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

Krishnakumar Balasubramanian† Saeed Ghadimi‡ Anthony Nguyen§

August 25, 2020

Abstract

In this paper, we study smooth stochastic multi-level composition optimization problems, where the objective function is a nested composition of T functions. We assume access to noisy evaluations of the functions and their gradients, through a stochastic first-order oracle. For solving this class of problems, we propose two algorithms using moving-average stochastic estimates, and analyze their convergence to an ϵ-stationary point of the problem. We show that the first algorithm, which is a generalization of [22] to the T level case, can achieve a sample complexity of $\mathcal{O}(1/\epsilon^6)$ by using mini-batches of samples in each iteration. By modifying this algorithm using linearized stochastic estimates of the function values, we improve the sample complexity to $\mathcal{O}(1/\epsilon^4)$. This modification also removes the requirement of having a mini-batch of samples in each iteration. To the best of our knowledge, this is the first time that such an online algorithm designed for the (un)constrained multi-level setting, obtains the same sample complexity of the smooth single-level setting, under mild assumptions on the stochastic first-order oracle.

1 Introduction

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

$$\min_{x\in X}\Big\{F(x) = f_1 \circ \cdots \circ f_T(x)\Big\}, \tag{1}$$

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 [8, 20, 21, 24, 27, 31], for a partial list. In this work, we consider stochastic first-order algorithms for solving (1) when $T \geq 1$. Note that the gradient of the function $F(x)$ in (1), 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. Because of the nested nature of the

*Authors listed by alphabetical order.

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

‡Department of Management Sciences, University of Waterloo. sghadimi@uwaterloo.ca.

§Department of Mathematics, University of California, Davis. anthonynguyen@math.ucdavis.edu.

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 (1) have been considered since the work of [17], 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 [5, 7, 11, 18, 22, 25, 34, 35, 37, 39] for such applications and various algorithmic approaches for solving problem (1). In particular [34] and [37] 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 [18] and [11], respectively, show that in the sample-average or empirical risk minimization setting, the argmin of the problem in (1) 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 (1) with level-independent convergence rates?* Recently, for the case of $T = 2$, [22] 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) with level-independent convergence rates in the stochastic first-order oracle setting, under mild assumptions. Our complexity results are summarized in Table 1. The first algorithm is based on an extension of the NASA algorithm from [22] (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 [10, 15, 29, 30], 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 [37] is provided in Table 1. We remark that the approach and the results in [37] are provided only for the unconstrained setting. We also highlight the related work of [39] 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 (1) 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 (1). The algorithm proposed in [39] is a proximal variant of SPIDER variance reduction technique [19] 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, [39] 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,

Method	Convergence Rate	Oracle Complexity
[37]	$\mathcal{O}\left(N^{-4/(7+T)}\right)$	$\mathcal{O}\left(1/\epsilon^{(7+T)/2}\right)$
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: Convergence rates and Oracle complexity results for finding an ϵ-pair $\bar{x}, \bar{z}$ of (1); see Definition 2.1 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 1 and Remark 3 for more details.

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 [30] 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 (16) is different from that used in [30], which makes an important part of our convergence analysis distinct. This enables us, unlike [30], to relax the boundedness assumption of the feasible set thereby making our method applicable to the unconstrained problems as well.

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

$$\check{x} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \mathbb{E}(y - a^\top x)^2,$$

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^* [26, 28]. In compressed sensing [9, 13], 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,

$$\bar{x} = \operatorname*{argmin}_{x \in \mathbb{R}^d : \|x\|_0 \leq k} \mathbb{E}(y - a^\top x)^2,$$

it is enough to require $N \approx k \log d$ (as opposed to $N \approx d$) samples for accurate reconstruction [1, 2]. 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 [6, 16, 33].

Recently, motivated by the success of deep learning, [7] 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,

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

was proposed in [7]; see also [23, 25, 36] for more details. Furthermore, the mapping G is assumed to be deep neural network with depth T'. That is, $G(z) = f_1 \circ f_2 \cdots, f_T'(z)$, 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

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

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). The statistical sample complexity of the above problem, for accurate reconstruction, requires the number of measurement to be of the order of k [7]. However, efficient algorithms for solving the above problem are less explored; see [23, 32] 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 paper, 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 paper is organized as follows. In Section 2, we present our first algorithm and analyze its convergence analysis for solving (1) with any $T \ge 1$. In Section 2, 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 4.

2 Multi-level Nested Averaging Stochastic Gradient Method

In this section, we present our first algorithm for solving problem (1). As mentioned in Section 1, the previously proposed stochastic gradient-type methods suffer in terms of the convergence rates when applied for solving this problem [37]. The main reason is the increased bias when estimating the stochastic gradient of F, for $T \ge 2$. 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 [22] 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 [37]. 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), 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

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 $k = 0, 1, 2, \ldots,$ **do**

1. Compute

$$u^k = \operatorname*{argmin}_{y\in X} \left\{ \langle z^k, y - x^k \rangle + \frac{\beta_k}{2}\|y - x^k\|^2 \right\}, \tag{2}$$

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

2. Set

$$x^{k+1} = (1 - \tau_k)x^k + \tau_k u^k, \tag{3}$$

$$z^{k+1} = (1 - \tau_k)z^k + \tau_k \prod_{i=1}^{T} J_{T+1-i}^{k+1}, \tag{4}$$

$$w_i^{k+1} = (1 - \tau_k)w_i^k + \tau_k \bar{G}_i^{k+1}, \qquad 1 \leq i \leq T, \tag{5}$$

where

$$\bar{G}_i^{k+1} = \frac{1}{b_k}\sum_{j=1}^{b_k} G_{i,j}^{k+1}. \tag{6}$$

end for

Output:

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 [22]. 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,

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

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

Assumption 2.1. *All functions $f_1, \ldots, f_T$ and their derivatives are Lipschitz continuous with Lipschitz constants L_{f_i} and $L_{\nabla f_i}$, respectively.*

Assumption 2.2. *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., $\{G_i^{k+1}\}_{i=1,\ldots,T}$ are independent and so are $\{J_i^{k+1}\}_{i=1,\ldots,T}$.*

Assumption 2.1 is a standard smoothness assumption made in the literature on nonlinear optimization. Similarly, Parts 1 and 2 in Assumption 2.2 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 [39] are stronger than our assumptions above, as the 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 [37]. In the next couple of technical results, we provide some properties of composite functions that are required for our subsequent results.

Lemma 2.1. *Define $F_i(x) = f_i \circ f_{i+1} \circ \cdots f_T(x)$. Under Assumption 2.1, 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 2.1, 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

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

Now, suppose that gradient of F_{i+1} is Lipschitz continuous for any $i \le T-1$. 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 [39], Lemma 5.2., with a slightly different proof.

Lemma 2.2. *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 $x \in X, w_j \in \mathbb{R}^{d_j}$ $\ j = i+1, \ldots, T$. Then under Assumption 2.1, we have*

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

Proof. We show the result by backward induction. The case $i = T$ is trivial. When $i = T-1$, under Assumption 2.1, we have

$$\|\nabla F_{T-1}(x) - \nabla f_T(x) \nabla f_{T-1}(w_T)\| = \|\nabla f_T(x)[\nabla f_{T-1}(f_T(x)) - \nabla f_{T-1}(w_T)]\| \le L_{\nabla f_{T-1}} L_{f_T} \|f_T(x) - w_T\|.$$

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

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

We then have

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

■

Lemma 2.3. *Under Assumption 2.1, for any $j \in \{1, \ldots, T-1\}$, we have*

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

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

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

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

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

where the third inequality follows by induction hypothesis. ■

Lemma 2.4. *Define*

$$\begin{aligned}
&R_1 = L_{\nabla f_1} L_{f_2} \cdots L_{f_T} \\
&R_j = L_{f_1} \ldots L_{f_{j-1}} L_{\nabla f_j} L_{f_{j+1}} \cdots L_{f_T} \quad 1 < j \le T-1 \\
&C_2 = R_1
\end{aligned}$$

$$C_j = R_1 L_{f_2\circ\cdots\circ f_{j-1}} + R_2 L_{f_3\circ\cdots\circ f_{j-1}} + \cdots + R_{j-2} L_{f_{j-1}} + R_{j-1} \text{ with } 2 < j \le T.$$

Assume that Assumption 2.1 holds. Then for $T \ge 3$,

$$\left\| \nabla F(x) - \nabla f_T(x) \prod_{i=2}^{T} \nabla f_{T+1-i}(w_{T+2-i}) \right\| \le \sum_{j=2}^{T-1} C_j \|f_j(w_{j+1}) - w_j\| + C_T \|f_T(x) - w_T\| \tag{7}$$

Proof. By Lemma 2.2 and Lemma 2.3, we have

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

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

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

The conclusion follows. To see this, term collecting $\|f_2(w_3) - w_2\|$, we have C_2. For $2 < j \le T$, 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). One can see [22] for a simple proof.

Lemma 2.5. *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 2.6. *Let $\{x^k\}_{k\ge0}$ and $\{w_i^k\}_{k\ge0}$ $1 \le i \le T$ be generated by Algorithm 1. Denote*

$$d^k = u^k - x^k, \qquad w_{T+1}^k \equiv x^k \quad \forall k \ge 0, \qquad A_{k,i} = f_i(w_{i+1}^{k+1}) - f_i(w_{i+1}^k) \quad 1 \le i \le T. \tag{8}$$

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

$$\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 \le (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \frac{1}{\tau_k}\|A_{k,i}\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + r_i^{k+1}, \tag{9}$$

$$\|w_i^{k+1} - w_i^k\|^2 \le \tau_k^2 \left[\|f_i(w_{i+1}^k) - w_i^k\|^2 + \|e_i^{k+1}\|^2 - 2\langle e_i^{k+1}, f_i(w_{i+1}^k) - w_i^k \rangle \right], \tag{10}$$

where

$$r_i^{k+1} = 2\tau_k \langle e_i^{k+1}, A_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k) \rangle, \qquad e_i^{k+1} = f_i(w_{i+1}^k) - \bar{G}_i^{k+1}. \tag{11}$$

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

$$\|f_T(x^{k+1}) - w_T^{k+1}\|^2 \leq (1-\tau_k)\|f_T(x^k) - w_T^k\|^2 + L_{f_T}\tau_k\|d^k\|^2 + \tau_k^2\|e_T^{k+1}\|^2 + r_T^{k+1}, \tag{12}$$

$$\begin{aligned}\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 &\leq (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + L_{f_i}^2\tau_k\left[\|f_{i+1}(w_{i+2}^k) - w_{i+1}^k\|^2 + \|e_{i+1}^{k+1}\|^2\right] \\ &\quad + \tau_k^2\|e_i^{k+1}\|^2 + \bar{r}_i^{k+1} \qquad\qquad 1 \leq i \leq T-1,\end{aligned} \tag{13}$$

where

$$\bar{r}_i^{k+1} = -2\tau_k L_{f_i}^2\langle e_{i+1}^{k+1}, f_{i+1}(w_{i+2}^k) - w_{i+1}^k\rangle + r_i^{k+1}. \tag{14}$$

Proof. Noting (5), (9), and (11), 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 (11), (9) follows by noting that

$$\begin{aligned}\|A_{k,i} + (1-b\tau_k)(f_i(w_{i+1}^k) - w_i^k)\|^2 &= \|A_{k,i}\|^2 + (1-\tau_k)^2\|f_i(w_{i+1}^k) - w_i^k\|^2 + 2(1-\tau_k)\langle A_{k,i}, f_i(w_{i+1}^k) - w_i^k\rangle \\ &\leq \|A_{k,i}\|^2 + (1-\tau_k)^2\|f_i(w_{i+1}^k) - w_i^k\|^2 + \left(\frac{1}{\tau_k} - 1\right)\|A_{k,i}\|^2 \\ &\quad + (1-\tau_k)\tau_k\|f_i(w_{i+1}^k) - w_i^k\|^2 \\ &= (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \frac{1}{\tau_k}\|A_{k,i}\|^2,\end{aligned} \tag{15}$$

due to Cauchy Schwartz and Young's inequalities. Also, (10) directly follows from (5) 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 (3), (8), and Lipschitz continuity of f_i, we have

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

The results then follows by noting (9) and (10).

■

We remark that the mini-batch sampling in (6) 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 (13). Moreover, we do not need this inequality for $i=1$ when establishing the convergence rate of Algorithm 1. Thus, when $T \leq 2$, this algorithm convergences without using mini-batch of samples in each iteration, as shown in [22].

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

$$W(x,z,w) = F(x) - F^* - \eta(x,z) + \sum_{i=1}^{T-1}\gamma_i\|f_i(w_{i+1}) - w_i\|^2 + \gamma_T\|f_T(x) - w_T\|^2 \tag{16}$$

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

Lemma 2.7. *Suppose that* $\{x^k, z^k, u^k, w_1^k, \ldots, w_T^k\}_{k\geq 0}$ *are generated by Algorithm 1 and Assumption 2.1 holds.*

a) If

$$\begin{aligned}\gamma_0 := 0, \qquad \gamma_1, \lambda > 0, \qquad \beta_k \equiv \beta \geq \lambda + \gamma_T,\\ \gamma_j - \gamma_{j-1}L_{f_{j-1}}^2 - \lambda > 0, \qquad 4(\beta - \lambda - \gamma_T)(\gamma_j - \gamma_{j-1}L_{f_{j-1}}^2 - \lambda) \geq TC_j^2 \qquad j = 2, \ldots, T,\end{aligned} \tag{17}$$

where C_j*'s are defined in Lemma 2.4, we have*

$$\lambda \sum_{k=0}^{N-1} \tau_k \left[\|d^k\|^2 + \sum_{i=1}^{T-1} \|f_i(w_{i+1}^k) - w_i^k\|^2 + \|f_T(x^k) - w_T^k\|^2\right] \leq W(x^0, z^0, w^0) + \sum_{k=0}^{N-1} R^{k+1}, \tag{18}$$

where

$$\begin{aligned}R^{k+1} &:= \tau_k^2 \sum_{i=1}^{T} \gamma_i \|e_i^{k+1}\|^2 + \tau_k \sum_{i=1}^{T-1} \gamma_i L_{f_i}^2 \|e_{i+1}^{k+1}\|^2 + \sum_{i=1}^{T-1} \gamma_i \bar{r}_i^{k+1} + \gamma_T r_T^{k+1} + \tau_k \langle d^k, \Delta^k\rangle,\\ &\quad + \frac{(L_{\nabla F} + L_{\nabla \eta})\tau_k^2}{2}\|d^k\|^2 + \frac{L_{\nabla \eta}}{2}\|z^{k+1} - z^k\|^2,\end{aligned} \tag{19}$$

$$\Delta^k := \nabla f_T(x^k) \prod_{i=2}^{T} \nabla f_{T+1-i}(w_{T+2-i}^k) - \prod_{i=1}^{T} J_{T-i+1}^{k+1}, \tag{20}$$

and $r_i^{k+1}, \bar{r}_i^{k+1}$ *are defined in (11) and (14), respectively.*

b) If parameters are chosen as

$$\begin{aligned}&\gamma_0 = 0, \qquad \gamma_1 = 1, \qquad \gamma_j := 2^{j-1}(L_{f_1} \cdots L_{f_{j-1}})^2 \quad 2 \leq j \leq T,\\ &\lambda = \frac{1}{2}\min_{1\leq i\leq T}(\gamma_i - \gamma_{i-1}L_{f_{i-1}}^2), \qquad \beta \geq \lambda + \gamma_T + \frac{T\max_{2\leq i\leq T} C_i^2}{4\lambda}.\end{aligned} \tag{21}$$

Then, conditions in (17) are satisfied.

Proof. First, note that by Lemma 2.1, we have

$$F(x^{k+1}) \leq F(x^k) + \langle \nabla F(x^k), x^{k+1} - x^k\rangle + \frac{L_{\nabla F}}{2}\|x^{k+1} - x^k\|^2 = F(x^k) + \tau_k\langle \nabla F(x^k), d^k\rangle + \frac{L_{\nabla F}\tau_k^2}{2}\|d^k\|^2. \tag{22}$$

Second, note that by the optimality condition of (2), we have

$$\langle z^k + \beta_k(u^k - x^k), x^k - u^k\rangle \geq 0, \qquad \langle z^k, d^k\rangle + \beta_k\|d^k\|^2 \leq 0. \tag{23}$$

Then, noting (3), (4), and in the view of Lemma 2.5, we obtain

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

$$\leq -\beta_k \tau_k \|d^k\|^2 - \tau_k \langle d^k, \prod_{i=1}^{T} J_{T-i+1}^{k+1} \rangle + \frac{L_{\nabla \eta}}{2} \left[\tau_k^2 \|d^k\|^2 + \|z^{k+1} - z^k\|^2 \right]. \tag{24}$$

Third, noting Lemma 2.6.b), we have

$$\sum_{i=1}^{T-1} \gamma_i \left[\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 - \|f_i(w_{i+1}^k) - w_i^k\|^2 \right] + \gamma_T \left[\|f_T(x^{k+1}) - w_T^{k+1}\|^2 - \|f_T(x^k) - w_T^k\|^2 \right]$$
$$\leq \sum_{i=1}^{T-1} \gamma_i \left\{ -\tau_k \left[\|f_i(w_{i+1}^k) - w_i^k\|^2 - L_{f_i}^2 \|f_{i+1}(w_{i+2}^k) - w_{i+1}^k\|^2 - L_{f_i}^2 \|e_{i+1}^{k+1}\|^2 \right] + \tau_k^2 \|e_i^{k+1}\|^2 + \bar{r}_i^{k+1} \right\}$$
$$+ \gamma_T \left\{ -\tau_k \left[\|f_T(x^k) - w_T^k\|^2 - L_{f_T}^2 \|d^k\|^2 \right] + \tau_k^2 \|e_T^{k+1}\|^2 + r_T^{k+1} \right\}$$
$$= -\tau_k \left\{ \gamma_1 \|f_1(w_2^k) - w_1^k\|^2 + \sum_{j=2}^{T-1} [\gamma_j - \gamma_{j-1} L_{f_{j-1}}^2] \|f_j(w_{j+1}^k) - w_j^k\|^2 + [\gamma_T - \gamma_{T-1} L_{f_{T-1}}^2] \|f_T(x^k) - w_T^k\|^2 \right\}$$
$$+ \tau_k \left[\sum_{i=1}^{T-1} \gamma_i L_{f_i}^2 \|e_{i+1}^{k+1}\|^2 + \gamma_T \|d^k\|^2 \right] + \tau_k^2 \sum_{i=1}^{T} \gamma_i \|e_i^{k+1}\|^2 + \sum_{i=1}^{T-1} \gamma_i \bar{r}_i^{k+1} + \gamma_T r_T^{k+1}. \tag{25}$$

Combining the above relation with (24), (22), noting definition of merit function in (16), and in the view of Lemma 2.4, we obtain

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

where R^{k+1} is defined in (19). Thus, if (17) holds, we have

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

Summing up the above inequalities and re-arranging the terms, we obtain (18). It can be easily verified that condition (17) is satisfied by the choice of parameters in (21).

■

We introduce the following additional lemmas.

Lemma 2.8. *Consider a sequence $\{\tau_k\}_{k\geq 0} \in (0, 1]$, and define*

$$\Gamma_k = \Gamma_1 \prod_{i=1}^{k-1} (1 - \tau_i) \qquad k \geq 2, \qquad \Gamma_1 = \begin{cases} 1 & \text{if } \tau_0 = 1, \\ 1 - \tau_0 & \text{otherwise.} \end{cases} \tag{26}$$

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

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

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

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

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 (26), 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. *Suppose that Assumption 2.2 holds and (for simplicity) $\tau_0 = 1$, $\beta_k = \beta > 0$ for all k. Then, for any $k \ge 1$, we have*

$$\beta^2 \mathbb{E}[\|d^k\|^2 | \mathscr{F}_k] \le \mathbb{E}[\|z^k\|^2 | \mathscr{F}_k] \le \prod_{i=1}^{T} \sigma_{J_i}^2, \tag{27}$$

$$\mathbb{E}[\|z^{k+1} - z^k\|^2 | \mathscr{F}_k] \le 4\tau_k^2 \prod_{i=1}^{T} \sigma_{J_i}^2. \tag{28}$$

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

$$b_k = \left\lceil \frac{\max_{1 \le i \le T} L_{f_i}^2}{\tau_k} \right\rceil \qquad k \ge 0, \tag{29}$$

we have

$$\mathbb{E}[R^{k+1} | \mathscr{F}_k] \le \tau_k^2 \left[\frac{1}{2} \left(\prod_{i=1}^{T} \sigma_{J_i}^2 \right) \left(\frac{L_{\nabla F} + (1+4\beta^2) L_{\nabla \eta}}{\beta^2} \right) + \sum_{i=1}^{T} \gamma_i \sigma_{G_i}^2 \right] := \tau_k^2 \sigma^2, \tag{30}$$

where R^{k+1} is defined in (19).

Proof. The first inequality in (27) directly follows by (23) and Cauchy-Schwarz inequality. Noting (4), the fact that $\tau_0 = 1$, and in the view of Lemma 2.8, 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

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

Noting (27), we have

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

Now, observe that by (11), (14), the choice of b_k in (29), and under Assumption 2.2, 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 (19), (27), (28), and the above observation, we obtain (30). ∎

Observe that Lemma 2.7 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. This is the main step in establishing the convergence of Algorithm 1. Indeed, $\bar{x}\in X$ is a stationary point of (1), if $u=\bar{x}$ and $\bar{z}=\nabla F(\bar{x})$, where

$$u = \operatorname*{argmin}_{y\in X} \left\{\langle \bar{z}, y-\bar{x}\rangle + \frac{1}{2}\|y-\bar{x}\|^2\right\}. \tag{31}$$

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

Definition 2.1. *A pair of $(\bar{x},\bar{z})$ generated by Algorithm 1 is called an ϵ-stationary pair, if $\mathbb{E}[\sqrt{V(\bar{x},\bar{z})}]\le\epsilon$, where*

$$V(x,z) = \|u-x\|^2 + \|z-\nabla F(x)\|^2, \tag{32}$$

and u is the solution to (31).

When $X=\mathbb{R}^{d_T}$, $V(x,z)$ provides an upper bound for the $\|\nabla F(x)\|^2$. One can see [22] for the relation between $V(\bar{x},\bar{z})$ and other common gradient-based termination criteria such as gradient mapping. Furthermore, as shown in [22], we have

$$V(x^k,z^k) = \max(1,\beta_k^2)\|u^k-x^k\|^2 + \|z^k-\nabla F(x^k)\|^2, \tag{33}$$

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. *Suppose that $\{x^k,z^k\}_{k\ge0}$ are generated by Algorithm 1, Assumption 2.1 and Assumption 2.2 holds. Also assume that the parameters satisfy (21) and step sizes $\{\tau_k\}$ are chosen such that*

$$\sum_{i=k+1}^{N}\tau_i\Gamma_i \le c\Gamma_{k+1} \quad \forall k\ge 0 \text{ and } \forall N\ge 1, c \text{ is a positive constant.} \tag{34}$$

(a) *For every $N\ge1$, we have*

$$\sum_{k=1}^{N}\tau_k\mathbb{E}[\|\nabla F(x^k)-z^k\|^2|\mathscr{F}_k] \le \mathcal{B}_1(\sigma^2), \tag{35}$$

where

$$\mathcal{B}_1(\sigma^2, N) = \frac{4cL^2(T-1)}{\lambda}\left[W(x^0, z^0, w^0) + \sigma^2 \sum_{k=0}^{N-1} \tau_k^2\right] + c\prod_{\ell=1}^{T} \sigma_{J_\ell}^2 \sum_{k=0}^{N-1} \tau_k^2, \tag{36}$$

σ^2 is defined in (30) and

$$L^2 = \max\left\{L_{\nabla F}^2, \max_{1\le i\le T} C_j^2\right\}. \tag{37}$$

(b) *As a consequence, we have*

$$\mathbb{E}[V(x^R, z^R)] \le \frac{1}{\sum_{k=1}^{N}\tau_k}\left\{\mathcal{B}_1(\sigma^2, N) + \frac{\max(1,\beta^2)}{\lambda}\left[W(x^0, z^0, w^0) + \sigma^2\sum_{k=0}^{N}\tau_k^2\right]\right\}, \tag{38}$$

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

$$\tau_0 = 1, \quad \tau_k = \frac{1}{\sqrt{N}} \quad \forall k = 1, \ldots, N, \tag{39}$$

we have

$$\mathbb{E}[\|\nabla F(x^R) - z^R\|^2] \le \frac{1}{\sqrt{N}}\left[\frac{4L^2(T-1)\left[W(x^0, z^0, w^0) + 2\sigma^2\right]}{\lambda} + 2\prod_{\ell=1}^{T}\sigma_{J_\ell}^2\right] := \frac{\mathcal{B}_2(\sigma^2, N)}{\sqrt{N}}, \tag{40}$$

$$\mathbb{E}[V(x^R, z^R)] \le \frac{1}{\sqrt{N}}\left[\mathcal{B}_2(\sigma^2, N) + \frac{\max(1,\beta^2)}{\lambda}\left[W(x^0, z^0, w^0) + 2\sigma^2\right]\right], \tag{41}$$

$$\mathbb{E}[\|f_i(w_{i+1}^R) - w_i^R\|^2] \le \frac{1}{\lambda\sqrt{N}}\left[W(x^0, z^0, w^0) + 2\sigma^2\right] \qquad i = 1, \ldots, T. \tag{42}$$

Proof. We first show part (a). Noting (4), 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 (19) and

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

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

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

where the inequality follows from convexity of $\|\cdot\|^2$ and Lipschitz continuity of gradient of F. Thus, in the view of Lemma 2.8, 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

$$\begin{aligned}
\sum_{k=1}^{N} \tau_k \|\nabla F(x^k) - z^k\|^2 &= 2\sum_{k=1}^{N} \tau_k \Gamma_k \sum_{i=0}^{k-1} \frac{\tau_i}{\Gamma_{i+1}} \left(\|\delta^i\|^2 + L^2_{\nabla F}\|d^i\|^2 + \bar{\Delta}_i + \frac{\tau_i}{2}\|\Delta^i\|^2\right) \\
&= 2\sum_{k=0}^{N-1} \frac{\tau_k}{\Gamma_{k+1}} \left(\sum_{i=k+1}^{N} \tau_i \Gamma_i\right) \left(\|\delta^k\|^2 + L^2_{\nabla F}\|d^k\|^2 + \bar{\Delta}_k + \frac{\tau_k}{2}\|\Delta^k\|^2\right) \\
&\leq 2c \sum_{k=0}^{N-1} \tau_k \left(\|\delta^k\|^2 + L^2_{\nabla F}\|d^k\|^2 + \bar{\Delta}_k + \frac{\tau_k}{2}\|\Delta^k\|^2\right), \qquad (43)
\end{aligned}$$

where the last inequality follows from (34).

Now, observe that under Assumption 2.2, we have

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

Moreover, by Lemma 2.4 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

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

Combining the above observations with (44) and in the view of (37), we obtain

$$\begin{aligned}
\sum_{k=1}^{N} \tau_k \mathbb{E}[\|\nabla F(x^k) - z^k\|^2 | \mathscr{F}_k] \leq 4cL(T-1) \sum_{k=0}^{N-1} \tau_k &\left(\sum_{j=2}^{T-1} \|f_j(w_{j+1}) - w_j\|^2 + \|f_T(x) - w_T\|^2 + \|d^k\|^2\right) \\
&+ c \prod_{\ell=1}^{T} \sigma^2_{J_\ell} \sum_{k=0}^{N-1} \tau_k^2. \qquad (44)
\end{aligned}$$

Then, (35) follows from the above inequality, (18), and (30).

Part (b) then follows from part (a), (33), (18), 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 (39) implies that

$$\sum_{k=1}^{N} \tau_k \geq \sqrt{N}, \quad \sum_{k=0}^{N} \tau_k^2 = 2, \quad \Gamma_k = \left(1 - \frac{1}{\sqrt{N}}\right)^{k-1},$$

$$\sum_{i=k+1}^{N} \tau_i \Gamma_i = \left(1 - \frac{1}{\sqrt{N}}\right)^k \frac{1}{\sqrt{N}} \sum_{i=0}^{N-k-1} \left(1 - \frac{1}{\sqrt{N}}\right)^i \leq \left(1 - \frac{1}{\sqrt{N}}\right)^k,$$

ensuring condition (34) with $c = 1$.

■

Remark 1. *The result in (41) implies that to find an ϵ-stationary point of (1) (see, Definition 2.1), 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 (29) and (39). This bound is much better than $\mathcal{O}\left(1/\epsilon^{(7+T)/2}\right)$ obtained in [37] when $T > 4$[1]. In particular, it exhibits the level-independent behavior as discussed in Section 1. Note that, we obtain constants of order ρ^T, for example, when $\sigma^2_{J_i}$ in (30) are all of equal. We emphasize that [37] and [39] also have such constant factors that depend exponentially on T, in their proofs and the final results.

Remark 2. *The bound in (42) 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, (40) 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 [37], it is still worse than $\mathcal{O}\left(1/\epsilon^4\right)$ obtained in [22] 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.

3 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) 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 (5) 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 (45). 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 [29] in other contexts. Furthermore, it was also used in [10, 15] and [30] recently for the two-level and multi-level cases respectively.

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 [37].

Assumption 3.1. *Denote $w^k_{T+1} \equiv x^k$. Instantiate the conditions in Assumption 2.2. In addition to that, the stochastic oracle satisfies, for $1 \leq i \leq T$,*

[1]Following the presentation in [37], we only present the ϵ-related T dependence for their result.

Algorithm 2 Multi-level Nested Linearized Averaging Stochastic Gradient Method

Set $b_k = 1$ in Algorithm 1 and replace (5) with

$$w_i^{k+1} = (1-\tau_k)w_i^k + \tau_k G_i^{k+1} + J_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k), \qquad 1 \le i \le T. \tag{45}$$

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

The above assumptions is 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 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 3.1. *Let $\{x^k\}_{k\ge 0}$ and $\{w_i^k\}_{k\ge 0}$ $1 \le i \le T$ be generated by Algorithm 2. Define, for $1 \le i \le T$,*

$$e_i^{k+1} := f_i(w_{i+1}^k) - G_i^{k+1},\ \hat{e}_i^{k+1} := \nabla f_i(w_{i+1}^k) - J_i^{k+1}, \tag{46}$$

$$A_{k,i} := f_i(w_{i+1}^{k+1}) - f_i(w_{i+1}^k) - \nabla f_i(w_{i+1}^k)(w_{i+1}^{k+1} - w_{i+1}^k). \tag{47}$$

a) Under Assumption 2.1, we have, for $1 \le i \le T$,

$$\begin{aligned}\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 \le (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \frac{L_{\nabla f_i}^2}{4\tau_k}\|w_{i+1}^{k+1} - w_{i+1}^k\|^4 + \tau_k^2\|e_i^{k+1}\|^2 + \dot{r}_i^{k+1} \\ + \|\hat{e}_i^{k+1}\|^2\|w_{i+1}^{k+1} - w_{i+1}^k\|^2,\end{aligned} \tag{48}$$

where,

$$\begin{aligned}\dot{r}_i^{k+1} := 2\tau_k\langle e_i^{k+1}, A_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k) + \hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle \\ + 2\langle \hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k), A_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k)\rangle.\end{aligned} \tag{49}$$

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

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

Proof. We first prove part a). When $1 \le i < T$, 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

$$\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2$$

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

Combining the above inequality with (15) and noting that under Assumption 2.1,

$$
\|A_{k,i}\| \le \frac{1}{2}\min\left\{4L_{f_i}\|w_{i+1}^{k+1} - w_{i+1}^k\|, L_{\nabla f_i}\|w_{i+1}^{k+1} - w_{i+1}^k\|^2\right\}, \tag{50}
$$

we obtain (48).

We now prove part b). Note that by the definition of (5) and (46), 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 3.2. *For two vectors x, y of equal dimension and any $\delta > 0$, we have*

$$
\|x+y\|^2 \le (1+\delta)\|x\|^2 + \left(1 + \frac{1}{\delta}\right)\|y\|^2 \tag{51}
$$

$$
\|x+y\|^4 \le (1+\delta)^3\|x\|^4 + \left(1 + \frac{1}{\delta}\right)^3\|y\|^4 \tag{52}
$$

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

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

(51) follows. Next, by (51) 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 3.3. *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 3.4. *Under Assumption 2.1 and Assumption 3.1, for $1 \le i \le T$, and with the choice of $\tau_0 = 1$ (for simplicity), we have*

$$\mathbb{E}[\|w_i^{k+1} - w_i^k\|^2 | \mathscr{F}_k] \le \tilde{c}_i \; \tau_k^2, \tag{53}$$
$$\mathbb{E}[\|w_i^{k+1} - w_i^k\|^4 | \mathscr{F}_k] \le c_i \; \tau_k^4, \tag{54}$$

where,

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

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

with

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

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 3.4. 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 $1 \le i \le T$,

$$D_{k,i} := A_{k,i} + \tau_k e_i^{k+1} + \hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k). \tag{55}$$

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

$$f_i(w_{i+1}^{k+1}) - w_i^{k+1} = (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k) + D_{k,i}. \tag{56}$$

We now prove (53). By equation (56), Lemma 3.2 using $\delta = \tau_k$, we obtain

$$\begin{aligned}\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 &\le (1-\tau_k^2)(1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \frac{(1+\tau_k)}{\tau_k}\|D_{k,i}\|^2 \\ &\le (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \frac{2}{\tau_k}\|D_{k,i}\|^2. \end{aligned}\tag{57}$$

Moreover, we have

$$\begin{aligned}\|D_{k,i}\|^2 &= \|A_{k,i}\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + \|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2 + 2\tilde{r}_{k,i}, \\ r'_{k,i} &= \langle A_{k,i}, \tau_k e_i^{k+1} + \hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle + \tau_k\langle e_i^{k+1}, \hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle\end{aligned}\tag{58}$$

which together with the fact that $\mathbb{E}[\tilde{r}_{k,i}|\mathscr{F}_k] = 0$ under Assumption 2.2, we have imply that

$$\begin{aligned}\mathbb{E}[\|D_{k,i}\|^2|\mathscr{F}_k] &= \mathbb{E}[\|A_{k,i}\|^2|\mathscr{F}_k] + \tau_k^2\mathbb{E}[\|e_i^{k+1}\|^2|\mathscr{F}_k] + \mathbb{E}[\|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2|\mathscr{F}_k] \\ &\le \tau_k^2\mathbb{E}[\|e_i^{k+1}\|^2|\mathscr{F}_k] + \left(4L_{f_i}^2 + \mathbb{E}[\|\hat{e}_i^{k+1}\|^2|\mathscr{F}_k]\right)\mathbb{E}[\|w_{i+1}^{k+1} - w_{i+1}^k\|^2|\mathscr{F}_k],\end{aligned}\tag{59}$$

where the second inequality follows from (50). Hence, noting the result from Proposition 3.1.a), $w_{T+1}^k = x^k$, and under Assumption 3.1, we have

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

Using (57) with $i = T$, the above inequality, and Lemma 2.8 with the choice of $\tau_0 = 1$, we have

$$\mathbb{E}[\|f_T(x^k) - w_T^k\|^2|\mathscr{F}_k] \le 2\left[\sigma_{G_T}^2 + \left(4L_{f_T}^2 + \varrho_{J_T}^2\right)\left(\prod_{i=1}^T \sigma_{J_i}^2\right)\beta^{-2}\right]. \tag{60}$$

Moreover, under Assumption 3.1 and Lemma 3.1.b), we have

$$\mathbb{E}[\|w_{i+1}^{k+1} - w_i^k\|^2|\mathscr{F}_k] \le \tau_k^2\mathbb{E}\left[2\|f_i(w_{i+1}^k) - w_i^k\|^2 + \|e_i^{k+1}\|^2 + \frac{2}{\tau_k^2}\|J_i^{k+1}\|^2\|w_{i+1}^{k+1} - w_{i+1}^k\|^2\Big|\mathscr{F}_k\right], \tag{61}$$

implying that

$$\mathbb{E}[\|w_T^{k+1} - w_T^k\|^2|\mathscr{F}_k] \le \tau_k^2\left[5\sigma_{G_T}^2 + 2(8L_{f_T}^2 + 2\varrho_{J_T}^2 + \sigma_{J_T}^2)\left(\prod_{i=1}^T \sigma_{J_i}^2\right)\beta^{-2}\right]. \tag{62}$$

This completes the proof of (53) 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 $\mathbb{E}[\|w_{i+1}^{k+1} - w_{i+1}^k\|^2|\mathscr{F}_k] \le \tilde{c}_{i+1}\tau_k^2$ for some $1 \le i < T$. Hence, by (58) and under Assumption 3.1, we have

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

which together with Lemma 2.8, imply that

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

Thus, by (61), we obtain

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

where after using Lemma 3.3, $\tilde{c}_i$ for $1 \le i \le T-2$, is as defined in the statement of Lemma 3.4. Hence, we obtain the claim in (53) by induction.

We now start proving (54). We start with $i = T$. By equation (56), Lemma 3.2 and setting $\delta = \tau_k$ we get

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

Now, by (58), 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

$$\begin{aligned}
\|D_{k,i}\|^4 &\le \|A_{k,i}\|^4 + \tau_k^4\|e_i^{k+1}\|^4 + \|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^4 + 4\tau_k^2\|e_i^{k+1}\|^2\|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2 \\
&\quad + 4\|A_{k,i}\|^2\left(\tau_k^2\|e_i^{k+1}\|^2 + \|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2\right) + 4r''_{k,i}, \qquad (63) \\
r''_{k,i} &= r'_{k,i}\left(\|A_{k,i}\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + \|\hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\|^2\right) + \tau_k\|A_{k,i}\|^2\langle e_i^{k+1}, \hat{e}_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k)\rangle.
\end{aligned}$$

By definition of d^k and Assumption 3.1, we obtain $\|A_{k,T}\| \le 2L_{f_T}\tau_k\|d^k\|$. By this inequality and by applying Lemma 3.2 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 3.1 and Proposition 3.1, 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 2.8, 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 3.1 and Lemma 3.1, 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 (54) 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 equation (56), 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 3.1, $\|A_{k,i}\| \le 2L_{f_i}\|w_{i+1}^{k+1} - w_{i+1}^k\|$. Using this fact and Lemma 3.2 with $\delta = 1$, in (63), 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 2.8, 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 3.1 and Lemma 3.1, 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 3.3, c_i for $1 \le i \le T-2$, is as defined in the statement of Lemma 3.4. Hence, we obtain the claim in (54) by induction.

∎

The next result is the counterpart of Lemma 2.7 for Algorithm 2.

Lemma 3.5. *Recall the definition of the merit function in (16). Define $w^k := (w_1^k, \ldots, w_T^k)$ for $k \ge 0$. Let $\{x^k, z^k, u^k, w_1^k, \ldots, w_T^k\}_{k\ge0}$ be the sequence generated by Algorithm 2. Suppose for $1 \le i \le T$, we have*

$$\max_{2\le j\le T} C_j^2 \le \frac{(\beta_k - \lambda)}{T}(\gamma_i b - \lambda) \tag{64}$$

where C_j's are defined in Lemma 2.4. Then, under Assumption 2.1 and Assumption 3.1, we have

$$\lambda\sum_{k=0}^{N-1}\tau_k\left[\|d^k\|^2 + \sum_{i=1}^{T-1}\|f_i(w_{i+1}^k) - w_i^k\|^2 + \|f_T(x^k) - w_T^k\|^2\right] \le W(x^0, z^0, w^0) + \sum_{k=0}^{N-1}\hat{R}^{k+1}, \tag{65}$$

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 (20) and (49). Furthermore, notice that (64) is satisfied, when we pick

$$\gamma_i = 1 \qquad \lambda = 1/2 \qquad \beta_k \equiv \beta \geq \frac{1}{2} + 2T \max_{2\leq j\leq T} C_j^2. \tag{66}$$

Proof. Noting Lemma 3.1 and definition of $\hat{r}_i^{k+1}$, we have

$$\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 - \|f_i(w_{i+1}^k) - w_i^k\|^2 \leq -\tau_k\|f_i(w_{i+1}^k) - w_i^k\|^2 + \hat{r}_i^{k+1},$$
$$\|f_T(x^{k+1}) - w_T^{k+1}\|^2 - \|f_T(x^k) - w_T^k\|^2 \leq -\tau_k\|f_T(x^k) - w_T^k\|^2 + \hat{r}_T^{k+1},$$

Combining the above inequalities with (22), (24), and noting definition of the merit function in (16), 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 (64) and last follows by Young's inequality. Thus, by summing up the above inequalities and re-arranging the terms, we obtain (65). also It is easy to see that (64) holds, by picking the parameters as in (66). ∎

In the next result, we show the error terms in the right hand side of (65) is bounded in the order of $\sum_{k=1}^N \tau_k^2$ in expectation.

Proposition 3.1. *Suppose $\beta_k = \beta > 0$ for all k and $\tau_0 = 1$. We then have*

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

where

$$\hat{\sigma}^2 := \sum_{i=1}^{T-1} \gamma_i \left(\frac{L^2_{\nabla f_i} c_{i+1}}{4} + \varrho^2_{J_i} \tilde{c}_{i+1} + \sigma^2_{G_i} \right) + \frac{\gamma_T L^2_{\nabla f_T} \sigma_d}{4} + 2L_{\nabla \eta} \left(\prod_{i=1}^{T} \sigma^2_{J_i} \right)$$
$$+ \frac{1}{2} \left[2\gamma_T \sigma^2_{J_T} + \frac{1}{\beta_k^2} \left(\prod_{i=1}^{T} \sigma^2_{J_i} \right) \left\{ 2\gamma_T \varrho^2_{J_T} + L_{\nabla F} + L_{\nabla \eta} + 2C_T L_{f_T} \right\} \right]. \tag{67}$$

Proof. Noting the convexity of $\|\cdot\|^4$, the first inequality follows similarly to that of Proposition 2.1 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 3.5, Lemma 3.4 and Assumption 3.1, we obtain σ^2 as in (67).

■

We remark that the c_{i+1} in the right hand side of (67) 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 3.1. *Suppose that $\{x^k, z^k\}_{k\ge 0}$ are generated by Algorithm 2, and Assumption 2.1 and Assumption 3.1 hold. Also assume the parameters satisfy (66) and the step sizes $\{\tau_k\}$ satisfy (34).*

(a) The results in parts a) and b) of (35) still hold by replacing σ^2 by $\hat{\sigma}^2$.

b) If the stepsizes are set to (39), the results of part c) of (35) also hold with replacing σ^2 by $\hat{\sigma}^2$.

Proof. The proof follows from the same arguments in the proof of (35) by noticing (65), and Proposition 3.1, hence, we skip the details.

■

Remark 3. *Note that Algorithm 2 does not use a mini-batch of samples in any iteration. Thus, (41) (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 (1), 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 [22] for the two-level composite problem which in turn is in the same order for single-level smooth stochastic optimization.*

4 Concluding remarks

In this paper, 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 [4, 14]. 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 [3, 12, 26, 28, 38] for the standard stochastic gradient algorithm when $T = 1$.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Jose Blanchet, Donald Goldfarb, Garud Iyengar, Fengpei Li, and Chaoxu Zhou. Unbiased simulation for optimizing stochastic function compositions. *arXiv preprint arXiv:1711.07564*, 2017.

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

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

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

[14] 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.

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

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

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

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

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

[20] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

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

[22] 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.

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

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

[25] 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.

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

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

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

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

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

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

[32] 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.

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

[34] 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.

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

[36] 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.

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

[38] 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.

[39] Junyu Zhang and Lin Xiao. Multi-level composite stochastic optimization via nested variance reduction. *arXiv preprint arXiv:1908.11468*, 2019.