})] - \eta_t W(x, x^{(t+1)}). \tag{3.27}$$

Due to the convexity of f_{0,ν_0} and f_{i,ν_i}, and since f_0, f_i are Lipschitz, and by the definition of ℓ_f, and the fact that $y^{(t+1)} \geqslant \mathbf{0}$, we have

$$
\begin{aligned}
&\langle v_t, x^{(t+1)} - x\rangle = \langle f'_{0,\nu_0}(x^{(t)}) + \textstyle\sum_{i\in[m]} f'_{i,\nu_i}(x^{(t)})y_i^{(t+1)}, x^{(t+1)} - x\rangle \\
&= \langle f'_{0,\nu_0}(x^{(t)}), x^{(t+1)} - x^{(t)} + x^{(t)} - x\rangle + \langle f'_{\nu}(x^{(t)})y^{(t+1)}, x^{(t+1)} - x^{(t)} + x^{(t)} - x\rangle \\
&\geqslant f_{0,\nu_0}(x^{(t)}) - f_{0,\nu_0}(x) + f_{0,\nu_0}(x^{(t+1)}) - f_{0,\nu_0}(x^{(t)}) - \frac{L_0}{2}\|x^{(t+1)} - x^{(t)}\|^2 \\
&+ \langle y^{(t+1)}, \ell_f(x^{(t+1)}) - f_\nu(x^{(t)})\rangle + \langle y^{(t+1)}, f_\nu(x^{(t)}) - f_\nu(x)\rangle \\
&= f_{0,\nu_0}(x^{(t+1)}) - f_{0,\nu_0}(x) + \langle \ell_f(x^{(t+1)}) - f_\nu(x), y^{(t+1)}\rangle - \underbrace{\frac{L_0}{2}\|x^{(t+1)} - x^{(t)}\|^2}_{O_{t+1}}. \qquad (3.28)
\end{aligned}
$$

Combining (3.27), (3.28), noting that $\delta_t^G = V_t - v_t$, we have

$$
\begin{aligned}
&f_{0,\nu_0}(x^{(t+1)}) - f_{0,\nu_0}(x) + \langle \ell_f(x^{(t+1)}) - f_\nu(x), y^{(t+1)}\rangle + \langle \delta_t^G, x^{(t+1)} - x\rangle \\
&\leqslant \eta_t W(x, x^{(t)}) - \eta_t W(x^{(t+1)}, x^{(t)}) - \eta_t W(x, x^{(t+1)}) + O_{t+1}. \qquad (3.29)
\end{aligned}
$$

Noting the definition of $Q_\nu(\cdot,\cdot)$ (see (3.17)) and, adding (3.26) and (3.29), we obtain

$$
\begin{aligned}
&Q_\nu(z^{(t+1)}, z) - \langle f_\nu(x^{(t+1)}), y\rangle + \langle \ell_f(x^{(t+1)}), y^{(t+1)}\rangle - \langle s^{(t)}, y^{(t+1)} - y\rangle + \langle \delta_t^G, x^{(t+1)} - x\rangle \\
&\leqslant \frac{\tau_t}{2}[\|y - y^{(t)}\|_2^2 - \|y^{(t+1)} - y^{(t)}\|_2^2 - \|y - y^{(t+1)}\|_2^2] \\
&+ \eta_t W(x, x^{(t)}) - \eta_t W(x^{(t+1)}, x^{(t)}) - \eta_t W(x, x^{(t+1)}) + O_{t+1}. \qquad (3.30)
\end{aligned}
$$

Note that we also have $f_{i,\nu_i}(x^{(t+1)}) - \ell_{f_i}(x^{(t+1)}) \leqslant \frac{L_i}{2}\|x^{(t+1)} - x^{(t)}\|^2$. Then, using Cauchy-Schwarz inequality and noting definitions of L_f, we have

$$
\langle y, f_\nu(x^{(t+1)}) - \ell_f(x^{(t+1)})\rangle \leqslant \|y\|_2 \underbrace{\frac{L_f}{2}\|x^{(t+1)} - x^{(t)}\|^2}_{C_{t+1}}.
$$

Noting the above relation and definitions of q_t and δ_{t+1}^F, we have

$$\begin{aligned}
&\langle \ell_f(x^{(t+1)}), y^{(t+1)}\rangle - \langle f_\nu(x^{(t+1)}), y\rangle - \langle s^{(t)}, y^{(t+1)} - y\rangle \\
\geqslant&\langle \ell_f(x^{(t+1)}), y^{(t+1)}\rangle - \langle \ell_f(x^{(t+1)}), y\rangle - \langle s^{(t)}, y^{(t+1)} - y\rangle - \|y\|_2 C_{t+1} \\
=&\langle \ell_f(x^{(t+1)}) - s^{(t)}, y^{(t+1)} - y\rangle - \|y\|_2 C_{t+1} \\
=&\langle \ell_f(x^{(t+1)}) - \ell_F(x^{(t)}) - \theta_t q_t, y^{(t+1)} - y\rangle - \|y\|_2 C_{t+1} \\
=&\langle q_{t+1}, y^{(t+1)} - y\rangle - \theta_t\langle q_t, y^{(t)} - y\rangle - \theta_t\langle q_t, y^{(t+1)} - y^{(t)}\rangle - \langle \delta_{t+1}^F, y^{(t+1)} - y\rangle - \|y\|_2 C_{t+1}. \quad (3.31)
\end{aligned}$$

Let $B \geqslant 0$ be a constant. Then

$$\begin{aligned}
\|y\|_2 C_{t+1} &= \frac{L_f}{2}(\|y\|_2 - B)\|x^{(t+1)} - x^{(t)}\|^2 + \frac{BL_f}{2}\|x^{(t+1)} - x^{(t)}\|^2 \\
&\leqslant \frac{L_f}{2}[\|y\|_2 - B]_+\|x^{(t+1)} - x^{(t)}\|^2 + \frac{BL_f}{2}\|x^{(t+1)} - x^{(t)}\|^2 \\
&\leqslant \frac{BL_f}{2}\|x^{(t+1)} - x^{(t)}\|^2 + \frac{L_f D_X}{2}[\|y\|_2 - B]_+\|x^{(t+1)} - x^{(t)}\|. \quad (3.32)
\end{aligned}$$

By (3.30), (3.31), and (3.32), noting the definition of O_{t+1} and using the relation $\frac{1}{2}\|a-b\|^2 \leqslant W(a,b)$, we have

$$\begin{aligned}
&Q_\nu(z^{(t+1)}, z) + \langle q_{t+1}, y^{(t+1)} - y\rangle - \theta_t\langle q_t, y^{(t)} - y\rangle + \langle \delta_t^G, x^{(t)} - x\rangle - \langle \delta_{t+1}^F, y^{(t+1)} - y\rangle \\
\leqslant&\theta_t\langle q_t, y^{(t+1)} - y^{(t)}\rangle - \langle \delta_t^G, x^{(t+1)} - x^{(t)}\rangle \\
&+\eta_t W(x, x^{(t)}) - \eta_t W(x, x^{(t+1)}) + \frac{\tau_t}{2}[\|y - y^{(t)}\|_2^2 - \|y^{(t+1)} - y^{(t)}\|_2^2 - \|y - y^{(t+1)}\|_2^2] \\
&-(\eta_t - L_0 - BL_f)W(x^{(t+1)}, x^{(t)}) + \frac{L_f D_X}{2}[\|y\|_2 - B]_+\|x^{(t+1)} - x^{(t)}\|. \quad (3.33)
\end{aligned}$$

Multiplying (3.33) by γ_t, summing them up from $t = 0$ to $T - 1$ with $T \geqslant 1$, we obtain

$$
\begin{aligned}
&\textstyle\sum_{t=0}^{T-1} \gamma_t Q_\nu(z^{(t+1)}, z) + \sum_{t=0}^{T-1} [\gamma_t \langle q_{t+1}, y^{(t+1)} - y\rangle - \gamma_t\theta_t\langle q_t, y^{(t)} - y\rangle] \\
&\quad \textstyle+ \sum_{t=0}^{T-1} \gamma_t [\langle \delta_t^G, x^{(t)} - x\rangle - \langle \delta_{t+1}^F, y^{(t+1)} - y\rangle] \\
\leqslant\ & \textstyle\sum_{t=0}^{T-1} [\gamma_t\theta_t\langle q_t - \bar{q}_t, y^{(t+1)} - y^{(t)}\rangle + \gamma_t\theta_t\langle \bar{q}_t, y^{(t+1)} - y^{(t)}\rangle + \langle \gamma_t\delta_t^G, x^{(t)} - x^{(t+1)}\rangle] \\
+\ & \textstyle\sum_{t=0}^{T-1} \left[\frac{\gamma_t\tau_t}{2}\|y - y^{(t)}\|_2^2 - \frac{\gamma_t\tau_t}{2}\|y - y^{(t+1)}\|_2^2\right] - \sum_{t=0}^{T-1} \frac{\gamma_t\tau_t}{2}\|y^{(t+1)} - y^{(t)}\|_2^2 \\
+\ & \textstyle\sum_{t=0}^{T-1} [\gamma_t\eta_t W(x, x^{(t)}) - \gamma_t\eta_t W(x, x^{(t+1)})] \\
-\ & \sum_{t=0}^{T-1} \left[\gamma_t(\eta_t - L_0 - BL_f)W(x^{(t+1)}, x^{(t)}) - \gamma_t \underbrace{\left(\frac{L_f D_X}{2}[\|y\|_2 - B]_+\right)}_{\mathcal{H}(y,B)} \|x^{(t+1)} - x^{(t)}\|\right],
\end{aligned} \tag{3.34}
$$

where $\mathcal{H}(y, B) := \frac{L_f D_X}{2}[\|y\|_2 - B]_+$. Now we focus our attention to handle the inner product terms of (3.34). Noting the definition of $\bar{q}_t$, we have

$$
\begin{aligned}
\|\bar{q}_t\|_2 &= \|\ell_f(x^{(t)}) - \ell_f(x^{(t-1)})\|_2 \\
&= \|f_\nu(x^{(t-1)}) + f'_\nu(x^{(t-1)})^T(x^{(t)} - x^{(t-1)}) - f_\nu(x^{(t-2)}) - f'_\nu(x^{(t-2)})^T(x^{(t-1)} - x^{(t-2)})\|_2 \\
&\leqslant \|f_\nu(x^{(t-1)}) - f_\nu(x^{(t-2)})\|_2 + \|f'_\nu(x^{(t-1)})^T(x^{(t)} - x^{(t-1)})\|_2 + \|f'_\nu(x^{(t-2)})^T(x^{(t-1)} - x^{(t-2)})\|_2 \\
&\leqslant 2M_f\|x^{(t-1)} - x^{(t-2)}\| + M_f\|x^{(t)} - x^{(t-1)}\|,
\end{aligned} \tag{3.35}
$$

where we used the fact that $\|f_\nu(x) - f_\nu(y)\| \leqslant M_f\|x - y\|$ and $\|[f'_\nu(x)]^T(y - x)\|_2 \leqslant M_f\|y - x\|$, which follows from an analogue for (3.2) and Theorem 3.1.1. Using the above relation for $\|\bar{q}_t\|_2$, we now obtain

$$
\begin{aligned}
&\gamma_t\theta_t\langle \bar{q}_t, y^{(t+1)} - y^{(t)}\rangle - \frac{\gamma_t\tau_t}{3}\|y^{(t+1)} - y^{(t)}\|_2^2 \\
&- \frac{\gamma_{t-2}(\eta_{t-2} - L_0 - BL_f)}{4}W(x^{(t-1)}, x^{(t-2)}) - \frac{\gamma_{t-1}(\eta_{t-1} - L_0 - BL_f)}{4}W(x^{(t)}, x^{(t-1)}) \\
\leqslant& \gamma_t\theta_t\|\bar{q}_t\|_2\|y^{(t+1)} - y^{(t)}\|_2 - \frac{\gamma_t\tau_t}{3}\|y^{(t+1)} - y^{(t)}\|_2^2 \\
&- \frac{\gamma_{t-2}(\eta_{t-2} - L_0 - BL_f)}{4}W(x^{(t-1)}, x^{(t-2)}) - \frac{\gamma_{t-1}(\eta_{t-1} - L_0 - BL_f)}{4}W(x^{(t)}, x^{(t-1)}) \\
\leqslant& 2M_f\gamma_t\theta_t\|x^{(t-1)} - x^{(t-2)}\|\|y^{(t+1)} - y^{(t)}\|_2 - \frac{\gamma_t\tau_t}{6}\|y^{(t+1)} - y^{(t)}\|_2^2 \\
&- \frac{\gamma_{t-2}(\eta_{t-2} - L_0 - BL_f)}{4}W(x^{(t-1)}, x^{(t-2)}) + M_f\gamma_t\theta_t\|x^{(t)} - x^{(t-1)}\|\|y^{(t+1)} - y^{(t)}\|_2 \\
&- \frac{\gamma_t\tau_t}{6}\|y^{(t+1)} - y^{(t)}\|_2^2 - \frac{\gamma_{t-1}(\eta_{t-1} - L_0 - BL_f)}{4}W(x^{(t)}, x^{(t-1)}) \\
\leqslant& 0,
\end{aligned} \tag{3.36}
$$

where the last inequality follows by applying the relation $W(x,y) \geqslant \frac{1}{2}\|x-y\|$, Young's inequality $(2ab \leqslant a^2 + b^2)$ applied twice, once with

$$
a = \left(\frac{\gamma_t\tau_t}{6}\right)^{1/2}\|y^{(t+1)} - y^{(t)}\|, \quad b = \left(\frac{\gamma_{t-2}(\eta_{t-2} - L_0 - BL_f)}{8}\right)^{1/2}\|x^{(t-1)} - x^{(t-2)}\|,
$$

and second time with

$$
a = \left(\frac{\gamma_t\tau_t}{6}\right)^{1/2}\|y^{(t+1)} - y^{(t)}\|, \quad b = \left(\frac{\gamma_{t-1}(\eta_{t-1} - L_0 - BL_f)}{8}\right)^{1/2}\|x^{(t)} - x^{(t-1)}\|,
$$

and the fact that

$$
\begin{aligned}
2M_f\gamma_t\theta_t &\leqslant \left\{\frac{\gamma_t\gamma_{t-2}\tau_t(\eta_{t-2} - L_0 - BL_f)}{12}\right\}^{1/2} \quad \Leftrightarrow \quad (2M_f)^2\frac{\theta_t}{\theta_{t-1}} \leqslant \frac{\tau_t(\eta_{t-2} - L_0 - BL_f)}{12}, \\
M_f^2\gamma_t^2\theta_t^2 &\leqslant \frac{\gamma_t\gamma_{t-1}\tau_t(\eta_{t-1} - L_0 - BL_f)}{12} \quad \Leftrightarrow \quad M_f^2\theta_t \leqslant \frac{\tau_t(\eta_{t-1} - L_0 - BL_f)}{12},
\end{aligned}
$$

where the equivalences follow due to (3.23). Using Young's inequality, Cauchy-Schwarz inequality and the relation $u^T v \leqslant \|u\|\|v\|_*$, we have

$$\gamma_t\theta_t\langle q_t - \bar{q}_t, y^{(t+1)} - y^{(t)}\rangle - \frac{\gamma_t\tau_t}{6}\|y^{(t+1)} - y^{(t)}\|_2^2 \leqslant \frac{3\gamma_t\theta_t^2}{2\tau_t}\|q_t - \bar{q}_t\|_2^2,$$

$$\langle\gamma_t\delta_t^G, x^{(t)} - x^{(t+1)}\rangle - \frac{\gamma_t(\eta_t - L_0 - BL_f)}{4}W(x^{(t+1)}, x^{(t)}) \leqslant \frac{2\gamma_t}{\eta_t - L_0 - BL_f}\|\delta_t^G\|_*^2, \tag{3.37}$$

(3.38)

Using (3.36) and (3.37) for $t = 0, \ldots, T-1$ inside (3.34) and noting (3.23), we have

$$\begin{aligned}
&\sum_{t=0}^{T-1}\gamma_t Q_\nu(z^{(t+1)}, z) + \gamma_{T-1}\langle q_T, y^{(T)} - y\rangle + \sum_{t=0}^{T-1}\gamma_t[\langle\delta_t^G, x^{(t)} - x\rangle - \langle\delta_{t+1}^F, y^{(t+1)} - y\rangle] \\
\leqslant&\gamma_0\eta_0 W(x, x^{(0)}) - \gamma_{T-1}\eta_{T-1}W(x, x^{(T)}) + \frac{\gamma_0\tau_0}{2}\|y - y^{(0)}\|_2^2 - \frac{\gamma_{T-1}\tau_{T-1}}{2}\|y - y^{(T)}\|_2^2 \\
&+ \sum_{t=0}^{T-1}\left[\frac{3\gamma_t\theta_t^2}{2\tau_t}\|q_t - \bar{q}_t\|_2^2 + \frac{2\gamma_t}{\eta_t - L_0 - BL_f}\|\delta_t^G\|_*^2 + \frac{2\gamma_t}{\eta_t - L_0 - BL_f}\mathcal{H}(y, B)^2\right] \\
&- \frac{\gamma_{T-2}(\eta_{T-2} - L_0 - BL_f)}{4}W(x^{(T-1)}, x^{(T-2)}) - \frac{\gamma_{T-1}(\eta_{T-1} - L_0 - BL_f)}{2}W(x^{(T)}, x^{(T-1)}),
\end{aligned} \tag{3.39}$$

where in the left hand side of the above relation, we used the fact that $q_0 = \ell_F(x^{(0)}) - \ell_F(x^{(-1)}) = \mathbf{0}$. Similarly, we see that $\bar{q}_0 = \mathbf{0}$. Hence, we can ignore $\|q_0 - \bar{q}_0\|_2^2$ term in the right hand side of the above relation, after which we obtain

$$\begin{aligned}
&- \gamma_{T-1}\langle\bar{q}_T, y^{(T)} - y\rangle - \frac{\gamma_{T-1}\tau_{T-1}}{3}\|y - y^{(T)}\|_2^2 \\
&- \frac{\gamma_{T-2}(\eta_{T-2} - L_0 - BL_f)}{4}W(x^{(T-1)}, x^{(T-2)}) - \frac{\gamma_{T-1}(\eta_{T-1} - L_0 - BL_f)}{2}W(x^{(T)}, x^{(T-1)}) \\
\leqslant\ & M_f\gamma_{T-1}\|x^{(T)} - x^{(T-1)}\|\|y^{(T)} - y\|_2 - \frac{\gamma_{T-1}\tau_{T-1}}{12}\|y - y^{(T)}\|_2^2 \\
&- \frac{\gamma_{T-1}(\eta_{T-1} - L_0 - BL_f)}{2}W(x^{(T)}, x^{(T-1)}) + 2M_f\gamma_{T-1}\|x^{(T-1)} - x^{(T-2)}\|\|y^{(T)} - y\|_2 \\
&- \frac{\gamma_{T-1}\tau_{T-1}}{6}\|y - y^{(T)}\|_2^2 - \frac{\gamma_{T-2}(\eta_{T-2} - L_0 - BL_f)}{4}W(x^{(T-1)}, x^{(T-2)}) - \frac{\gamma_{T-1}\tau_{T-1}}{12}\|y^{(T)} - y\|_2^2 \\
\leqslant\ & -\frac{\gamma_{T-1}\tau_{T-1}}{12}\|y^{(T)} - y\|_2^2,
\end{aligned} \tag{3.40}$$

where the last relation follows from (3.24), Young's inequality and the fact that

$$2M_f\gamma_{T-1} \leqslant \left\{\frac{\gamma_{T-2}\gamma_{T-1}\tau_{T-1}(\eta_{T-2}-L_0-BL_f)}{12}\right\}^{1/2} \Leftrightarrow \quad \frac{(2M_f)^2}{\theta_{T-1}} \leqslant \frac{\tau_{T-1}(\eta_{T-2}-L_0-BL_f)}{12}$$

$$M_f\gamma_{T-1} \leqslant \left\{\frac{\gamma_{T-1}^2\tau_{T-1}(\eta_{T-1}-L_0-BL_f)}{12}\right\}^{1/2} \Leftrightarrow \quad M_f^2 \leqslant \frac{\tau_{T-1}(\eta_{T-1}-L_0-BL_f)}{12}.$$

Moreover, again using Young's inequality and Cauchy-Schwarz inequality, we have

$$-\gamma_{T-1}\langle q_T-\bar{q}_T, y^{(T)}-y\rangle - \frac{\gamma_{T-1}\tau_{T-1}}{6}\|y-y^{(T)}\|_2^2 \leqslant \frac{3\gamma_{T-1}}{2\tau_{T-1}}\|q_T-\bar{q}_T\|_2^2. \tag{3.41}$$

Using (3.40) and (3.41) in relation (3.39), noting that $q_0-\bar{q}_0=\mathbf{0}$ and replacing the definition of $\mathcal{H}(y,B)$, we obtain (3.25), which completes the proof. ∎

Lemma 3.2.6. Suppose all conditions required for Lemma 3.2.5 hold. Then, for all $T\geqslant 1$, we have

$$\begin{aligned}
\mathbb{E}[f_0(\bar{x}_T)-f_0(x^*)] \leqslant \frac{1}{\Gamma_T}\Big[&\gamma_0\eta_0 W(x^*,x^{(0)}) + \frac{\gamma_0\eta_0}{2}\|y^{(0)}\|_2^2 + \textstyle\sum_{t=0}^{T-1}\frac{2\gamma_t}{\eta_t-L_0-BL_f}\mathbb{E}[\|\delta_t^G\|_*^2] \\
&+ \left(\textstyle\sum_{t=1}^{T-1}\frac{12\gamma_t\theta_t^2}{\tau_t}+\frac{12\gamma_{T-1}}{\tau_{T-1}}\right)(\sigma_{f,\nu}^2+D_X^2\|\sigma_\nu\|_2^2)\Big] \\
&+ [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}],
\end{aligned} \tag{3.42}$$

$$\begin{aligned}
\mathbb{E}[\|[f(\bar{x}_T)]_+\|_2] \leqslant \frac{1}{\Gamma_T}\Big[&\gamma_0\tau_0\|y^{(0)}\|_2^2 + 3(\|y^*\|_2+1)^2\gamma_0\tau_0 + \gamma_0\eta_0 W(x^*,x^{(0)}) \\
&+ \textstyle\sum_{t=0}^{T-1}\frac{2\gamma_t}{\eta_t-L_0-BL_f}\left\{\mathbb{E}[\|\delta_t^G\|_*^2] + \left(\frac{L_f D_X}{2}[\|y^*\|_2+1-B]_+\right)^2\right\} && (3.43) \\
&+ \left(\textstyle\sum_{t=1}^{T-1}\frac{12\gamma_t\theta_t^2}{\tau_t}+\sum_{t=0}^{T-1}\frac{\gamma_t}{\tau_t}+\frac{12\gamma_{T-1}}{\tau_{T-1}}\right)(\sigma_{f,\nu}^2+D_X^2\|\sigma_\nu\|_2^2)\Big] && (3.44) \\
&+ [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}],
\end{aligned}$$

where $\Gamma_T := \sum_{t=0}^{T-1}\gamma_t$ and $\sigma_\nu = (\sigma_{1,\nu_1},\ldots,\sigma_{m,\nu_m})$ with σ_{i,ν_i} as defined in (3.12).

Proof. First, observe that $y^{(t+1)}$ is a constant conditioned on random variable $\xi_{[t-1]}, u_{[t-1]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}$. In particular,

$$\mathbb{E}[\langle\delta_t^G, x^{(t)}-x\rangle] = \mathbb{E}\langle\mathbb{E}_{|\xi_{[t-1]},u_{[t-1]},\bar{\xi}_{[t-1]},\bar{u}_{[t-1]}}[\delta_t^G], x^{(t)}-x\rangle = 0 \tag{3.45}$$

for any non-random x. This follows due to the following relation

$$
\begin{aligned}
&\mathbb{E}_{|\xi_{[t-1]},u_{[t-1]},\bar{\xi}_{[t-1]},\bar{u}_{[t-1]}}[\delta_t^G] \\
&= \mathbb{E}_{|\xi_{[t-1]},u_{[t-1]},\bar{\xi}_{[t-1]},\bar{u}_{[t-1]}}[G_{0,\nu_0}(x^{(t)},\xi_0^{(t)},u_0^{(t)}) - f'_{0,\nu_0}(x^{(t)})] \\
&\quad + \textstyle\sum_{i=1}^m y_i^{(t+1)} \mathbb{E}_{|\xi_{[t-1]},u_{[t-1]},\bar{\xi}_{[t-1]},\bar{u}_{[t-1]}}[G_{i,\nu_i}(x^{(t)},\xi_i^{(t)},u_i^{(t)}) - f'_{i,\nu_i}(x^{(t)})] \\
&= \mathbf{0}.
\end{aligned}
$$

Similarly, we have

$$
\mathbb{E}[\langle \delta_{t+1}^F, y^{(t+1)} - y\rangle] = \mathbb{E}[\langle \mathbb{E}_{|\xi_{[t]},u_{[t]},\bar{\xi}_{[t-1]},\bar{u}_{[t-1]}}[\delta_{t+1}^F], y^{(t+1)} - y\rangle] = 0, \tag{3.46}
$$

for any non-random y. Here, we note that

$$
\begin{aligned}
\mathbb{E}_{|\xi_{[t]},u_{[t]},\bar{\xi}_{[t-1]},\bar{u}_{[t-1]}}[\delta_{t+1}^F] &= \mathbb{E}_{|\xi_{[t]},u_{[t]},\bar{\xi}_{[t-1]},\bar{u}_{[t-1]}}[F_\nu(x^{(t)},\bar{\xi}^{(t)},\bar{u}^{(t)})] - f_\nu(x^{(t)}) \\
&\quad + (\mathbb{E}_{|\xi_{[t]},u_{[t]},\bar{\xi}_{[t-1]},\bar{u}_{[t-1]}}[\mathbf{G}_\nu(x^{(t)},\bar{\xi}^{(t)},\bar{u}^{(t)})] - f'_\nu(x^{(t)}))^T(x^{(t+1)} - x^{(t)}) = \mathbf{0},
\end{aligned} \tag{3.47}
$$

where the first term in RHS is $\mathbf{0}$ due to $\mathbb{E}_{\xi,u}F_\nu(x,\xi,u) = f_\nu(x)$, the second term is $\mathbf{0}$ due to the $\mathbb{E}_{\xi,u}\mathbf{G}_\nu(x,\xi,u) = f'_\nu(x)$ and the common fact for both the terms that $x^{(t)}, x^{(t+1)}$ are constants for given $\xi_{[t]}, u_{[t]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}$. We now note that

$$
\begin{aligned}
\mathbb{E}[\|\delta_t^F\|_2^2] &\leqslant 2\mathbb{E}[\|F_\nu(x^{(t-1)},\bar{\xi}^{(t-1)},\bar{u}^{(t-1)}) - f_\nu(x^{(t-1)})\|_2^2] \\
&\quad + 2\mathbb{E}[\|[\mathbf{G}_\nu(x^{(t-1)},\bar{\xi}^{(t-1)},\bar{u}^{(t-1)}) - f'_\nu(x^{(t-1)})]^T(x^{(t)} - x^{(t-1)})\|_2^2] \\
&\leqslant 2\sigma_{f,\nu}^2 + 2\mathbb{E}\left[\sum_{i=1}^m \left\{(G_{i,\nu_i}(x^{(t-1)},\bar{\xi}_i^{(t-1)},\bar{u}_i^{(t-1)}) - f'_{i,\nu_i}(x^{(t-1)}))^T(x^{(t)} - x^{(t-1)})\right\}^2\right] \\
&\leqslant 2\sigma_{f,\nu}^2 + 2\mathbb{E}\left[\sum_{i=1}^m \|G_{i,\nu_i}(x^{(t-1)},\bar{\xi}_i^{(t-1)},\bar{u}_i^{(t-1)}) - f'_{i,\nu_i}(x^{(t-1)})\|_*^2 \|x^{(t)} - x^{(t-1)}\|^2\right] \\
&\leqslant 2\sigma_{f,\nu}^2 + 2D_X^2\|\sigma_\nu\|_2^2.
\end{aligned} \tag{3.48}
$$

Then, in view of above relation and definitions of $q_t, \bar{q}_t$, we have

$$\begin{aligned}\mathbb{E}[\|q_t - \bar{q}_t\|_2^2] &= \mathbb{E}[\|\ell_F(x^{(t)}) - \ell_f(x^{(t)}) - \ell_F(x^{(t-1)}) + \ell_f(x^{(t-1)})\|_2^2] \\ &\leqslant 2\mathbb{E}[\|\delta_t^F\|_2^2] + 2\mathbb{E}[\|\delta_{t-1}^F\|_2^2] \leqslant 8(\sigma_{f,\nu}^2 + D_X^2\|\sigma_\nu\|_2^2). \end{aligned} \tag{3.49}$$

Taking the expectation on both sides of (3.25) and using relation (3.45), (3.46) and (3.49), we have for all non-random $z \in \{(x, y) : x \in X, y \geqslant \mathbf{0}\}$,

$$\begin{aligned}&\mathbb{E}\left[\textstyle\sum_{t=0}^{T-1} \gamma_t Q_\nu(z^{(t+1)}, z)\right] \\ \leqslant&\gamma_0\eta_0 W(x, x^{(0)}) - \gamma_{T-1}\eta_{T-1}\mathbb{E}[W(x, x^{(T)})] + \frac{\gamma_0\tau_0}{2}\|y - y^{(0)}\|_2^2 \\ &+ \textstyle\sum_{t=0}^{T-1} \frac{2\gamma_t}{\eta_t - L_0 - BL_f}\left[\mathbb{E}[\|\delta_t^G\|_*^2] + \left(\frac{L_f D_X}{2}[\|y\|_2 - B]_+\right)^2\right] \\ &+ \left(\textstyle\sum_{t=1}^{T-1} \frac{12\gamma_t\theta_t^2}{\tau_t} + \frac{12\gamma_{T-1}}{\tau_{T-1}}\right)(\sigma_{f,\nu}^2 + D_X^2\|\sigma_\nu\|_2^2)\end{aligned} \tag{3.50}$$

where we dropped $\|y - y^{(T)}\|_2^2$. By Lemma 3.2.1, we have

$$Q(z^{(t+1)}, z) - [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}] \leqslant Q_\nu(z^{(t+1)}, z).$$

Using this relation, multiplying both sides by γ_t, summing from $t = 0, \ldots, T-1$, and taking expectation on both sides, we have

$$\mathbb{E}\left[\textstyle\sum_{t=0}^{T-1} \gamma_t Q(z^{(t+1)}, z)\right] - [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\Gamma_T \leqslant \mathbb{E}\left[\textstyle\sum_{t=0}^{T-1} \gamma_t Q_\nu(z^{(t+1)}, z)\right]. \tag{3.51}$$

Using this relation, the convexity of $f_0(\cdot)$ and $f(\cdot)$, and noting the definition of Γ_T, we have for all non-random $y \geqslant \mathbf{0}$ and $x \in X$,

$$\begin{aligned}&\Gamma_T\mathbb{E}[f_0(\bar{x}_T) + \langle y, f(\bar{x}_T)\rangle - f_0(x) - \langle \bar{y}_T, f(x)\rangle] - [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\Gamma_T \\ \leqslant& \mathbb{E}\left[\textstyle\sum_{t=0}^{T-1} \gamma_t Q(z^{(t+1)}, z)\right] - [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\Gamma_T \\ \leqslant& \mathbb{E}\left[\textstyle\sum_{t=0}^{T-1} \gamma_t Q_\nu(z^{(t+1)}, z)\right].\end{aligned} \tag{3.52}$$

Combining (3.50), (3.51) and (3.52), then choosing $x = x^*$, $y = \mathbf{0}$ (which are non-random) throughout the combined relation, observing that $[0 - B]_+ = 0$ for any $B \geqslant 0$, we have

$$
\begin{aligned}
&\Gamma_T \mathbb{E}[f_0(\bar{x}_T) - f_0(x^*) - \langle \bar{y}_T, f(x^*)\rangle] - [\nu_0^2 L_0 n + M_X n (\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\Gamma_T \\
\leqslant\ & \mathbb{E}\left[\textstyle\sum_{t=0}^{T-1} \gamma_t Q_\nu(z^{(t+1)}, (x^*, \mathbf{0}))\right] \\
\leqslant\ & \gamma_0 \eta_0 W(x^*, x^{(0)}) - \gamma_{T-1}\eta_{T-1}\mathbb{E}[W(x^*, x^{(T)})] + \frac{\gamma_0 \tau_0}{2}\|y^{(0)}\|_2^2 + \textstyle\sum_{t=0}^{T-1} \frac{2\gamma_t}{\eta_t - L_0 - BL_f}\mathbb{E}[\|\delta_t^G\|_*^2] \\
+\ & \left(\textstyle\sum_{t=1}^{T-1} \frac{12\gamma_t \theta_t^2}{\tau_t} + \frac{12\gamma_{T-1}}{\tau_{T-1}}\right)(\sigma_{f,\nu}^2 + D_X^2 \|\sigma_\nu\|_2^2).
\end{aligned} \tag{3.53}
$$

Ignoring the $\mathbb{E}[W(x^*, x^{(T)})]$ term and noting that $f(x^*) \leqslant \mathbf{0}$ and $\bar{y}_T \geqslant \mathbf{0}$ implies $\langle \bar{y}_T, f(x^*)\rangle \leqslant 0$, we have (3.42).

Now, we focus our attention to the infeasibility bound. First, we define $R := \|y^*\|_2 + 1$. Second, define an auxilliary sequence $\{y_t^v\}$ in the following way: $y_0^v = y^{(0)}$ and for all $t \geqslant 0$, define

$$
y_{t+1}^v := \operatorname*{arg\,min}_{y \in \mathcal{B}_+^2(R)} \frac{1}{\tau_{t-1}}\langle \delta_t^F, y\rangle + \frac{1}{2}\|y - y_t^v\|_2^2,
$$

where we recall that $\mathcal{B}_+^2(R) = \{x \in \mathbb{R}^n : \|x\|_2 \leqslant R, x \geqslant \mathbf{0}\}$. Then in view of Lemma 3.2.3, in particular relation (3.21), for all $y \in \mathcal{B}_+^2(R)$ we have

$$
\frac{1}{\tau_t}\langle \delta_{t+1}^F, y_{t+1}^v - y\rangle \leqslant \frac{1}{2}\|y - y_{t+1}^v\|_2^2 - \frac{1}{2}\|y - y_{t+2}^v\|_2^2 + \frac{1}{2\tau_t^2}\|\delta_{t+1}^F\|_2^2. \tag{3.54}
$$

Multiplying (3.54) by $\gamma_t \tau_t$, taking a sum from $t = 0$ to $T - 1$ and noting the second relation in (3.23), we obtain

$$
\textstyle\sum_{t=0}^{T-1} \gamma_t \langle \delta_{t+1}^F, y_{t+1}^v - y\rangle \leqslant \frac{\gamma_0 \tau_0}{2}\|y - y_1^v\|_2^2 + \sum_{t=0}^{T-1} \frac{\gamma_t}{2\tau_t}\|\delta_{t+1}^F\|_2^2, \tag{3.55}
$$

for all $y \in \mathcal{B}_+^2(R)$. Summing (3.55) and (3.25), we obtain

$$\begin{aligned}
&\textstyle\sum_{t=0}^{T-1} \gamma_t Q_\nu(z^{(t+1)}, z) + \sum_{t=0}^{T-1} \gamma_t [\langle \delta_t^G, x^{(t)} - x\rangle - \langle \delta_{t+1}^F, y^{(t+1)} - y_{t+1}^v\rangle] \\
&\leqslant \frac{\gamma_0 \tau_0}{2}[\|y - y^{(0)}\|_2^2 + \|y - y_1^v\|_2^2] + \gamma_0 \eta_0 W(x, x^{(0)}) \\
&+ \textstyle\sum_{t=1}^{T-1} \frac{3\gamma_t \theta_t^2}{2\tau_t} \|q_t - \bar{q}_t\|_2^2 + \frac{3\gamma_{T-1}}{2\tau_{T-1}} \|q_T - \bar{q}_T\|_2^2 \\
&+ \textstyle\sum_{t=0}^{T-1} \left[\frac{2\gamma_t}{\eta_t - L_0 - BL_f} \left\{\|\delta_t^G\|_*^2 + \left(\frac{L_f D_X}{2}[\|y\|_2 - B]_+\right)^2\right\} + \frac{\gamma_t}{2\tau_t} \|\delta_{t+1}^F\|_2^2\right],
\end{aligned} \tag{3.56}$$

for all $z \in \{(x, y) : x \in X, y \in \mathcal{B}_+^2(R)\}$. Note that given $\xi_{[t]}, u_{[t]}$ and $\bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}$, we have $y^{(t+1)}, y_{t+1}^v, x^{(t+1)}, x^{(t)}$ are constants. Hence, we have

$$\mathbb{E}[\langle \delta_{t+1}^F, y^{(t+1)} - y_{t+1}^v\rangle] = \mathbb{E}[\langle \mathbb{E}_{|\xi_{[t]}, u_{[t]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}}[\delta_{t+1}^F], y^{(t+1)} - y_{t+1}^v\rangle] = 0, \tag{3.57}$$

where second equality follows from (3.47). Choosing $z = \widehat{z} := (x^*, \widehat{y})$ in (3.56) where $\widehat{y} := (\|y^*\|_2 + 1)[f(\bar{x}_T)]_+ \|[f(\bar{x}_T)]_+\|_2^{-1} \in \mathcal{B}_+^2(R)$, taking expectation on both sides and noting (3.57), (3.48), (3.49), first relation in (3.45), we have

$$\begin{aligned}
\mathbb{E}\left[\textstyle\sum_{t=0}^{T-1} \gamma_t Q_\nu(z^{(t+1)}, \hat{z})\right] &\leqslant \frac{\gamma_0 \tau_0}{2} \mathbb{E}[\|\hat{y} - y^{(0)}\|_2^2 + \|\hat{y} - y_1^v\|_2^2] + \gamma_0 \eta_0 W(x^*, x^{(0)}) \\
&+ \textstyle\sum_{t=0}^{T-1} \frac{2\gamma_t}{\eta_t - L_0 - BL_f} \left\{\mathbb{E}[\|\delta_t^G\|_*^2] + \left(\frac{L_f D_X}{2}[\|y^*\|_2 + 1 - B]_+\right)^2\right\} \\
&+ \left(\textstyle\sum_{t=1}^{T-1} \frac{12\gamma_t \theta_t^2}{\tau_t} + \sum_{t=0}^{T-1} \frac{\gamma_t}{\tau_t} + \frac{12\gamma_{T-1}}{\tau_{T-1}}\right) (\sigma_{f,\nu}^2 + D_X^2 \|\sigma_\nu\|_2^2).
\end{aligned} \tag{3.58}$$

By Lemma 3.2.1, we then have $Q(z^{(t+1)}, \hat{z}) - [\nu_0^2 L_0 n + M_X n (\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}] \leqslant Q_\nu(z^{(t+1)}, \hat{z})$. Multiplying both sides by γ_t, summing from $t = 0$ to $T - 1$, taking expectation of both sides and dividing by Γ_T, we have

$$\frac{1}{\Gamma_T} \mathbb{E}\left[\textstyle\sum_{t=0}^{T-1} \gamma_t Q(z^{(t+1)}, \hat{z})\right] - [\nu_0^2 L_0 n + M_X n (\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}] \leqslant \frac{1}{\Gamma_T} \mathbb{E}\left[\textstyle\sum_{t=0}^{T-1} \gamma_t Q_\nu(z^{(t+1)}, \hat{z})\right]. \tag{3.59}$$

Noting the convexity of Q in the first argument, we obtain

$$\mathbb{E}[Q(\bar{z}_T, \hat{z})] \leqslant \frac{1}{\Gamma_T} \mathbb{E}\left[\textstyle\sum_{t=0}^{T-1} \gamma_t Q(z^{(t+1)}, \hat{z})\right]. \tag{3.60}$$

Now observe that we have $\mathcal{L}(\bar{x}_T, y^*) - \mathcal{L}(x^*, y^*) \geqslant 0$ which implies that $f_0(\bar{x}_T) + \langle y^*, f(\bar{x}_T)\rangle - f_0(x^*) \geqslant 0$, which follows from complementary slackness. In view of the relation

$$\langle y^*, f(\bar{x}_T)\rangle \leqslant \langle y^*, [f(\bar{x}_T)]_+\rangle \leqslant \|y^*\|_2 \|[f(\bar{x}_T)]_+\|_2,$$

the above inequality implies that

$$f_0(\bar{x}_T) + \|y^*\|_2 \|[f(\bar{x}_T)]_+\|_2 - f_0(x^*) \geqslant 0. \tag{3.61}$$

Moreover, we have that

$$\begin{aligned} Q(\bar{z}_T, \hat{z}) &= \mathcal{L}(\bar{x}_T, \hat{y}) - \mathcal{L}(x^*, \bar{y}_T) \\ &\geqslant \mathcal{L}(\bar{x}_T, \hat{y}) - \mathcal{L}(x^*, y^*) \\ &= f_0(\bar{x}_T) + (\|y^*\|_2 + 1)\|[f(\bar{x}_T)]_+\|_2 - f_0(x^*), \end{aligned}$$

which along with (3.61) implies that

$$Q(\bar{z}_T, \hat{z}) \geqslant \|[f(\bar{x}_T)]_+\|_2.$$

The above relation, (3.58), (3.59) and (3.60) together yield

$$\begin{aligned} \mathbb{E}[\|[f(\bar{x}_T)]_+\|_2] \leqslant \frac{1}{\Gamma_T}\Bigg[& \frac{\gamma_0\tau_0}{2}\mathbb{E}[\|\hat{y} - y^{(0)}\|_2^2 + \|\hat{y} - y_1^v\|_2^2] + \gamma_0\eta_0 W(x^*, x^{(0)}) \\ & + \textstyle\sum_{t=0}^{T-1} \frac{2\gamma_t}{\eta_t - L_0 - BL_f}\left\{\mathbb{E}[\|\delta_t^G\|_*^2] + \left(\frac{L_f D_X}{2}[\|y^*\|_2 + 1 - B]_+\right)^2\right\} \\ & + \left(\textstyle\sum_{t=1}^{T-1} \frac{12\gamma_t\theta_t^2}{\tau_t} + \sum_{t=0}^{T-1}\frac{\gamma_t}{\tau_t} + \frac{12\gamma_{T-1}}{\tau_{T-1}}\right)(\sigma_{f,\nu}^2 + D_X^2\|\sigma_\nu\|_2^2)\Bigg] \\ & + [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]. \end{aligned}$$

Noting the bound $\|\hat{y} - y_1^v\| \leqslant 2R$ and $\|\hat{y} - y^{(0)}\|_2^2 \leqslant 2\|y^{(0)}\|_2^2 + 2\|\hat{y}\|_2^2 \leqslant 2\|y^{(0)}\|_2^2 + 2R^2$ in the above relation and recalling that $R = \|y^*\|_2 + 1$, we obtain (3.43). Hence, we conclude the proof. ■

Lemma 3.2.7. Assume that $\{\gamma_t, \tau_t, \eta_t\}$ satisfy

$$\frac{96\|\sigma_\nu\|_2^2}{\tau_t(\eta_t - L_0 - BL_f)} < 1, \tag{3.62}$$

for all $t \leqslant T-1$ and constants R_1 and R_2 satisfying the following conditions exist:

$$R_1 \geqslant \left(1 - \frac{96\|\sigma_\nu\|_2^2}{\tau_t(\eta_t - L_0 - BL_f)}\right)^{-1} \left[2\sigma_{0,\nu_0}^2 + \frac{48\|\sigma_\nu\|_2^2}{\gamma_t\tau_t}\left\{\gamma_0\eta_0 W(x^*, x^{(0)}) + \frac{\gamma_0\tau_0}{2}\|y^* - y^{(0)}\|_2^2 \right.\right.$$
$$+ \frac{\gamma_t\tau_t}{12}\|y^*\|_2^2 + \textstyle\sum_{i=0}^t \frac{2\gamma_i}{\eta_i - L_0 - BL_f}\left(\frac{L_f D_X}{2}[\|y^*\|_2 - B]_+\right)^2$$
$$\left.\left. + \left(\textstyle\sum_{i=1}^t \frac{12\gamma_i\theta_i^2}{\tau_i} + \frac{12\gamma_t}{\tau_t}\right)(\sigma_{f,\nu}^2 + D_X^2\|\sigma_\nu\|_2^2) + [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\Gamma_{t+1}\right\}\right] \tag{3.63}$$

for all $t \leqslant T-1$ and

$$R_2 \geqslant \left(1 - \frac{96\|\sigma_\nu\|_2^2}{\tau_t(\eta_t - L_0 - BL_f)}\right)^{-1} \frac{96\|\sigma_\nu\|_2^2\gamma_i}{\gamma_t\tau_t(\eta_i - L_0 - BL_f)} \tag{3.64}$$

for all $t \leqslant T-1$ and $i \leqslant t-1$. Then, we have

$$\mathbb{E}[\|\delta_t^G\|_*^2] \leqslant R_1(1 + R_2)^t, \tag{3.65}$$

for all $t \leqslant T-1$. In particular, if $\|\sigma_\nu\|_2 = 0$, then we can set $R_1 = 2\sigma_{0,\nu_0}^2$ and $R_2 = 0$ implying $\mathbb{E}[\|\delta_t^G\|_*^2] \leqslant 2\sigma_{0,\nu_0}^2$.

Proof. First note that by Lemma 3.2.1, we have

$$Q(z^{(i+1)}, z) - [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}] \leqslant Q_\nu(z^{(i+1)}, z).$$

Multiplying the above by γ_i and summing up $i = 0$ to t, we have

$$\textstyle\sum_{i=0}^t \gamma_i Q(z^{(i+1)}, z) - [\nu_0^2 L_0 n + M_X n(\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\Gamma_{t+1} \leqslant \sum_{i=0}^t \gamma_i Q_\nu(z^{(i+1)}, z).$$

Replacing T for $t + 1(\geqslant 1)$ in (3.25), we have

$$\textstyle\sum_{i=0}^t \gamma_i Q_\nu(z^{(i+1)}, z) + \sum_{i=0}^t \gamma_i[\langle\delta_i^G, x^{(i)} - x\rangle - \langle\delta_{i+1}^F, y^{(i+1)} - y\rangle]$$
$$\leqslant \gamma_0\eta_0 W(x, x^{(0)}) - \gamma_t\eta_t W(x, x^{(t+1)}) + \frac{\gamma_0\tau_0}{2}\|y - y^{(0)}\|_2^2 - \frac{\gamma_t\tau_t}{12}\|y - y^{(t+1)}\|_2^2$$
$$+ \textstyle\sum_{i=0}^t \frac{2\gamma_i}{\eta_i - L_0 - BL_f}\left[\|\delta_i^G\|_*^2 + \left(\frac{L_f D_X}{2}[\|y\|_2 - B]_+\right)^2\right]$$
$$+ \textstyle\sum_{i=1}^t \frac{3\gamma_i\theta_i^2}{2\tau_i}\|q_i - \bar{q}_i\|_2^2 + \frac{3\gamma_t}{2\tau_t}\|q_{t+1} - \bar{q}_{t+1}\|_2^2. \tag{3.66}$$

Observe that $Q(z^{(i+1)}, z^*) \geqslant 0$ for $i = 0, \ldots, t$ by our saddle point assumption where $z^* = (x^*, y^*)$. Choosing $z = z^*$ (both non-random) in the above relations, taking expectation, using (3.45) with $x = x^*$ and (3.46) with $y = y^*$, disregarding the term $-\gamma_t \eta_t \mathbb{E}[W(x^*, x^{(t+1)})]$ and noting (3.49), we have

$$
\begin{aligned}
&- [\nu_0^2 L_0 n + M_X n (\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\Gamma_{t+1} + \tfrac{\gamma_t \tau_t}{12}\mathbb{E}\|y^* - y^{(t+1)}\|_2^2 \\
&\leqslant \gamma_0 \eta_0 W(x^*, x^{(0)}) + \frac{\gamma_0 \tau_0}{2}\|y^* - y^{(0)}\|^2 \\
&+ \textstyle\sum_{i=0}^t \tfrac{2\gamma_i}{\eta_i - L_0 - BL_f}\left[\mathbb{E}[\|\delta_i^G\|_*^2] + \left(\tfrac{L_f D_X}{2}[\|y^*\|_2 - B]_+\right)^2\right] \\
&+ \left(\textstyle\sum_{i=1}^t \tfrac{12\gamma_i \theta_i^2}{\tau_i} + \tfrac{12\gamma_t}{\tau_t}\right)(\sigma_{f,\nu}^2 + D_X^2\|\sigma_\nu\|_2^2). \qquad (3.67)
\end{aligned}
$$

Now, let us define $\delta_{t,i}^G := G_{i,\nu_i}(x^{(t)}, \xi_i^{(t)}, u_i^{(t)}) - f'_{i,\nu_i}(x^{(t)})$ for $i = 0, \ldots, m$. As a consequence, we have $\delta_t^G = \delta_{t,0}^G + \sum_{i=1}^m y_i^{(t+1)}\delta_{t,i}^G$. Then, we have

$$
\begin{aligned}
\mathbb{E}[\|\delta_t^G\|_*^2] &= \mathbb{E}[\|\delta_{t,0}^G + \textstyle\sum_{i=1}^m y_i^{(t+1)}\delta_{t,i}^G\|_*^2] \\
&\overset{(i)}{\leqslant} 2\mathbb{E}[\|\delta_{t,0}^G\|_*^2] + 2\mathbb{E}[\|\textstyle\sum_{i=1}^m y_i^{(t+1)}\delta_{t,i}^G\|_*^2] \\
&\leqslant 2\mathbb{E}[\|\delta_{t,0}^G\|_*^2] + 2\mathbb{E}[(\textstyle\sum_{i=1}^m \|y_i^{(t+1)}\delta_{t,i}^G\|)^2] \\
&\overset{(ii)}{\leqslant} 2[\sigma_{0,\nu_0}^2 + \mathbb{E}[\|y^{(t+1)}\|_2^2(\textstyle\sum_{i=1}^m \|\delta_{t,i}^G\|_*^2)]] \\
&\overset{(iii)}{\leqslant} 2[\sigma_{0,\nu_0}^2 + \mathbb{E}[\|y^{(t+1)}\|_2^2(\textstyle\sum_{i=1}^m \mathbb{E}_{|\xi_{[t-1]}, u_{[t-1]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}}[\|\delta_{t,i}^G\|_*^2])]] \\
&\overset{(iv)}{\leqslant} 2[\sigma_{0,\nu_0}^2 + \mathbb{E}[\|y^{(t+1)}\|_2^2 \textstyle\sum_{i=1}^m \sigma_{i,\nu_i}^2]] \\
&= 2(\sigma_{0,\nu_0}^2 + \|\sigma_\nu\|_2^2 \mathbb{E}\|y^{(t+1)}\|_2^2) \\
&\leqslant 2\sigma_{0,\nu_0}^2 + 4\|\sigma_\nu\|_2^2(\|y^*\|_2^2 + \mathbb{E}[\|y^{(t+1)} - y^*\|_2^2]). \qquad (3.68)
\end{aligned}
$$

Here, relation (i) follows due to the fact that $\|a+b\|_*^2 \leqslant (\|a\|_* + \|b\|_*)^2 \leqslant 2\|a\|_*^2 + 2\|b\|_*^2$, relation (ii) follows due to Cauchy-Schwarz inequality, relation (iii) follows due to the fact that $y^{(t+1)}$ is a constant conditioned on random variables $\xi_{[t-1]}, u_{[t-1]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}$ and relation (iv) follows from the fact that $x^{(t)}$ is a constant conditioned on random variables $\xi_{[t-1]}, u_{[t-1]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}$.

Adding $\frac{\gamma_t\tau_t}{12}\|y^*\|_*^2$ to both sides of (3.67), then multiplying it by $\frac{48\|\sigma_\nu\|_2^2}{\gamma_t\tau_t}$ and observing (3.68), we have

$$\begin{aligned}\mathbb{E}[\|\delta_t^G\|_*^2] \leqslant 2\sigma_{0,\nu_0}^2 + \frac{48\|\sigma_\nu\|_2^2}{\gamma_t\tau_t}\Big\{&\gamma_0\eta_0 W(x^*, x^{(0)}) + \frac{\gamma_0\tau_0}{2}\|y^* - y^{(0)}\|_2^2 + \frac{\gamma_t\tau_t}{12}\|y^*\|_2^2 \\ &+ \textstyle\sum_{i=0}^t \frac{2\gamma_i}{\eta_i - L_0 - BL_f}\left(\frac{L_f D_X}{2}[\|y^*\|_2 - B]_+\right)^2 \\ &+ \left(\textstyle\sum_{i=1}^t \frac{12\gamma_i\theta_i^2}{\tau_i} + \frac{12\gamma_t}{\tau_t}\right)(\sigma_{f,\nu}^2 + D_X^2\|\sigma_\nu\|_2^2) + [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\Gamma_{t+1}\Big\} \\ &+ \textstyle\sum_{i=0}^t \frac{96\|\sigma_\nu\|_2^2\gamma_i}{\gamma_t\tau_t(\eta_i - L_0 - BL_f)}\mathbb{E}[\|\delta_i^G\|_*^2].\end{aligned}$$

In view of (3.62), we have that the coefficient of the δ_t^G term on the right hand side of the above relation is strictly less than 1. Moving the δ_t^G term to the left hand side and noting the conditions imposed on constants R_1, R_2, we have

$$\mathbb{E}[\|\delta_t^G\|_*^2] \leqslant R_1 + R_2 \textstyle\sum_{i=0}^{t-1} \mathbb{E}[\|\delta_i^G\|_*^2],$$

for all $t \leqslant T - 1$. Using Lemma 3.2.4 for the above relation, we have (3.65). Hence we conclude the proof. ∎

Theorem 3.2.8. *Suppose the functions f_i, for $i = 0, \ldots, m$, are convex and satisfy Assumptions 4, 6 and 5. Let $B \geqslant 1$ be a given constant and define $\mathcal{H}_* := (L_f D_X[\|y^*\|_2 + 1 - B]_+)/2$. Set $y^{(0)} = \mathbf{0}$ and $\{\gamma_t, \theta_t, \eta_t, \tau_t\}$ in Algorithm 3 according to the following: $\gamma_t = 1$, $\eta_t = L_0 + BL_f + \eta$, and $\theta_t = 1$, $\tau_t = \tau$, where*

$$\eta := \max\left\{\frac{\sqrt{2T[\mathcal{H}_*^2 + \sigma_{0,\nu_0}^2 + 48B^2\|\sigma_\nu\|_2^2]}}{D_X}, \frac{6B\max\{2M_f, 4\|\sigma_\nu\|_2\}}{D_X}\right\},$$

$$\tau := \max\left\{\frac{\sqrt{96T}\sigma_{X,f}}{B}, \frac{2D_X\max\{2M_f, 4\|\sigma_\nu\|_2\}}{B}\right\}.$$

Then, we have

$$
\begin{aligned}
\mathbb{E}[f_0(\bar{x}_T) - f_0(x^*)] \leqslant & \frac{(L_0 + BL_f)D_X^2 + \max\{12M_f, 24\|\sigma_\nu\|_2\}BD_X}{T} \\
& + \frac{1}{\sqrt{T}}\left\{\frac{\sqrt{2}\zeta^2 D_X}{\sqrt{\mathcal{H}_*^2 + \sigma_{0,\nu_0}^2 + 48B^2\|\sigma_\nu\|_2^2}} + \frac{\sqrt{3}B\sigma_{X,f}}{\sqrt{2}}\right\} \\
& + \frac{1}{\sqrt{T}}\sqrt{2(\mathcal{H}_*^2 + \sigma_{0,\nu_0}^2 + 48B^2\|\sigma_\nu\|_2^2)}D_X \\
& + [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}],
\end{aligned}
\tag{3.69}
$$

and

$$
\begin{aligned}
\mathbb{E}[\|[f(\bar{x}_T)]_+\|_2] \leqslant & + \frac{1}{\sqrt{T}}\left\{\left[\frac{12\sqrt{6}(\|y^*\|_2 + 1)^2}{B} + \frac{13B}{4\sqrt{6}}\right]\sigma_{X,f}\right. \\
& + \sqrt{2}D_X\left[\sqrt{\mathcal{H}_*^2 + \sigma_{0,\nu_0}^2 + 48B^2\|\sigma_\nu\|_2^2}\right. \\
& \left.\left. + \frac{\zeta^2 + \mathcal{H}_*^2}{\sqrt{\mathcal{H}_*^2 + \sigma_{0,\nu_0}^2 + 48B^2\|\sigma_\nu\|_2^2}}\right]\right\} \\
& + [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}] + \\
& \frac{(L_0 + BL_f)D_X^2 + \max\{12M_f, 24\|\sigma_\nu\|_2\}D_X\left(B + \frac{(\|y^*\|_2+1)^2}{B}\right)}{T},
\end{aligned}
\tag{3.70}
$$

where $\zeta \coloneqq 2e\{\sigma_{0,\nu_0}^2 + \|\sigma_\nu\|_2^2(14\|y^*\|_2^2 + 75B^2) + 2\sqrt{3}\|\sigma_\nu\|_2(2B\mathcal{H}_* + B\sigma_{0,\nu_0} + \sqrt{48}B^2\|\sigma_\nu\|_2)$
$+ \sqrt{6}D_X^{-1}\|\sigma_\nu\|_2 B[\nu_0^2 L_0 n + M_X n(\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\sqrt{T}\}^{1/2}$.

Hence, by choosing,

$$
\nu_0 \leqslant \min\left\{\frac{1}{\sqrt{2L_0 n\sqrt{T}}}, \frac{2}{(n+3)^{3/2}}, \frac{1}{L_i(n+6)^{3/2}}\right\}, \tag{3.71}
$$

$$
\nu_i \leqslant \min\left\{\frac{2}{(n+3)^{3/2}}, \frac{1}{2M_i\sqrt{(n+2)m}},\right. \tag{3.72}
$$

$$
\left.\frac{1}{\sqrt{L_i n\sqrt{m}}}, \frac{1}{\sqrt{2L_i n M_X\sqrt{Tm}}}, \frac{1}{L_i(n+6)^{3/2}\sqrt{m}}\right\}, \tag{3.73}
$$

for $i \in [m]$, the number of calls to the stochastic zeroth-order oracle required by Algorithm 3 to find an ε-approximately optimal solution of (1.7) is of the order

$$\mathcal{O}\left(\frac{(m+1)n}{\epsilon^2}\right).$$

We are now ready to prove Theorem 3.2.8.

Proof of Theorem 3.2.8. It is easy to verify that $\{\gamma_t, \theta_t, \eta_t, \tau_t\}$ set according to Theorem 3.2.8 satisfies (3.23). Note that (3.24) is satisfied if $\mathcal{M}^2 \leqslant \frac{\tau_t(\eta_{t-2}-L_0-BL_f)}{12}$ where $\mathcal{M} := 2M_f$. This follows due to the fact that $\{\eta_t\}$ is a non-decreasing sequence and $\theta_t = 1$ for all $t \geqslant 0$. Then we have

$$\frac{\tau_t(\eta_{t-2} - L_0 - BL_f)}{12} \geqslant \frac{6\mathcal{M}B}{D_X}\frac{2\mathcal{M}D_X}{B} \times \frac{1}{12} = \mathcal{M}^2.$$

Also, since $(\eta_t - L_0 - BL_f) \geqslant \frac{24B\|\sigma_\nu\|_2}{D_X}$ and $\tau_t \geqslant \frac{8D_X\|\sigma_\nu\|_2}{B}$, we have

$$\tau_t(\eta_t - L_0 - BL_f) \geqslant 192\|\sigma_\nu\|_2^2$$

for all $t \geqslant 0$. In view of the above relation, we have

$$\frac{96\|\sigma_\nu\|_2^2}{\tau_t(\eta_t - L_0 - BL_f)} \leqslant \frac{1}{2}, \tag{3.74}$$

hence (3.62) is satisfied. We also need to show the existence of R_1 and R_2 satisfying (3.63) and (3.64), respectively. Using the fact that γ_t, η_t and τ_t are constants for all $t \geqslant 0, \tau\eta \geqslant \frac{96T\sigma_{X,f}\|\sigma_\nu\|_2}{D_X}$ and noting (3.74), we obtain

$$\left(1 - \frac{96\|\sigma_\nu\|_2^2}{\tau_t(\eta_t - L_0 - BL_f)}\right)^{-1}\frac{96\|\sigma_\nu\|_2^2\gamma_i}{\gamma_t\tau_t(\eta_i - L_0 - BL_f)} \leqslant 2\frac{96\|\sigma_\nu\|_2^2}{\tau\eta} \leqslant 2\frac{\|\sigma_\nu\|_2 D_X}{T\sigma_{X,f}} \leqslant \frac{2}{T},$$

where in the last relation, we used the fact that $\sigma_{X,f} \geqslant D_X\|\sigma_\nu\|_2$. In view of the above relation and (3.64), we can set

$$R_2 := \frac{2}{T}. \tag{3.75}$$

Noting (3.63) along with the fact that $\mathcal{H}_* \geqslant \frac{L_f D_X[\|y^*\|_2 - B]_+}{2}$, setting $y^{(0)} = \mathbf{0}$, using (3.74), (3.62), $\gamma_t\tau_t = \tau \geqslant \frac{\sqrt{96T}\sigma_{X,f}}{B}$, $\sum_{i=0}^t \frac{\gamma_i}{\eta_i - L_0 - BL_f} = \frac{t+1}{\eta} \leqslant \frac{\sqrt{T}D_X}{\sqrt{2[\mathcal{H}_*^2 + \sigma_{0,\nu_0}^2 + 48B^2\|\sigma_\nu\|_2^2]}}$, and $\sum_{i=1}^t \frac{\gamma_i\theta_i^2}{\tau_i} + \frac{\gamma_t}{\tau_t} = \frac{t+1}{\tau} \leqslant \frac{T}{\tau}$ for all $t \leqslant T-1$, we can see that the RHS of (3.63) is at most

$$
\begin{aligned}
&2\bigg[2\sigma_{0,\nu_0}^2 + 48\|\sigma_\nu\|_2^2\bigg\{\frac{7}{12}\|y^*\|_2^2 + \frac{\eta}{\tau}D_X^2 + \frac{\sqrt{2T}D_X\mathcal{H}_*^2}{\sqrt{\mathcal{H}_*^2 + \sigma_{0,\nu_0}^2 + 48B^2\|\sigma_\nu\|_2^2}}\frac{B}{\sqrt{96T}\sigma_{X,f}} + 12\sigma_{X,f}^2\frac{T}{\tau^2} \\
&+ \frac{B[\nu_0^2 L_0 n + M_X n(\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\sqrt{T}}{4\sqrt{6}\sigma_{X,f}}\bigg\}\bigg] \\
&\leqslant 2\bigg[2\sigma_{0,\nu_0}^2 + 48\|\sigma_\nu\|_2^2\bigg\{\frac{7}{12}\|y^*\|_2^2 + \frac{\eta}{\tau}D_X^2 + \frac{D_X B\mathcal{H}_*}{\sqrt{48}\sigma_{X,f}} + 12T\sigma_{X,f}^2\frac{B^2}{96T\sigma_{X,f}^2} \\
&+ \frac{B[\nu_0^2 L_0 n + M_X n(\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\sqrt{T}}{4\sqrt{6}\sigma_{X,f}}\bigg\}\bigg] \\
&\leqslant 2\bigg[2\sigma_{0,\nu_0}^2 + 48\|\sigma_\nu\|_2^2\bigg\{\frac{7}{12}\|y^*\|_2^2 + \frac{D_X}{\sigma_{X,f}}\left(B\sqrt{\frac{[\mathcal{H}_*^2 + \sigma_{0,\nu_0}^2 + 48B^2\|\sigma_\nu\|_2^2]}{48}} + \frac{B\mathcal{H}_*}{\sqrt{48}}\right) \\
&+ \frac{6\max\{\mathcal{M}, 4\|\sigma_\nu\|_2\}BD_X}{2\max\{\mathcal{M}, 4\|\sigma_\nu\|_2\}}\frac{B}{D_X} + \frac{B^2}{8} + \frac{B[\nu_0^2 L_0 n + M_X n(\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\sqrt{T}}{4\sqrt{6}\sigma_{X,f}}\bigg\}\bigg] \\
&\leqslant 2\bigg[2\sigma_{0,\nu_0}^2 + 28\|\sigma_\nu\|_2^2\|y^*\|_2^2 + 150B^2\|\sigma_\nu\|_2^2 + \sqrt{48}\|\sigma_\nu\|_2[2B\mathcal{H}_* + (B\sigma_{0,\nu_0} + \sqrt{48}B^2\|\sigma_\nu\|_2)] \\
&+ 2\sqrt{6}D_X^{-1}\|\sigma_\nu\|_2 B[\nu_0^2 L_0 n + M_X n(\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\sqrt{T}\bigg] \\
&=: R_1
\end{aligned} \tag{3.76}
$$

where in the last inequality, we used the fact that $\frac{\|\sigma_\nu\|_2 D_X}{\sigma_{X,f}} \leqslant 1$. Note that the last term in the above sequence of relations is a constant satisfying the requirement in (3.63). Hence, we can set

$$
\begin{aligned}
R_1 := 2\bigg[&2\sigma_{0,\nu}^2 + 28\|\sigma_\nu\|_2^2\|y^*\|_2^2 + 150B^2\|\sigma_\nu\|_2^2 + \sqrt{48}\|\sigma_\nu\|_2[2B\mathcal{H}_* + (B\sigma_{0,\nu} + \sqrt{48}B^2\|\sigma_\nu\|_2)] \\
&+ 2\sqrt{6}D_X^{-1}\|\sigma_\nu\|_2 B[\nu_0^2 L_0 n + M_X n(\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}]\sqrt{T}\bigg].
\end{aligned} \tag{3.77}
$$

Then using Lemma 3.2.7 and noting (3.75), we have for all $t \leqslant T-1$

$$\mathbb{E}[\|\delta_t^G\|_*^2] \leqslant \begin{cases} 4\sigma_{0,\nu_0}^2 & \text{if } \|\sigma_\nu\|_2 = 0; \\ R_1\left(1+\frac{2}{T}\right)^{T-1} \leqslant R_1 e^2 & \text{otherwise.} \end{cases}$$

Noting the above relation, (3.77) and the definition of ζ, we have

$$\mathbb{E}[\|\delta_t^G\|_*^2] \leqslant \zeta^2, \quad \forall t \leqslant T-1. \tag{3.78}$$

Hence, according to (3.42) with $y^{(0)} = \mathbf{0}$ and using (3.78), we have

$$\mathbb{E}[f_0(\bar{x}_T) - f_0(x^*)] \leqslant \frac{1}{T}\left[(\eta + L_0 + BL_f)W(x^*, x^{(0)}) + \frac{2T\zeta^2}{\eta} + 12\sigma_{X,f}^2\frac{T}{\tau}\right]$$
$$+ [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}].$$

Using the bound $W(x^*, x^{(0)}) \leqslant D_X^2$, we obtain (3.69). From (3.43) and (3.78), we have for $T \geqslant 1$

$$\mathbb{E}\|[f(\bar{x}_T)]_+\|_2 \leqslant \frac{1}{T}\left[3(\|y^*\|_2+1)^2\tau + (\eta + L_0 + BL_f)W(x^*, x^{(0)}) + \frac{2(\zeta^2 + \mathcal{H}_*^2)T}{\eta} + \frac{13\sigma_{X,f}^2 T}{\tau}\right]$$
$$+ [\nu_0^2 L_0 n + M_X n(\textstyle\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}].$$

Using bounds $W(x^*, x^{(0)}) \leqslant D_X^2$, we obtain (3.70). Define

$$\bar{\sigma_f}^2 := 2(1+\sigma_f^2), \tag{3.79}$$

$$\bar{\sigma}_0^2 := 1 + 10(n+4)[\sigma_0^2 + [L_0(1+D_X) + M_0]^2], \tag{3.80}$$

$$\bar{\sigma_i}^2 := \frac{1}{m} + 10(n+4)[\sigma_i^2 + [L_i(1+D_X) + M_i]^2] \quad \text{for } i \in \{1, \dots, m\}, \tag{3.81}$$

$$\bar{\sigma}^2 = 1 + 10(n+4)[\|\sigma\|_2^2 + 2L_f^2(1+D_X)^2 + 2M_f^2], \tag{3.82}$$

$$\overline{\sigma_{X,f}} = (2(1+\sigma_f^2) + D_X^2\bar{\sigma}^2)^{1/2}, \tag{3.83}$$

$$\bar{\zeta} := 2e\left\{\bar{\sigma_0}^2 + \bar{\sigma}^2(14\|y^*\|_2^2 + 75B^2) + 2\sqrt{3}\bar{\sigma}(2B\mathcal{H}_* + B\bar{\sigma}_0 + \sqrt{48}B^2\bar{\sigma}) + \sqrt{6}D_X^{-1}\bar{\sigma}B\right\}^{1/2}. \tag{3.84}$$

By choice of ν_0, ν_i for $i \in [m]$, definition of $\sigma_{f,\nu}^2$, $\tilde{B}_i$, σ_{i,ν_i}^2, and σ_ν, we have

$$\begin{aligned}
\sigma_{f,\nu}^2 &\leqslant 2 + 2\sigma_f^2 =: \overline{\sigma}_f^2, \\
\nu_0^2 L_0 n + M_X n \left(\sum_{i=1}^m \nu_i^4 L_i^2\right)^{1/2} &\leqslant \frac{1}{\sqrt{T}}, \\
\tilde{B}_i &\leqslant L_i(1 + D_X) + M_i, \\
\sigma_{0,\nu_0}^2 &\leqslant 1 + 10(n+4)[\sigma_0^2 + [L_0(1+D_X) + M_0]^2], \\
\sigma_{i,\nu_i}^2 &\leqslant \frac{1}{m} + 10(n+4)[\sigma_i^2 + [L_i(1+D_X) + M_i]^2] =: \overline{\sigma}_i^2 \quad \text{ for } i \in [m], \\
\|\sigma_\nu\|_2^2 &\leqslant 1 + 10(n+4)[\|\sigma\|_2^2 + 2L_f^2(1+D_X)^2 + 2M_f^2] =: \overline{\sigma}^2.
\end{aligned}$$

Using these relations, we see that $\sigma_{X,f} \leqslant \overline{\sigma_{X,f}}$ and $\zeta \leqslant \overline{\zeta}$. Hence, we have

$$\begin{aligned}
\mathbb{E}[f_0(\bar{x_T}) - f_0(x^*)] \leqslant{} & \frac{(L_0 + BL_f)D_X^2 + \max\{6\mathcal{M}, 24\overline{\sigma}\}BD_X}{T} \\
& + \frac{1}{\sqrt{T}}\left\{\frac{\sqrt{2}D_X\overline{\zeta^2}}{\sqrt{\mathcal{H}_*^2 + \sigma_0^2 + 48B^2\|\sigma\|_2^2}} + \frac{\sqrt{3}B\overline{\sigma_{X,f}}}{\sqrt{2}}\right\} \\
& + \frac{1}{\sqrt{T}}\left[\sqrt{2(\mathcal{H}_*^2 + \bar{\sigma}_0^2 + 48B^2\bar{\sigma}^2)}D_X + 1\right]
\end{aligned} \tag{3.85}$$

and

$$\begin{aligned}
\mathbb{E}[\|[f(\bar{x}_T)]_+\|_2] \leqslant{} & \frac{(L_0 + BL_f)D_X^2 + \max(6\mathcal{M}, 24\overline{\sigma})D_X\left(B + \frac{(\|y^*\|_2+1)^2}{B}\right)}{T} \\
& \frac{1}{\sqrt{T}}\left\{\left[\frac{12\sqrt{6}(\|y^*\|_2 + 1)^2}{B} + \frac{13B}{4\sqrt{6}}\right]\overline{\sigma_{X,f}}\right\} \\
& + \frac{1}{\sqrt{T}}\left\{\sqrt{2}D_X\left[\sqrt{\mathcal{H}_*^2 + \overline{\sigma}_0^2 + 48B^2\overline{\sigma}^2} + \frac{\overline{\zeta}^2 + \mathcal{H}_*^2}{\sqrt{\mathcal{H}_*^2 + \sigma_0^2 + 48B^2\|\sigma\|_2^2}}\right]\right\} \\
& + \frac{1}{\sqrt{T}}.
\end{aligned} \tag{3.86}$$

As a consequence, to obtain an $(\varepsilon, \varepsilon)$-optimal solution with Algorithm 1, we need the number of iterations to be

$$T := \max\Bigg\{ \frac{25}{\varepsilon^2}, \frac{5(L_0 + BL_f)D_X^2 + 5\max(6\mathcal{M}, 24\overline{\sigma})D_X\left(B + \frac{(\|y^*\|_2+1)^2}{B}\right)}{\varepsilon},$$

$$\frac{\overline{\sigma_{X,f}^2}}{\varepsilon^2}\left[\frac{60\sqrt{6}(\|y^*\|_2+1)^2}{B} + \frac{65B}{4\sqrt{6}}\right]^2,$$

$$\frac{50}{\varepsilon^2}\left[D_X\sqrt{\mathcal{H}_*^2 + \bar{\sigma}_0^2 + 48B^2\overline{\sigma}^2} + \frac{D_X(\overline{\zeta}^2 + \mathcal{H}_*^2)}{\sqrt{\mathcal{H}_*^2 + \sigma_0^2 + 48B^2\|\sigma\|_2^2}}\right]^2\Bigg\}. \tag{3.87}$$

Now, by the choice of ν_o and ν_i in (3.72) and (3.73) respectively, we see that the oracle complexity is given by $\mathcal{O}((m+1)n)/\epsilon^2)$. ■

Remark 3.2.1. *Although the parameter settings of Theorem 3.2.8 and the right hand side of* (3.69) *and* (3.70) *appear complicated to parse, the important take away message is that the right hand side of* (3.69) *and* (3.70) *are of the order* $\mathcal{O}(1/\sqrt{T})$ *which leads to the oracle complexity described above. Furthermore, the order of* ϵ *in the oracle complexity is of the same order as that in [25] for the stochastic first-order setting. The presence of* $(m+1)n$ *in the oracle complexity is due to the fact that we are required to estimate* $m+1$ *gradient vectors, each of dimension* n*. This also illustrates that the oracle complexity in the zeroth-order setting is linear in the number of constraints* m*, for a fixed dimensionality* n*. The dimension dependency is unavoidable even in the unconstrained setting, as showed via lower bounds in [46, 74].*

3.3. Meta-Algorithm for Nonconvex Setting

We now consider the case when objective function f_0, and the constraint functions $f_1, \ldots, f_m$ are nonconvex. In this case, [25], proposed a two-step meta-algorithm: (i) construct a sequence of convex relaxations for the nonconvex problem, and (ii) leverage the algorithm developed for the convex setting. Given our Algorithm 3, we leverage this framework to solve (1.7) in the nonconvex setting. Before proceeding, we need a notion of optimality for the nonconvex setting, which we discuss below.

We first define the exact Karush-Kuhn-Tucker (KKT) condition for (1.7) as follows. For a convex set X, we denote interior as int$X$, the normal cone at $x \in X$ as $N_X(x)$, and its dual cone as $N_X^*(x)$. Let $\oplus$ denote the Minkowski sum of two sets. We refer to the distance between two sets $A, B \subset \mathbb{R}^n$ as $d(A, B) := \inf_{a\in A, b\in B} \|a - b\|$.

DEFINITION 3.3.1. We say that $x^* \in X$ is a critical KKT point of (1.7) if $f_i(x^*) \leqslant 0$ and $\exists y^* := [y_1^*, \ldots, y_m^*]^T \geqslant \mathbf{0}$ such that

$$y_i^* f_i(x^*) = 0, \quad i \in [m],$$
$$d(\nabla f_0(x^*) + \textstyle\sum_{i=1}^m y_i^* \nabla f_i(x^*) \oplus N_X(x^*), \mathbf{0}) = 0.$$

The parameters $\{y_i^*\}_{i\in[m]}$ are called *Lagrange multipliers*. For brevity, we use the notation y^* and $[y_1^*, \ldots, y_m^*]^T$ interchangeably. With this definition, we also have the following approximate KKT condition which is the standard approximate optimality condition for solving (1.7) in the nonconvex setting.

DEFINITION 3.3.2. We say that a point $\hat{x} \in X$ is an (ε, δ)-KKT point in expectation for (1.7) if there exists $(\bar{x}, \bar{y})$ such that $f(\bar{x}) \leqslant \mathbf{0}, \bar{y} \geqslant \mathbf{0}$ and

$$\mathbb{E}[\textstyle\sum_{i=1}^m |\bar{y}_i f_i(\bar{x})|] \leqslant \varepsilon,$$
$$\mathbb{E}[(d(\nabla f_0(\bar{x}) + \textstyle\sum_{i=1}^m \bar{y}_i \nabla f_i(\bar{x}) \oplus N_X(\bar{x}), \mathbf{0}))^2] \leqslant \varepsilon,$$
$$\mathbb{E}[\|\bar{x} - \hat{x}\|^2] \leqslant \delta.$$

PROPOSITION 3.3.1. *Consider solving* (1.7) *with both the objective and the constraint function being nonconvex and satisfying Assumptions 4, 6 and 5. Then, by running Algorithm 4 with* $K = \mathcal{O}(1/\epsilon)$*, we obtain* $(\epsilon, 2\epsilon/2\mu_0 c_1)$*-KKT point. Hence, the total number of calls to the stochastic zeroth-order oracle is given by*

$$\mathcal{O}\left(\frac{(m+1)n}{\epsilon^3}\right).$$

PROOF OF PROPOSITION 3.3.1. The claim follows immediately by Theorem 3.2.8 and Corollary 3.19 from [25]. ∎

To the best of our knowledge, we are not aware of a non-asymptotic result on the oracle complexity of stochastic zeroth-order optimization with stochastic zeroth-order functional constraints, in both the convex and nonconvex settings.

Algorithm 4 Meta-Algorithm for Nonconvex Setting

Input: Input x_0

1: **for** $k = 1, \ldots, K$ **do**

2: Set:

$$f_0(x; x_{k-1}) := f_0(x) + 2\mu_0 W(x, x_{k-1}),$$

$$f_i(x; x_{k-1}) := f_i(x) + 2\mu_i W(x, x_{k-1}), \quad i \in [m].$$

3: Obtain an ϵ-approximately optimal solution to the problem:

$$\arg\min_{x \in X} f_0(x; x_{k-1}) \tag{3.88}$$

$$\text{s.t.} \quad f_i(x; x_{k-1}) \leqslant 0, \quad i \in [m]. \tag{3.89}$$

by using `SZO-ConEx` in Algorithm 3. Denote it by x_k, for $k = 1, \ldots, K$.

4: Randomly choose $\hat{k} \in \{1, \ldots, K\}$

5: **return** $x_{\hat{k}}$.

3.4. Conclusion

In this project, we proposed and analyzed stochastic zeroth-order optimization algorithms for nonlinear optimization problems with functional constraints. We consider the case when both the objective function and the constraint functions are observed only via noisy function queries. Our algorithm is based on leveraging the constraint extrapolation technique proposed by [25] and the Gaussian smoothing technique. We characterize the oracle complexity of the proposed algorithm in both the convex and nonconvex setting. We also apply our methodology for the problem of hyperparameter tuning for the HMC algorithm and demonstrate its superior performance. For future work, we plan to develop parallel versions of our algorithm for the case when the objective functions and the constraint functions are available only locally in different machines. It is also interesting to develop lower bounds on the oracle complexity of stochastic zeroth-order optimization algorithms in the constrained setting. Finally, it is of great interest to find other applications of the proposed methodology in statistical machine learning, robotics, and other scientific and engineering fields.

Bibliography

[1] Luigi Acerbi and Wei Ji Ma. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. In *Advances in neural information processing systems*, pages 1836–1846, 2017.

[2] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Conference on Learning Theory*, pages 28–40, 2010.

[3] Alekh Agarwal, Sahand Negahban, and Martin right. Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. In *Advances in Neural Information Processing Systems*, pages 1538–1546, 2012.

[4] Alekh Agarwal, Sahand Negahban, and Martin Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, pages 2452–2482, 2012.

[5] Nadir Amaioua, Charles Audet, Andrew R Conn, and Sébastien Le Digabel. Efficient solution of quadratically constrained quadratic subproblems within the mesh adaptive direct search algorithm. *European Journal of Operational Research*, 268(1):13–24, 2018.

[6] Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In *Conference on Learning Theory*, pages 115–137, 2019.

[7] Francesco Archetti and Antonio Candelieri. *Bayesian optimization and data science.* Springer, 2019.

[8] Setareh Ariafar, Jaume Coll-Font, Dana H Brooks, and Jennifer G Dy. ADMMBO: Bayesian Optimization with Unknown Constraints using ADMM. *Journal of Machine Learning Research*, 20(123):1–26, 2019.

[9] Yossi Arjevani, Yair Carmon, John Duchi, Dylan Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint*

arXiv:1912.02365, 2019.

[10] Charles Audet and John E Dennis Jr. A pattern search filter method for nonlinear programming without derivatives. *SIAM Journal on Optimization*, 14(4):980–1010, 2004.

[11] Charles Audet and John E Dennis Jr. Mesh adaptive direct search algorithms for constrained optimization. *SIAM Journal on optimization*, 17(1):188–217, 2006.

[12] Charles Audet and John E Dennis Jr. A progressive barrier for derivative-free nonlinear programming. *SIAM Journal on Optimization*, 20(1):445–472, 2009.

[13] Charles Audet and Warren Hare. Derivative-free and blackbox optimization. 2017.

[14] Charles Audet, Sébastien Le Digabel, and Mathilde Peyrega. Linear equalities in blackbox optimization. *Computational Optimization and Applications*, 61(1):1–23, 2015.

[15] Charles Audet and Christophe Tribes. Mesh-based Nelder–Mead algorithm for inequality constrained optimization. *Computational Optimization and Applications*, 71(2):331–352, 2018.

[16] F Augustin and YM Marzouk. NOWPAC: A provably convergent derivative-free nonlinear optimizer with path-augmented constraints. *arXiv preprint arXiv:1403.1931*, 2014.

[17] François Bachoc, Céline Helbert, and Victor Picheny. Gaussian process optimization with failures: Classification and convergence proof. *Journal of Global Optimization*, 78(3):483–506, 2020.

[18] Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[19] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3459–3468, 2018.

[20] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order Nonconvex Stochastic Optimization: Handling Constraints, High-Dimensionality and Saddle-Points. *arXiv preprint arXiv:1809.06474v2*, 2019.

[21] Amir Beck. *First-order methods in optimization.* SIAM, 2017.

[22] Jose Blanchet, Donald Goldfarb, Garud Iyengar, Fengpei Li, and Chaoxu Zhou. Unbiased simulation for optimizing stochastic function compositions. *arXiv preprint arXiv:1711.07564*,

2017.

[23] Julius R Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744, 1954.

[24] Holger Boche, Robert Calderbank, Gitta Kutyniok, and Jan Vybíral. *Compressed sensing and its applications*. Springer, 2015.

[25] Digvijay Boob, Qi Deng, and Guanghui Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *arXiv preprint arXiv:1908.02734*, 4, 2019.

[26] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 537–546. JMLR. org, 2017.

[27] Vivek S Borkar. *Stochastic approximation: A dynamical systems viewpoint*, volume 48. Springer, 2009.

[28] Richard P Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

[29] Luis Felipe Bueno, Ana Friedlander, José Mario Martinez, and FNC Sobral. Inexact restoration method for derivative-free optimization with smooth constraints. *SIAM Journal on Optimization*, 23(2):1189–1213, 2013.

[30] Árpád Bűrmen, Janez Puhan, and Tadej Tuma. Grid restrained Nelder-Mead algorithm. *Computational optimization and applications*, 34(3):359–375, 2006.

[31] Emmanuel Candes, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[32] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. STAN: A Probabilistic Programming Language. *Journal of Statistical Software*, 76, 2017.

[33] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.

[34] Krzysztof Choromanski, Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, Deepali Jain, Yuxiang Yang, Atil Iscen, Jasmine Hsu, and Vikas Sindhwani. Provably robust blackbox

optimization for reinforcement learning. In *Conference on Robot Learning*, pages 683–696. PMLR, 2020.

[35] Andrew Conn, Katya Scheinberg, and Luis Vicente. *Introduction to Derivative-Free Optimization*, volume 8. SIAM, 2009.

[36] Andrew R Conn and Sébastien Le Digabel. Use of quadratic models with mesh-adaptive direct search for constrained black box optimization. *Optimization Methods and Software*, 28(1):139–158, 2013.

[37] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[38] Darinka Dentcheva, Spiridon Penev, and Andrzej Ruszczyński. Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, 69(4):737–760, 2017.

[39] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *Annals of Statistics*, 48(3):1348–1382, 2020.

[40] Sébastien Le Digabel and Stefan M Wild. A taxonomy of constraints in simulation-based optimization. *arXiv preprint arXiv:1505.07881*, 2015.

[41] Jürgen Dippon. Accelerated randomized stochastic optimization. *The Annals of Statistics*, 31(4):1260–1281, 2003.

[42] David Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[43] Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *Proceedings of the 35th International Conference on Machine Learning-Volume 119*, 2019.

[44] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[45] John Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

[46] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[47] Kwassi Joseph Dzahini, Michael Kokkolaras, and Sébastien Le Digabel. Constrained stochastic blackbox optimization using a progressive barrier and probabilistic estimates. *arXiv preprint arXiv:2011.04225*, 2020.

[48] N Echebest, María Laura Schuverdt, and Raúl Pedro Vignau. An inexact restoration derivative-free filter method for nonlinear programming. *Computational and Applied Mathematics*, 36(1):693–718, 2017.

[49] Yonina Eldar and Gitta Kutyniok. *Compressed sensing: Theory and applications*. Cambridge university press, 2012.

[50] David Eriksson and Matthias Poloczek. Scalable Constrained Bayesian Optimization. *arXiv preprint arXiv:2002.08526*, 2020.

[51] Yuri Ermoliev. Methods of stochastic programming. *Nauka, Moscow*, 1976.

[52] Yuri Ermoliev and Vladimir Norkin. Sample average approximation method for compound stochastic optimization problems. *SIAM Journal on Optimization*, 23(4):2231–2263, 2013.

[53] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal nonconvex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018.

[54] Giovanni Fasano, Giampaolo Liuzzi, Stefano Lucidi, and Francesco Rinaldi. A line-search based derivative-free approach for nonsmooth constrained optimization. *SIAM journal on optimization*, 24(3):959–992, 2014.

[55] Peter I Frazier. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[56] Wenbo Gao, Laura Graesser, Krzysztof Choromanski, Xingyou Song, Nevena Lazic, Pannag Sanketi, Vikas Sindhwani, and Navdeep Jaitly. Robotic table tennis with model-free reinforcement learning. *arXiv preprint arXiv:2003.14398*, 2020.

[57] Jacob Gardner, Matt Kusner, Kilian Weinberger, and John Cunningham. Bayesian optimization with inequality constraints. In *International Conference on Machine Learning*, pages 937–945, 2014.

[58] Michael A Gelbart, Ryan P Adams, Matthew W Hoffman, and Zoubin Ghahramani. A general framework for constrained bayesian optimization using information-based search. *Journal of Machine Learning Research*, 17(160):1–53, 2016.

[59] Michael A Gelbart, Jasper Snoek, and Ryan P Adams. Bayesian optimization with unknown constraints. In *30th Conference on Uncertainty in Artificial Intelligence, UAI 2014*, pages 250–259, 2014.

[60] Andrew Gelman and Donald B Rubin. A single series from the Gibbs sampler provide a false sense of security. *Bayesian Statistics*, 4, 1992.

[61] John Geweke. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. 1991.

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

[63] Saeed Ghadimi, Andrzej Ruszczynski, and Mengdi Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020.

[64] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[65] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D Sculley. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1487–1495, 2017.

[66] Robert B Gramacy, Genetha A Gray, Sébastien Le Digabel, Herbert Lee, Pritam Ranjan, Garth Wells, and Stefan M Wild. Modeling an augmented Lagrangian for blackbox constrained optimization. *Technometrics*, 58(1):1–11, 2016.

[67] Serge Gratton and Luís Nunes Vicente. A merit function approach for direct search. *Siam journal on optimization*, 24(4):1980–1998, 2014.

[68] Stewart Greenhill, Santu Rana, Sunil Gupta, Pratibha Vellanki, and Svetha Venkatesh. Bayesian optimization for adaptive experimental design: A review. *IEEE Access*, 8:13937–13948, 2020.

[69] Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained bayesian optimization for automatic chemical design. *arXiv preprint arXiv:1709.05501*, 2017.

[70] Paul Hand and Vladislav Voroninski. Global guarantees for enforcing deep generative priors by empirical risk. In *Conference On Learning Theory*, pages 970–978, 2018.

[71] José Miguel Hernández-Lobato, Michael Gelbart, Matthew Hoffman, Ryan Adams, and Zoubin Ghahramani. Predictive entropy search for Bayesian optimization with unknown constraints. In *International conference on machine learning*, pages 1699–1707. PMLR, 2015.

[72] Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[73] Robert Hooke and Terry A Jeeves. Direct search solution of numerical and statistical problems. *Journal of the ACM (JACM)*, 8(2):212–229, 1961.

[74] Kevin G Jamieson, Robert D Nowak, and Benjamin Recht. Query complexity of derivative-free optimization. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pages 2672–2680, 2012.

[75] Noémie Jaquier and Leonel Rozo. High-Dimensional Bayesian Optimization via Nested Riemannian Manifolds. *Advances in Neural Information Processing Systems*, 33, 2020.

[76] Noémie Jaquier, Leonel Rozo, Sylvain Calinon, and Mathias Bürger. Bayesian optimization meets Riemannian manifolds in robot learning. In *Conference on Robot Learning*, pages 233–246. PMLR, 2020.

[77] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[78] Tamara G Kolda, Robert Michael Lewis, and Virginia Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM review*, 45(3):385–482, 2003.

[79] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[80] Remi Lam and Karen Willcox. Lookahead Bayesian optimization with inequality constraints. In *Advances in Neural Information Processing Systems*, pages 1890–1900, 2017.

[81] Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

[82] Ben Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*, volume 39. Springer, 2015.

[83] Benjamin Letham, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis*, 14(2):495–519, 2019.

[84] Robert Michael Lewis and Virginia Torczon. A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Optimization*, 12(4):1075–1089, 2002.

[85] Jiaxiang Li, Krishnakumar Balasubramanian, and Shiqian Ma. Stochastic zeroth-order Riemannian derivative estimation and optimization. *arXiv preprint arXiv:2003.11238*, 2020.

[86] Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O Hero III, and Pramod K Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.

[87] Giampaolo Liuzzi and Stefano Lucidi. A derivative-free algorithm for inequality constrained nonlinear programming via smoothing of an ℓ_∞ penalty function. *SIAM Journal on Optimization*, 20(1):1–29, 2009.

[88] Giampaolo Liuzzi, Stefano Lucidi, and Marco Sciandrone. Sequential penalty derivative-free methods for nonlinear constrained optimization. *SIAM Journal on Optimization*, 20(5):2614–2635, 2010.

[89] Nimalan Mahendran, Ziyu Wang, Firas Hamze, and Nando De Freitas. Adaptive MCMC with Bayesian optimization. In *Artificial Intelligence and Statistics*, pages 751–760. PMLR, 2012.

[90] Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search provides a competitive approach to reinforcement learning. *arXiv preprint arXiv:1803.07055*, 2018.

[91] JTMT Meer, HV Duijne, R Nieuwenhuis, and HHM Rijnaarts. Prevention and reduction of pollution of groundwater at contaminated megasites: integrated management strategy, and its application on megasite cases. *Groundwater Science and Policy: An International Overview*, pages 405–420, 2008.

[92] Jonas Mockus. Application of Bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, 4(4):347–365, 1994.

[93] Jonas Mockus. *Bayesian approach to global optimization: Theory and applications*, volume 37. Springer Science & Business Media, 2012.

[94] Abdelkader Mokkadem and Mariane Pelletier. A companion for the Kiefer–Wolfowitz–Blum stochastic approximation algorithm. *The Annals of Statistics*, 35(4):1749–1772, 2007.

[95] Juliane Müller and Joshua D Woodbury. GOSAC: global optimization with surrogate approximation of constraints. *Journal of Global Optimization*, 69(1):117–136, 2017.

[96] Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

[97] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[98] Arkadij Semenovič Nemirovski and David Borisovich Yudin. Problem complexity and method efficiency in optimization. *Wiley-Interscience*, 1983.

[99] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[100] Gregory Ongie, Ajil Jalal, Christopher Metzler Richard Baraniuk, Alexandros Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 2020.

[101] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[102] Victor Picheny, Robert B Gramacy, Stefan Wild, and Sebastien Le Digabel. Bayesian optimization under mixed constraints with a slack-variable augmented Lagrangian. In *Advances in neural information processing systems*, pages 1435–1443, 2016.

[103] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[104] Tony Pourmohamad and Herbert Lee. The statistical filter approach to constrained optimization. *Technometrics*, 62(3):303–312, 2020.

[105] Michael JD Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162, 1964.

[106] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[107] Yangjun Ruan, Yuanhao Xiong, Sashank Reddi, Sanjiv Kumar, and Cho-Jui Hsieh. Learning to learn by zeroth-order oracle. *arXiv preprint arXiv:1910.09464*, 2019.

[108] David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[109] Andrzej Ruszczyński. A linearization method for nonsmooth stochastic programming problems. *Mathematics of Operations Research*, 12(1):32–49, 1987.

[110] Andrzej Ruszczyński. A stochastic subgradient method for nonsmooth nonconvex multi-level composition optimization. *arXiv preprint, arXiv:2001.10669*, 2020.

[111] Anit Kumar Sahu, Manzil Zaheer, and Soummya Kar. Towards gradient free and projection free stochastic optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3468–3477. PMLR, 2019.

[112] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[113] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[114] Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24. PMLR, 2013.

[115] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory.* SIAM, 2014.

[116] Jasper Snoek, Hugo Larochelle, and Ryan Prescott Adams. Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 2012.

[117] Benjamin Solnik, Daniel Golovin, Greg Kochanski, John Elliot Karro, Subhodeep Moitra, and D Sculley. Bayesian optimization for a better dessert. *Proceedings of the 2017 NIPS Workshop on Bayesian Optimization*, 2017.

[118] Ganlin Song, Zhou Fan, and John Lafferty. Surfing: Iterative optimization over incrementally trained deep networks. In *Advances in Neural Information Processing Systems*, pages 15034–15043, 2019.

[119] James C Spall. A stochastic approximation technique for generating maximum likelihood parameter estimates. In *1987 American control conference*, pages 1161–1167. IEEE, 1987.

[120] James C Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, volume 65. John Wiley & Sons, 2005.

[121] WGRFR Spendley, George R Hext, and Francis R Himsworth. Sequential application of simplex designs in optimisation and evolutionary operation. *Technometrics*, 4(4):441–461, 1962.

[122] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

[123] Anke Tröltzsch. A sequential quadratic programming algorithm for equality-constrained optimization without derivatives. *Optimization Letters*, 10(2):383–399, 2016.

[124] Mathukumalli Vidyasagar. *An introduction to compressed sensing*. SIAM, 2019.

[125] Mengdi Wang, Ethan Fang, and Han Liu. Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

[126] Mengdi Wang, Ji Liu, and Ethan Fang. Accelerating stochastic composition optimization. In *Advances in Neural Information Processing Systems*, pages 1714–1722, 2016.

[127] Ziyu Wang, Shakir Mohamed, and Nando Freitas. Adaptive Hamiltonian and Riemann Manifold Monte Carlo. In *International conference on machine learning*, pages 1462–1470. PMLR,

2013.

[128] Xiaohan Wei, Zhuoran Yang, and Zhaoran Wang. On the statistical rate of nonlinear recovery in generative models with heavy-tailed data. In *International Conference on Machine Learning*, pages 6697–6706, 2019.

[129] Shuoguang Yang, Mengdi Wang, and Ethan Fang. Multilevel stochastic gradient methods for nested composition optimization. *SIAM Journal on Optimization*, 29(1):616–659, 2019.

[130] Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat Erdogdu. An analysis of constant step size sgd in the non-convex regime: Asymptotic normality and bias. *arXiv preprint arXiv:2006.07904*, 2020.

[131] Junyu Zhang and Lin Xiao. Multi-level composite stochastic optimization via nested variance reduction. *arXiv preprint arXiv:1908.11468*, 2019.