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

Krishnakumar Balasubramanian[†] Saeed Ghadimi[‡] Anthony Nguyen[§]

February 15, 2022

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 [20] to the T level case, can achieve a sample complexity of $\mathcal{O}_T(1/\epsilon^6)$ by using mini-batches of samples in each iteration, where $\mathcal{O}_T$ hides constants that depend on T. By modifying this algorithm using linearized stochastic estimates of the function values, we improve the sample complexity to $\mathcal{O}_T(1/\epsilon^4)$. This modification not only removes the requirement of having a mini-batch of samples in each iteration, but also makes the algorithm parameter-free and easy to implement. 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 standard assumptions (unbiasedness and boundedness of the second moments) 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 functions, the composite function F is bounded below by $F^*>-\infty$, 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 $f_i(y) = \mathbb{E}_{\xi_i}[G_i(y,\xi_i)]$ 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 [5, 18, 19, 21, 26, 33], 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), is $\nabla F(x) = \nabla f_T(y_T)\nabla f_{T-1}(y_{T-1})\cdots\nabla f_1(y_1)$, where ∇f_i denotes the transpose of the Jacobian of f_i, $y_i = f_{i+1}\circ\cdots\circ f_T(x)$ for $1\leq i<T$, and $y_T=x$. Our goal is to solve the above optimization

[*] Authors are 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.

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 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 [14], recently there has been a renewed interest in this problem due to applications arising in mathematical finance, nonparametric statistics, deep generative modeling and reinforcement learning. We refer the reader to [3, 4, 9, 15, 20, 24, 34, 35, 36, 38] for such applications and various algorithmic approaches for solving problem (1). In particular [34] and [36] 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 [15] and [9], 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., dependence of the convergence rate on 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 $T = 2$, [20] proposed a single time-scale Nested Averaged Stochastic Approximation (NASA) algorithm. The authors showed that by modifying the specific Lyapunov function, defined in [29] for nonsmooth single-level stochastic optimization, the convergence analysis of the NASA algorithm can be established such that its complexity bound matches the case of $T = 1$. This resolved the above question for $T = 2$. However, constructing similar algorithms for the case of general T 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 algorithms are applicable to both unconstrained and constrained cases, as we do not make any boundedness assumption on the feasible set X. Their complexity results are summarized in Table 1. The first algorithm is based on an extension of the NASA algorithm from [20] (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 by adding a linear bias correction term in evaluating the inner function values, motivated by the recent work [31] for nonsmooth multi-level composition 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. Furthermore, it works with any positive constant step-size parameter choice (independent of problem parameters), thus making it easy to implement.

Comparisons to related works. A summary of our results, in comparison to the most related work of [36] is provided in Table 1. We use $\mathcal{O}(\cdot)$ to represent the fact that the constants involved are only numerical constants that are independent of T. However, when the constants involved are dependent on T, we use $\mathcal{O}_T(\cdot)$.

The approach and the results in [36] are provided only for the unconstrained setting. Furthermore, [36] requires an additional bounded fourth moment assumption on the stochastic Jacobian matrices. In an earlier version of our work uploaded to arXiv, we also made the same assumption, which however, we do not require in this work, thereby widening the applicability of the

Method	[36]	Algorithm 1	Algorithm 2
Convergence Rate	$\mathcal{O}_T\left(N^{-4/(7+T)}\right)$	$\mathcal{O}_T\left(N^{-1/2}\right)$	
Oracle Complexity	$\mathcal{O}_T\left(1/\epsilon^{(7+T)/2}\right)$	$\mathcal{O}_T\left(1/\epsilon^{6}\right)$	$\mathcal{O}_T\left(1/\epsilon^{4}\right)$
Mini-batch	No	Yes	No
Feasible Set	$X = \mathbb{R}^d$	General case	
Oracle Assumption	Finite 4th moment	Finite 2nd moment	

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 $\mathbb{E}[\sqrt{V(x,z)}]$ and oracle complexity refers to the number of calls to the stochastic first-order oracle to obtain a ϵ-pair. The constants in [36] and our work have exponential dependency on T in the worst case. See Remark 1 and Remark 3 for more details.

proposed method. We also highlight the related work of [38] 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 [38] is a proximal variant of SPIDER variance reduction technique [16] 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, [38] 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. 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 use 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.

As mentioned above, our Algorithm 2 is related to a concurrent work [31]. In this work, the author focuses on nonsmooth multi-level composition problems 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. By further making the smoothness assumption, the author also establishes a sample complexity of $\mathcal{O}_T(1/\epsilon^4)$, similar to that of Algorithm 2 in Theorem 3.1. However, our convergence analysis here is distinct since we do not require the boundedness assumption of the feasible set which makes our method applicable to both unconstrained constrained problems.[1]

After our first draft appeared on arXiv, [6] also proposed an approach for stochastic multi-level compositional optimization problems and obtained similar rates as us, albeit only for unconstrained problems and under the stronger assumption that the stochastic functions $G_i(y, \xi_i)$ are Lipschitz, almost surely.[2]

[1] We also remark that the finite-time convergence analysis of [31], from our communication with the author, was not complete in the first version released on arXiv. However, more recently, after release of the first version of our paper on arXiv, the author has refined the convergence analysis in [31].

[2] It is worth noting that [6] was released several months after the first draft of [31].

1.1 Motivating Applications

We now provide two motivating applications of stochastic multi-level composition optimization problems.

1.1.1 Risk-averse Optimization

Our first motivating example to consider multilevel stochastic composite optimization problems is from the field of risk-averse stochastic optimization [32, 36]. Specifically, the mean-deviation risk-averse optimization is given by the following form:

$$\min_x \left\{ \mathbf{E}[U(x,\xi)] + \lambda \left(\mathbf{E}\left[\max\left\{0, U(x,\xi) - \mathbf{E}[U(x,\xi)]\right\}^2 \right] \right)^{1/2} \right\}. \tag{2}$$

As noted in [31, 36], the problem in (2) is a stochastic three-level composition optimization problem with

$$\begin{aligned} &f_3(x) := (x, \mathbf{E}[U(x,\xi)]), \qquad f_2(y_3,y_4) := \left(y_4, \mathbf{E}\left[\max\left\{0, U(y_3,\xi) - y_4\right\}^2 \right] \right), \\ &f_1(y_1,y_2) := y_1 + \lambda\sqrt{y_2+\delta}. \end{aligned}$$

Here, $\delta > 0$ is added to make the square root function smooth. We consider a semi-parametric data generating process given by a single-index model of the form $b = g(a^\top x^*) + \zeta$, where $g : \mathbb{R} \to \mathbb{R}$ is called the link function. Such single-index models are widely used in statistics, machine learning and economics [28]. Here, X is the input data which is assumed to be independent of the noise ζ. The goal is to estimate the index β^* in a risk-averse manner, as they are well-known to provide stable solutions [36]. In this case, $\xi := (a,b)$ and the function $U(x,\xi)$ depends on the loss function. We will revisit this example in Section 4 for numerical experiments.

1.1.2 Training large-scale Graph Neural Networks (GNNs)

Our second motivating example is training GNNs, which has been formulated as a stochastic multi-level compositional optimization problem in [7]. Each layer of a GNN is given by a matrix $H^{(i-1)} = L\sigma(H^{(i)})W^{(i)} \in \mathbb{R}^{n\times p}$, for $2 \le i \le T$. Here, L is the normalized graph Laplacian matrix (calculated as $D^{-1/2}AD^{-1/2}$ or $D^{-1}A$, where D is the degree matrix and A is the adjacency matrix given the data matrix $U \in \mathbb{R}^{n\times d}$), $W^{(i)}$ is the weight matrix at layer i, and σ is the activation function, which either is a sigmoidal function $\sigma(s) := 1/(1+e^{-s})$ or the ReLU function $\sigma(s) := \max\{0,s\}$ operating entry-wise on matrices. Furthermore, $H^{(T)} := U$. When the size of the data set n is large, subsampling methods are used to train the GNN [7].

In our notation, the optimization variable $x := \{W^{(1)}, \ldots, W^{(T-1)}\}$. The function f_i is given by $f^{(i)} = L\sigma(H^{(i+1)})W^{(i)}$, for $i = 2, \ldots T-1$, with $f^{(T)} := U$. Furthermore, $f^{(1)}$ will be the user-defined loss function based on the label vector $Y \in \mathbb{R}^n$. The stochasticity in the problem is due to the fact that the data is subsampled when constructing the graph. Specifically, we have the random function given by $\tilde{L}^{(i-1)}\sigma(H^{(i)})W^{(i)}$, where $\tilde{L}^{(i-1)}$ is a stochastic approximation of the matrix L such that $\mathbf{E}[\tilde{L}^{(i-1)}] = L$. We refer the interested reader to [7, Section 3], for additional details. We also remark that while the ReLU activation does not satisfy the smoothness assumptions we make in this work, the sigmoidal function does.

1.2 Organization

The rest of our paper is organized as follows. In Section 2, we present our first algorithm and analyze its convergence for solving (1) with any $T \geq 1$. In Section 3, we present a modification of this algorithm and show that it can recover the best-known sample complexity for (single-level) smooth stochastic optimization. In Section 4, we present some numerical experiments and conclude the paper with some remarks in Section 5.

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 [36]. 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 [20] 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 [36]. Our approach is formally presented in Algorithm 1.

Algorithm 1 Multi-level Nested Averaging Stochastic Gradient Method

```
  Input: Positive integer sequences $\{b_k, \tau_k\}_{k\geq 0}$, step-size parameter $\beta$, and initial points $x^0 \in X$, $z^0 \in \mathbb{R}^{d_T}$, $w_i^0 \in \mathbb{R}^{d_{i-1}}$ $1 \leq i \leq T$, and a probability mass function $P_R(\cdot)$ supported over $\{1, 2, \dots, N\}$, where $N$ is the number of iterations.
  0. Generate a random integer number $R$ according to $P_R(\cdot)$.
  for $k = 0, 1, 2, \dots, R$ do
    1. Compute
         $u^k = \underset{y \in X}{\operatorname{argmin}} \left\{ \langle z^k, y - x^k \rangle + \frac{\beta}{2} \|y - x^k\|^2 \right\},$    (3)
       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
         $x^{k+1} = (1 - \tau_k) x^k + \tau_k u^k,$    (4)
         $z^{k+1} = (1 - \tau_k) z^k + \tau_k \prod_{i=1}^{T} J_{T+1-i}^{k+1},$    (5)
         $w_i^{k+1} = (1 - \tau_k) w_i^k + \tau_k \bar{G}_i^{k+1}, \qquad 1 \leq i \leq T,$    (6)
       where
         $\bar{G}_i^{k+1} = \frac{1}{b_k} \sum_{j=1}^{b_k} G_{i,j}^{k+1}.$    (7)
  end for
  Output: $(x^R, z^R, w_1^R, \dots, w_T^R)$.
```

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 (3), 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 for the sake of completeness. Second, when $T = 2$ and $b_k = 1$, this algorithm reduces to the NASA algorithm presented in [20]. 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. *For $i \in \{1, \ldots, T\}$, we have $\mathbb{E}[J_i^{k+1} | \mathscr{F}_k] = \nabla f_i(w_{i+1}^k)$, and $\mathbb{E}[G_i^{k+1} | \mathscr{F}_k] = f_i(w_{i+1}^k)$.*
2. *For $i \in \{1, \ldots, T\}$, we have $\mathbb{E}[\|G_i^{k+1} - f_i(w_{i+1}^k)\|^2 | \mathscr{F}_k] \leq \sigma_{G_i}^2$, $\mathbb{E}[\|J_i^{k+1} - \nabla f_i(w_{i+1}^k)\|^2 | \mathscr{F}_k] \leq \hat{\sigma}_{J_i}^2$, and $\mathbb{E}[\|J_i^{k+1}\|^2 | \mathscr{F}_k] \leq \sigma_{J_i}^2$. Here $\| \cdot \|$ denotes the Euclidean norm for vectors and the Frobenius 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 reemphasize that the assumptions made in [38] 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 convergence results in the multi-level case; similar assumptions have been made, for example, in [36]. 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 is that of F_{T-1} since for any $x, y \in X$, we have

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

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

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 [38], 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 the gradient term $\nabla \bar{f}_i(x, \bar{w}_i) = \nabla f_T(x) \nabla f_{T-1}(w_T) \cdots \nabla f_i(w_{i+1})$ with $\bar{w}_i = (w_{i+1}, \ldots, w_T)$ for any $x \in X, w_j \in \mathbb{R}^{d_j}$ $j = i+1, \ldots, T$. Then under Assumption 2.1, we have*

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

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

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

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

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

We then have

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

■

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

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

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 $j+1, j \in \{1, \dots, T-2\}$. 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 the induction hypothesis. ■

Lemma 2.4. *Define*

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

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

$$\left\| \nabla F(x) - \nabla f_T(x) \prod_{i=2}^{T} \nabla f_{T+1-i}(w_{T+2-i}) \right\| \leq \sum_{j=2}^{T-1} C_j \|f_j(w_{j+1}) - w_j\| + C_T \|f_T(x) - w_T\| \tag{8}$$

Proof. By Lemma 2.2 and Lemma 2.3, we have

$$\begin{aligned}
&\left\| \nabla F(x) - \nabla f_T(x) \prod_{i=2}^{T} \nabla f_{T+1-i}(w_{T+2-i}) \right\| \leq \sum_{j=1}^{T-1} R_j \|f_{j+1} \circ \cdots \circ f_T(x) - w_{j+1}\| \\
&= \sum_{j=1}^{T-2} R_j \|f_{j+1} \circ \cdots \circ f_T(x) - w_{j+1}\| + R_{T-1} \|f_T(x) - w_T\| \\
&= \sum_{j=1}^{T-2} R_j \|f_{j+1}(w_{j+2}) - w_{j+1}\| + \sum_{j=1}^{T-2} R_j \sum_{\ell=j+2}^{T} \left(\prod_{i=j+1}^{\ell-1} L_{f_i} \right) \|f_\ell(w_{\ell+1}) - w_\ell\|
\end{aligned}$$

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

Aggregating the constants for $\|f_j(w_{j+1}) - w_j\|$, we get the result. ■

The following result also shows the Lipschitz continuity of the gradient of the objective function of the subproblem (3). One can see [20] 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 $\{x^k\}_{k\geq 0}$ and $\{w_i^k\}_{k\geq 0}$ $1 \leq i \leq T$ be generated by Algorithm 1. Denote*

$$d^k = u^k - x^k, \qquad w_{T+1}^k \equiv x^k \quad \forall k \geq 0, \qquad A_{k,i} = f_i(w_{i+1}^{k+1}) - f_i(w_{i+1}^k) \quad 1 \leq i \leq T. \tag{9}$$

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

$$\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 \leq (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \frac{1}{\tau_k}\|A_{k,i}\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + r_i^{k+1}, \tag{10}$$

$$\|w_i^{k+1} - w_i^k\|^2 \leq \tau_k^2\left[\|f_i(w_{i+1}^k) - w_i^k\|^2 + \|e_i^{k+1}\|^2 - 2\langle e_i^{k+1}, f_i(w_{i+1}^k) - w_i^k\rangle\right], \tag{11}$$

where

$$r_i^{k+1} = 2\tau_k\langle e_i^{k+1}, A_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k)\rangle, \qquad e_i^{k+1} = f_i(w_{i+1}^k) - \bar{G}_i^{k+1}. \tag{12}$$

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

$$\|f_T(x^{k+1}) - w_T^{k+1}\|^2 \leq (1-\tau_k)\|f_T(x^k) - w_T^k\|^2 + L_{f_T}\tau_k\|d^k\|^2 + \tau_k^2\|e_T^{k+1}\|^2 + r_T^{k+1}, \tag{13}$$

$$\begin{aligned}\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 &\leq (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + \bar{r}_i^{k+1} \\ &\quad + L_{f_i}^2\tau_k\left[\|f_{i+1}(w_{i+2}^k) - w_{i+1}^k\|^2 + \|e_{i+1}^{k+1}\|^2\right] \qquad 1 \leq i \leq T-1,\end{aligned} \tag{14}$$

where

$$\bar{r}_i^{k+1} = -2\tau_k L_{f_i}^2\langle e_{i+1}^{k+1}, f_{i+1}(w_{i+2}^k) - w_{i+1}^k\rangle + r_i^{k+1}. \tag{15}$$

Proof. Noting (6), (10), and (12), 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 (12), (10) follows by noting that

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

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

due to Cauchy-Schwarz and Young's inequalities. Also, (11) directly follows from (6) since

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

To show part b), note that by (4), (9), and Lipschitz continuity of f_i, we have

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

for $1 \leq i \leq T-1$. The results then follows by noting (10) and (11). ■

We remark that the mini-batch sampling in (7) 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 (14). 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 converges without using mini-batch of samples in each iteration, as shown in [20].

Recalling the definition of F^* from Section 1 and denoting $w := (w_1, \ldots, w_T)$, we define, for some positive constants $\gamma = (\gamma_1, \ldots, \gamma_T)$, the merit function

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

which will be used in our next result for establishing convergence analysis of Algorithm 1. It is worth noting that $W_\gamma(x,z,w) \geq 0$ due to that facts that $F(x) \geq F^*, \eta(x,z) \leq 0$ (by Lemma 2.5), and $\gamma > 0$. The precise values of the constants $\gamma_1, \ldots, \gamma_T$ will be set later in our analysis. We should also mention that the above summation can start from $i=2$, in which case the convergence analysis is slightly simpler. However, we use (16) in our analysis since, as a byproduct, it gives us an online certificate for the stochastic values of the objective function. The above function is an extension of the one used in [20] for the case of $T=2$, to the multi-level setting of $T \geq 1$. A variant of this function (including only the first two terms in (16)) was used in the literature as early as [29] and was used later in [30] for nonsmooth single-level stochastic optimization.

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

a) If

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

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

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

$$+\sum_{k=0}^{N-1} R^{k+1}, \tag{18}$$

where

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

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

and $r_i^{k+1}, \bar{r}_i^{k+1}$ are defined in (12) and (15), respectively.

b) If parameters are chosen as

$$\gamma_j := 2^{j-1}(L_{f_1} \cdots L_{f_{j-1}})^2 \sqrt{T} \quad 2 \le j \le T, \qquad \beta \ge \lambda + \gamma_T + \frac{T \max_{2 \le i \le T} C_i^2}{4\lambda},$$
$$\gamma_1 = \lambda = \frac{1}{2} \min_{2 \le i \le T} (\gamma_i - \gamma_{i-1} L_{f_{i-1}}^2) = \frac{\min_{2 \le i \le T} \gamma_i}{4}. \tag{21}$$

Then, conditions in (17) are satisfied.

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

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

Second, note that by the optimality condition of (3), we have

$$\langle z^k + \beta(u^k - x^k), x^k - u^k \rangle \ge 0, \text{ which implies } \langle z^k, d^k \rangle + \beta \|d^k\|^2 \le 0. \tag{23}$$

Then, noting (4), (5), and in the view of Lemma 2.5, we obtain

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

Third, noting Lemma 2.6.b), we have

$$\sum_{i=1}^{T-1} \gamma_i \left[\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 - \|f_i(w_{i+1}^k) - w_i^k\|^2 \right]$$

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

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

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

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

$$
\begin{aligned}
&- \left(\frac{\beta - \gamma_T}{T}\right) \|d^k\|^2 - (\gamma_j - \gamma_{j-1} L_{f_{j-1}}^2) \|f_j(w_{j+1}^k) - w_j^k\|^2 + C_j \|d^k\| \|f_j(w_{j+1}^k) - w_j^k\| \\
&\le -\lambda \Big[\frac{1}{T} \|d^k\|^2 + \|f_j(w_{j+1}^k) - w_j^k\|^2\Big] \qquad \forall j \in \{1, \ldots, T\},
\end{aligned}
$$

which together with the above inequality imply that

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

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

The next technical result helps us to simplify our convergence analysis.

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

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

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 $q_{k+1} \leq (1-\tau_k)q_k + p_k \quad k \geq 0$ for sequences $\{q_k, p_k\}_{k\geq 0}$. Then, we have

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

Proof. To show part a), note that

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

To show part b), by dividing both sides of the inequality by Γ_{k+1} and noting (26), we have

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

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

The next result shows the boundedness of the error terms in the right hand side of (18) in expectation. This is an essential step in establishing the convergence analysis of the algorithm.

Proposition 2.1. *Suppose that Assumption 2.2 holds and (for simplicity) $\tau_0 = 1$, $\beta > 0$ for all k. Then, for any $k \geq 1$, we have*

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

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

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

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

we have

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

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

Proof. The first inequality in (27) directly follows by (23) and Cauchy-Schwarz inequality. Noting (5), 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

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

Noting (27), 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 (12), (15), the choice of b_k in (29), and under Assumption 2.2, we have

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

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

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 are 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 $\bar{u} = \bar{x}$ and $\bar{z} = \nabla F(\bar{x})$, where

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

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* $\mathbb{E}[\sqrt{V(\bar{x}, \bar{z})}] \leq \epsilon$*, where*

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

and $\bar{u}$ *is the solution to (31).*

We emphasize that in Definition 2.1, we consider a unified termination criterion for both the unconstrained and constrained cases. When $X = \mathbb{R}^{d_T}$, $V(\bar{x}, \bar{z})$ provides an upper bound for the $\|\nabla F(\bar{x})\|^2$. This can be simply seen from the fact that $\bar{u} - \bar{x} = \bar{z}$ in (31) for unconstrained problems and hence from (32), we have

$$V(\bar{x}, \bar{z}) = \|\bar{z}\|^2 + \|\bar{z} - \nabla F(\bar{x})\|^2 \geq \frac{1}{2}\|\nabla F(\bar{x})\|^2.$$

We also refer the reader to [20] for the relation between $V(\bar{x}, \bar{z})$ and other common gradient-based termination criteria used in the literature such as gradient mapping ([17, 22, 23]) and proximal mapping ([12]). Furthermore, as shown in [20], we have

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

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

Theorem 2.1. *Suppose that* $\{x^k, z^k\}_{k \geq 0}$ *are generated by Algorithm 1, Assumption 2.1 and Assumption 2.2 hold. Also assume that the parameters satisfy (21) and step sizes* $\{\tau_k\}$ *are chosen such that*

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

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

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

where

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

σ^2 *is defined in (30) and*

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

(b) *As a consequence, we have*

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

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

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

we have

$$\begin{aligned}
\mathbb{E}[\|\nabla F(x^R) - z^R\|^2] &\leq \frac{1}{\sqrt{N}} \left[\frac{4L^2(T-1)\left[W_\gamma(x^0, z^0, w^0) + 2\sigma^2\right]}{\lambda} + 2\prod_{\ell=1}^{T} \sigma_{J_\ell}^2 \right] \\
&:= \frac{\mathcal{B}_2(\sigma^2)}{\sqrt{N}}, &(40)\\
\mathbb{E}[V(x^R, z^R)] &\leq \frac{1}{\sqrt{N}} \left[\mathcal{B}_2(\sigma^2) + \frac{\max(1, \beta^2)}{\lambda} \left[W_\gamma(x^0, z^0, w^0) + 2\sigma^2\right] \right], &(41)\\
\mathbb{E}[\|f_i(w_{i+1}^R) - w_i^R\|^2] &\leq \frac{1}{\lambda\sqrt{N}} \left[W_\gamma(x^0, z^0, w^0) + 2\sigma^2\right] \qquad i = 1, \dots, T. &(42)
\end{aligned}$$

Proof. We first show part (a). Noting (5), 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

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

where the inequality follows from convexity of $\|\cdot\|^2$ and Lipschitz continuity of gradient of F. Thus, in the view of Lemma 2.8, 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 $\sum_{k=1}^{N} \tau_k\|\nabla F(x^k) - z^k\|^2 =$

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

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

where the last inequality follows from (34).

Now, observe that under Assumption 2.2, we have

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

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

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

Combining the above observations with (43) and in the view of (37), we obtain

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

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

Part (b) then follows from part (a), (33), (18), and noting that

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

Part (c) also follows by noting that choice of τ_k in (39) implies that

$$
\begin{aligned}
&\sum_{k=1}^{N}\tau_k\geq\sqrt{N},\quad \sum_{k=0}^{N}\tau_k^2=2,\quad \Gamma_k=\left(1-\frac{1}{\sqrt{N}}\right)^{k-1},\\
&\sum_{i=k+1}^{N}\tau_i\Gamma_i=\left(1-\frac{1}{\sqrt{N}}\right)^{k}\frac{1}{\sqrt{N}}\sum_{i=0}^{N-k-1}\left(1-\frac{1}{\sqrt{N}}\right)^{i}\leq\left(1-\frac{1}{\sqrt{N}}\right)^{k},
\end{aligned}
$$

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*

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

due to (29) and (39). This bound is much better than $\mathcal{O}_T\left(1/\epsilon^{(7+T)/2}\right)$ obtained in [36] when $T > 4$[3] . 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 $\sigma_{J_i}^2$ in (30) are all equal. We emphasize that [36] and [38] 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 [36], it is still worse than $\mathcal{O}_T\left(1/\epsilon^4\right)$ obtained in [20] 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 best known rate of convergence for problem (1) for any $T \geq 1$, under Assumptions 2.1 and 2.2. Indeed, when $T > 2$, we have accumulated errors in estimating the inner function values. Hence, in Algorithm 1 we use mini-batch sampling in (6) 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 (44). This modification reduces the bias error in estimating the inner function values which together with a refined convergence analysis enable us to obtain a sample complexity of $\mathcal{O}_T(1/\epsilon^4)$ with Algorithm 2, for any $T \geq 1$ without using any mini-batches. Furthermore, Algorithm 2 works with any constant choice of step-size parameter β (independent of the problem parameters), making it easy to implement. As mentioned previously, Algorithm 2 is motivated by the algorithm in [31] proposed for solving nonsmooth multi-level stochastic composition problems. However, [31] assumes that all functions f_i explicitly depend on the decision variable x which makes the composition function a variant of the general case in (1). It is also worth mentioning that other linearization techniques have been used in [8, 13] in estimating the stochastic inner function values for weakly convex two-level composition problems.

To establish the rate of convergence of Algorithm 2, we first need the next result which provides the recursion on the errors in estimating the inner function values.

Lemma 3.1. *Let $\{x^k\}_{k\geq 0}$ and $\{w_i^k\}_{k\geq 0}$ be generated by Algorithm 2. Define, for $1 \leq i \leq T$,*

$$e_i^{k+1} := f_i(w_{i+1}^k) - G_i^{k+1},\ \hat{e}_i^{k+1} := \nabla f_i(w_{i+1}^k) - J_i^{k+1}, \tag{45}$$

$$\hat{A}_{k,i} := f_i(w_{i+1}^{k+1}) - f_i(w_{i+1}^k) - \nabla f_i(w_{i+1}^k)^\top (w_{i+1}^{k+1} - w_{i+1}^k). \tag{46}$$

[3] Following the presentation in [36], 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 (6) with

$$w_i^{k+1} = (1-\tau_k)w_i^k + \tau_k G_i^{k+1} + (J_i^{k+1})^\top (w_{i+1}^{k+1} - w_{i+1}^k), \qquad 1 \le i \le T. \tag{44}$$

---

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

$$\begin{aligned}\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 &\le (1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + \dot{r}_i^{k+1}\\ &+ \left[8L_{f_i}^2 + L_{\nabla f_i}\|f_i(w_{i+1}^k) - w_i^k\| + \|\hat{e}_i{}^{k+1}\|^2\right]\|w_{i+1}^{k+1} - w_{i+1}^k\|^2, \end{aligned}\tag{47}$$

$$\begin{aligned}\dot{r}_i^{k+1} &:= 2\tau_k\langle e_i^{k+1}, \hat{A}_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k) + (\hat{e}_i{}^{k+1})^\top(w_{i+1}^{k+1} - w_{i+1}^k)\rangle\\ &+ 2\langle(\hat{e}_i{}^{k+1})^\top(w_{i+1}^{k+1} - w_{i+1}^k), \hat{A}_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k)\rangle. \end{aligned}\tag{48}$$

b) Furthermore, we have for $1 \le i \le T$, $\|w_i^{k+1} - w_i^k\|^2 \le$

$$\tau_k^2\left[2\|f_i(w_{i+1}^k) - w_i^k\|^2 + \|e_i^{k+1}\|^2 + \frac{2}{\tau_k^2}\|J_i^{k+1}\|^2\|w_{i+1}^{k+1} - w_{i+1}^k\|^2\right] + 2\ddot{r}_i^{k+1},$$

where $\ddot{r}_i^{k+1} := \tau_k\langle -e_i^{k+1}, \tau_k(f_i(w_{i+1}^k) - w_i^k) + (J_i^{k+1})^\top(w_{i+1}^{k+1} - w_{i+1}^k)\rangle$.

Proof. We first prove part a).

When $1 \le i < T$, by definition of $\hat{A}_{k,i}, \hat{e}_i{}^{k+1}, G_i^{k+1}, w_i^{k+1}$, and $\dot{r}_i^{k+1}$, we have

$$\begin{aligned}&\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2\\ =&\|\hat{A}_{k,i} + f_i(w_{i+1}^k) + \nabla f_i(w_{i+1}^k)^\top(w_{i+1}^{k+1} - w_{i+1}^k)\\ &\quad - (1-\tau_k)w_i^k - \tau_k G_i^{k+1} - (J_i^{k+1})^\top(w_{i+1}^{k+1} - w_{i+1}^k)\|^2\\ =&\|\hat{A}_{k,i} + (\widehat{e}_i{}^{k+1})^\top(w_{i+1}^{k+1} - w_{i+1}^k) + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k) + \tau_k e_i^{k+1}\|^2 \qquad (49)\\ =&\|(\widehat{e}_i{}^{k+1})^\top(w_{i+1}^{k+1} - w_{i+1}^k)\|^2 + \|\hat{A}_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k)\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + \dot{r}_i^{k+1}\\ \le&\|\hat{A}_{k,i} + (1-\tau_k)(f_i(w_{i+1}^k) - w_i^k)\|^2 + \tau_k^2\|e_i^{k+1}\|^2 + \dot{r}_i^{k+1} + \|\widehat{e}_i{}^{k+1}\|^2\|w_{i+1}^{k+1} - w_{i+1}^k\|^2\\ \le&(1-\tau_k)\|f_i(w_{i+1}^k) - w_i^k\|^2 + \|\hat{A}_{k,i}\|^2 + 2(1-\tau_k)\langle\hat{A}_{k,i}, f_i(w_{i+1}^k) - w_i^k\rangle + \tau_k^2\|e_i^{k+1}\|^2\\ &\quad + \dot{r}_i^{k+1} + \|\widehat{e}_i{}^{k+1}\|^2\|w_{i+1}^{k+1} - w_{i+1}^k\|^2. \qquad (50)\end{aligned}$$

Now, noting that under Assumption 2.1, we have

$$\|\hat{A}_{k,i}\| \le \frac{1}{2}\min\left\{4L_{f_i}\|w_{i+1}^{k+1} - w_{i+1}^k\|, L_{\nabla f_i}\|w_{i+1}^{k+1} - w_{i+1}^k\|^2\right\}, \tag{51}$$

and using Cauchy–Schwarz inequality in (50), we obtain (47).

To show part b), noting definition of (44) and (45), Cauchy-Schwartz and Young's inequality, for $1 \le i \le T$,

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

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

■

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 3.2. *Under Assumption 2.1, Assumption 2.2, for $1 \le i \le T$, and with the choice of $\tau_0 = 1$ (for simplicity), we have*

$$
\mathbb{E}[\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 | \mathscr{F}_k] \le \sigma_{G_i}^2 + (4L_{f_i}^2 + \hat{\sigma}_{J_i}^2) c_{i+1}, \tag{52}
$$

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

where, for $1 \le i \le T$,

$$
c_i := 3\sigma_{G_T}^2 + 2(4L_{f_T}^2 + \hat{\sigma}_{J_T}^2 + \sigma_{J_T}^2) c_{i+1}, \quad \textit{with} \quad c_{T+1} := \left(\prod_{i=1}^{T} \sigma_{J_i}^2\right) \beta^{-2}. \tag{54}
$$

Proof. Recall the definitions of $\hat{A}_{k,i}, e_i^{k+1}, \hat{e}_i^{k+1}$ and, for $1 \le i \le T$, define

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

Then, by (49), for $1 \le i \le T$, we have

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

which together with the convexity of $\|\cdot\|^2$, imply that

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

Moreover, we have

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

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

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

where the second inequality follows from (51). Hence, noting (27), $w_{T+1}^k = x^k$, we have

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

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

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

Moreover, by Lemma 3.1.b) and under Assumption 2.2, we have $\mathbb{E}[\|w_{i+1}^{k+1} - w_i^k\|^2 | \mathscr{F}_k] \leq$

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

implying that

$$\mathbb{E}[\|w_T^{k+1} - w_T^k\|^2 | \mathscr{F}_k] \leq \tau_k^2 \left[3\sigma_{G_T}^2 + 2(4L_{f_T}^2 + \hat{\sigma}_{J_T}^2 + \sigma_{J_T}^2)\left(\prod_{i=1}^{T} \sigma_{J_i}^2\right) \beta^{-2}\right]. \tag{62}$$

This completes the proof of (52) and (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 $\mathbb{E}[\|w_{i+1}^{k+1} - w_{i+1}^k\|^2 | \mathscr{F}_k] \leq c_{i+1}\tau_k^2$ for some $1 \leq i < T$. Hence, by (58) and under Assumption 2.2, we have

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

which together with Lemma 2.8, imply that

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

Thus, by (61), we obtain

$$\mathbb{E}[\|w_i^{k+1} - w_i^k\|^2 | \mathscr{F}_k] \leq \tau_k^2[3\sigma_{G_i}^2 + 2(4L_{f_i}^2 + \hat{\sigma}_{J_i}^2 + \sigma_{J_i}^2)c_{i+1}],$$

which together with the definition of c_i in (54), complete the proof. ■

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

Lemma 3.3. *Recall the definition of the merit function in (16). Define $w^k := (w_1^k, \ldots, w_T^k)$ for $k \geq 0$. Let $\{x^k, z^k, u^k, w_1^k, \ldots, w_T^k\}_{k\geq 0}$ be the sequence generated by Algorithm 2. Suppose that*

$$\gamma_1 \geq \lambda > 0, \quad \beta > \lambda, \quad (\beta - \lambda)(\gamma_j - \lambda) \geq 4TC_j^2, \qquad j \in \{2, \ldots, T\}, \tag{63}$$

where C_j's are defined in Lemma 2.4. Then, under Assumption 2.1 and Assumption 2.2, we have

$$\begin{aligned} &\lambda \sum_{k=0}^{N-1} \tau_k \left[\|d^k\|^2 + \sum_{i=1}^{T-1} \|f_i(w_{i+1}^k) - w_i^k\|^2 + \|f_T(x^k) - w_T^k\|^2\right] \\ &\leq W_\gamma(x^0, z^0, w^0) + \sum_{k=0}^{N-1} \hat{R}^{k+1}, \end{aligned} \tag{64}$$

where, for any $k \geq 0$,

$$\hat{R}^{k+1} := \left(\sum_{i=1}^{T} \gamma_i \hat{r}_i^{k+1}\right) + \frac{\tau_k^2}{2}\left[(L_{\nabla F} + L_{\nabla\eta}\right] + \tau_k\langle d^k, \Delta_k\rangle + \frac{L_{\nabla\eta}}{2}\|z^{k+1} - z^k\|^2, \tag{65}$$

$$\hat{r}_i^{k+1} = \left[8L_{f_i}^2 + L_{\nabla f_i}\|f_i(w_{i+1}^k) - w_i^k\| + \|\hat{e}_i{}^{k+1}\|^2\right] \|w_{i+1}^{k+1} - w_{i+1}^k\|^2$$

$$+ \tau_k^2 \|e_i^{k+1}\|^2 + \dot{r}_i^{k+1},$$

and Δ_k *and* $\dot{r}_i^{k+1}$ *are, respectively, defined in (20) and (48). Furthermore, notice that* (63) *is satisfied, when we pick*

$$\gamma_1 = \lambda = \sqrt{T}, \qquad \beta = 2\sqrt{T}, \qquad \gamma_j = \sqrt{T}(1 + 4C_j^2) \qquad 2 \leq j \leq T. \tag{66}$$

Proof. Noting Lemma 3.1 and definition of $\hat{r}_i^{k+1}$, we have, $\forall i \in \{1, 2, \dots, T\}$,

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

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

$$\begin{aligned}
&W_\gamma(x^{k+1}, z^{k+1}, w^{k+1}) - W_\gamma(x^k, z^k, w^k) \\
\leq &-\beta\tau_k \|d^k\|^2 + \sum_{j=2}^{T-1} \tau_k C_j \|d^k\| \|f_j(w_{j+1}^k) - w_j^k\| + \tau_k C_T \|d^k\| \|f_T(x^k) - w_T^k\| \\
&- \sum_{j=1}^{T-1} \gamma_j \tau_k \|f_j(w_{j+1}^k) - w_j^k\|^2 - \gamma_T \tau_k \|f_T(x^k) - w_T^k\|^2 + \hat{R}^{k+1}
\end{aligned}$$

Now if condition (63) holds, for any $i \in \{1, \dots, T\}$, we have

$$\begin{aligned}
&-\frac{\beta}{T}\|d^k\|^2 + C_i \|d^k\| \|f_i(w_{i+1}^k) - w_i^k\| - \gamma_i \|f_i(w_{i+1}^k) - w_i^k\|^2 \\
\leq &-\lambda\Big[\frac{1}{T}\|d^k\|^2 + \|f_i(w_{i+1}^k) - w_i^k\|^2\Big].
\end{aligned}$$

Combining the above inequalities, we obtain

$$\begin{aligned}
&W_\gamma(x^{k+1}, z^{k+1}, w^{k+1}) - W_\gamma(x^k, z^k, w^k) \\
\leq &-\lambda\tau_k \Big[\|d^k\|^2 + \sum_{j=1}^{T-1} \|f_j(w_{j+1}^k) - w_j^k\|^2 + \|f_T(x^k) - w_T^k\|^2\Big] + \hat{R}^{k+1}.
\end{aligned}$$

Thus, by summing up the above inequalities and re-arranging the terms, we obtain (64). Finally, it is easy to see that (63) holds, by picking the parameters as in (66). ∎

In the next result, we show the error terms in the right hand side of (64) is bounded in the order of $\sum_{k=1}^N \tau_k^2$ in expectation.

Proposition 3.1. *Let* $\hat{R}^k$ *be defined in (65). The, under Assumption 2.2, we have*

$$\mathbb{E}[\hat{R}^{k+1} | \mathscr{F}_k] \leq \hat{\sigma}^2 \tau_k^2, \qquad \forall k \geq 1,$$

where

$$\begin{aligned}
\hat{\sigma}^2 := &\sum_{i=1}^T \gamma_i \left(\Big[8L_{f_i}^2 + L_{\nabla f_i}\sqrt{\sigma_{G_i}^2 + (4L_{f_i}^2 + \hat{\sigma}_{J_i}^2)c_{i+1}} + \hat{\sigma}_{J_i}^2\Big] c_{i+1} + \sigma_{G_i}^2 \right) \\
&+ \frac{1}{2\beta^2}\left(\prod_{i=1}^T \sigma_{J_i}^2\right)[(1 + 4\beta^2)L_{\nabla\eta} + L_{\nabla F}].
\end{aligned} \tag{67}$$

Proof. Under Assumption 2.2, we have, for $1 \le i \le T$,

$$\mathbb{E}[\Delta_k | \mathscr{F}_k] = 0, \quad \mathbb{E}[\dot{r}_i^{k+1} | \mathscr{F}_k] = 0, \quad \mathbb{E}[\|\hat{e}_i^{k+1}\|^2 | \mathscr{F}_k] \le \sigma_{G_i}^2, \quad \mathbb{E}[\|e_i^{k+1}\|^2 | \mathscr{F}_k] \le \hat{\sigma}_{J_i}^2.$$

Moreover, by Lemma 3.2 and Holder's inequality, we have $\mathbb{E}[\|w_i^{k+1} - w_i^k\|^2 | \mathscr{F}_k] \le c_i \, \tau_k^2$ and

$$\begin{aligned}
&\mathbb{E}[\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\| \|w_i^{k+1} - w_i^k\|^2 | \mathscr{F}_k] \\
\le &\left(\mathbb{E}[\|f_i(w_{i+1}^{k+1}) - w_i^{k+1}\|^2 | \mathscr{F}_k] \right)^{\frac{1}{2}} \mathbb{E}[\|w_i^{k+1} - w_i^k\|^2 | \mathscr{F}_k] \\
\le &c_i \sqrt{\sigma_{G_i}^2 + (4L_{f_i}^2 + \hat{\sigma}_{J_i}^2) c_{i+1} \tau_k^2}
\end{aligned}$$

The result then follows by noting the definition of $\hat{\sigma}^2$ in (67) ■

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 2.2 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 τ_k is set as in (39), the results of part c) of (35) also hold with $\hat{\sigma}^2$ replacing σ^2.

Proof. The proof follows from the same arguments as in the proof of (35) by noticing (64), 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 $\mathcal{O}(c^T T^4/\epsilon^4) = \mathcal{O}_T(1/\epsilon^4)$ which is better in the order of magnitude than the complexity bound of Algorithm 1. Furthermore, this bound matches the complexity bound obtained in [20] for the two-level composite problem which in turn is in the same order for single-level smooth stochastic optimization. Finally, it is worth noting that this complexity bound for Algorithm 2 is obtained without assuming boundedness of the feasible set or any dependence of the parameter β on Lipschitz constants. Indeed, β can be set to any positive number in the order of $\mathcal{O}(\sqrt{T})$ due to (66), and τ_k depends only on the total number of iterations N due to (39). This makes Algorithm 2 parameter-free and easy to implement.*

4 Numerical Experiments

In this section, we provide numerical results for the risk-averse stochastic optimization problem introduced in Section 1.1.1. The link function g is set to be the square function and $U(x, \xi) := -(b - g(a^\top x))^2$. In this case, (2) becomes a non-convex stochastic three-level composition optimization problem. For our experiments, we assume $a \in \mathbb{R}^d$ is a zero-mean Gaussian random vector with covariance matrix $\Sigma_{j,k} = 0.5 e^{-\frac{|j-k|}{d}}$, following [36]. Furthermore, ζ is a standard normal random variable. The true parameter $x^* \in \mathbb{R}^d$ is drawn from a standard Gaussian distribution and fixed.

We compare our Algorithm 2 with the accelerated T-level stochastic compositional gradient descent (a-TSCGD) from [36]. For our algorithm, the parameter τ_k was set at $c/\sqrt{N}$ (with c being 0.5, 1 and 1.5) and the step-size β was set to 4 (as it is close to $2\sqrt{T}$ and empirically worked the best). The parameters for a-TSCGD were set according to the suggestion from [36]. We estimated the expected gradient size empirically, based on an independent dataset of size 10,000, so as to reduce

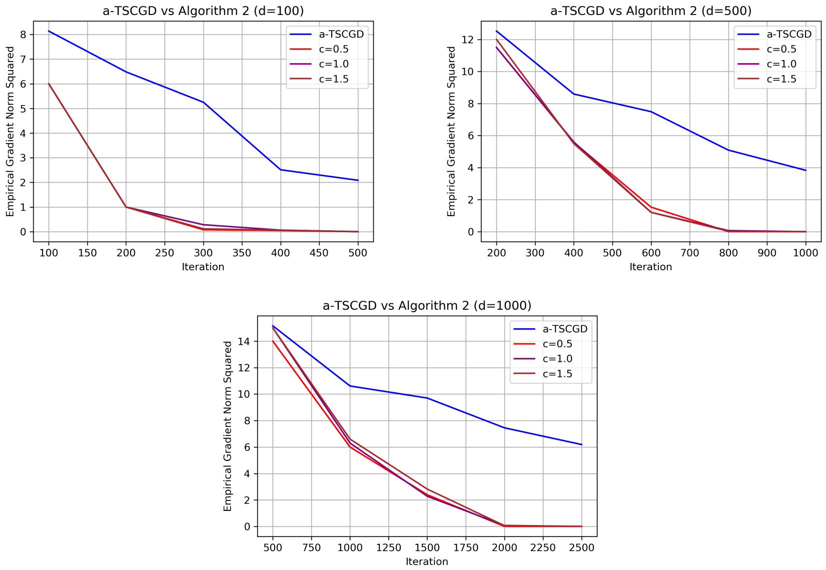

Figure 1: Comparison between Algorithm 2 and a-TSCGD [36]: Empirical gradient size squares versus iterations for $d = 100$ (top left), $d = 500$ (bottom) and $d = 1000$ (top right). Here, c refers to the choice of numerator in the tuning parameter τ_k, given by $\tau_k =: c/\sqrt{N}$.

any fluctuations in this estimation process. Furthermore, we reported the average over 100 Monte-Carlo trails, to reduce the fluctuations over the data generating process. Figure 4 plots the empirical gradient norm squared as a function of iteration, for the values of dimension $d \in \{100, 500, 1000\}$. As can be seem from the plots, Algorithm 2 outperforms a-TSCGD from [36] numerically as well. Furthermore, our algorithm is almost insensitive to the choice of c in the definition of τ_k.

5 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 $\mathcal{O}_T(1/\epsilon^6)$ for finding an ϵ-stationary point of the multi-level composite problem. By modifying this algorithm with a linearization technique, we show that we can improve the sample complexity to $\mathcal{O}_T(1/\epsilon^4)$ which seems to be unimprovable even for single-level stochastic optimization problems, without further assumptions [2, 11]. For future work, it would be 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 [1, 10, 25, 27, 37] for the standard stochastic gradient algorithm when $T = 1$.

Acknowledgment

We would like to thank Prof. Andrzej Ruszczyński for helpful comments on the related works. Krishnakumar Balasubramanian was supported in part by UC Davis CeDAR (Center for Data Science and Artificial Intelligence Research) Innovative Data Science Seed Funding Program. Saeed Ghadimi was supported in part by an NSERC Discovery Grant.

References

[1] 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.

[2] 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.

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

[4] 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.

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

[6] Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.

[7] Weilin Cong, Rana Forsati, Mahmut Kandemir, and Mehrdad Mahdavi. Minimal variance sampling with provable guarantees for fast training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1393–1403, 2020.

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

[9] 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.

[10] 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.

[11] 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.

[12] D. Drusvyatskiy and A.S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):693–1050, 2018.

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

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

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

[16] 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.

[17] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for constrained nonconvex stochastic programming. *Mathematical Programming*, 155(1-2):267–305, 2016.

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

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

[20] 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.

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

[22] Y. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.

[23] Y. Nesterov. Gradient methods for minimizing composite objective functions. *Mathematical Programming*, 140(1):125–161, 2013.

[24] 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.

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

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

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

[28] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*. Number 12. Cambridge university press, 2003.

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

[30] Andrzej Ruszczyński. Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization. *Optimization Letters*, pages 1–11, 2020.

[31] Andrzej Ruszczynski. A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization. *SIAM Journal on Control and Optimization*, 59(3):2301–2320, 2021.

[32] Andrzej Ruszczyński and Alexander Shapiro. Optimization of convex risk functions. *Mathematics of operations research*, 31(3):433–452, 2006.

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

[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*, 2016.

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

[37] 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. *Advances in Neural Information Processing Systems (forthcoming)*, 2021.

[38] Junyu Zhang and Lin Xiao. Multilevel composite stochastic optimization via nested variance reduction. *SIAM Journal on Optimization*, 31(2):1131–1157, 2021.