SIAM J. OPTIM.
Vol. 29, No. 1, pp. 616–659

MULTILEVEL STOCHASTIC GRADIENT METHODS FOR NESTED COMPOSITION OPTIMIZATION*

SHUOGUANG YANG†, MENGDI WANG‡, AND ETHAN X. FANG§

Abstract. Stochastic gradient methods are scalable for solving large-scale optimization problems that involve empirical expectations of loss functions. Existing results mainly apply to optimization problems where the objectives are one- or two-level expectations. In this paper, we consider the multilevel composition optimization problem that involves compositions of multilevel component functions and nested expectations over a random path. This finds applications in risk-averse optimization and sequential planning. We propose a class of multilevel stochastic gradient methods that are motivated by the method of multitimescale stochastic approximation. First, we propose a basic T-level stochastic compositional gradient algorithm. Then we develop accelerated multilevel stochastic gradient methods by using an extrapolation-interpolation scheme to take advantage of the smoothness of individual component functions. When all component functions are smooth, we show that the convergence rate improves to $\mathcal{O}(n^{-4/(7+T)})$ for general objectives and $\mathcal{O}(n^{-4/(3+T)})$ for strongly convex objectives. We also provide almost sure convergence and rate of convergence results for nonconvex problems. The proposed methods and theoretical results are validated using numerical experiments.

Key words. stochastic gradient, stochastic optimization, convex optimization, sample complexity, simulation, statistical learning

AMS subject classifications. 90C15, 90C25, 90C06, 68W27

DOI. 10.1137/18M1164846

1. Introduction.

Over the past decade, stochastic gradient-type methods have attracted significant attention from various communities—such as mathematical programming, signal processing, and machine learning—mainly due to their practical efficiency in minimizing expected-value objective functions or empirical sums of a large number of loss functions [2, 5, 11, 12, 13, 15, 19, 23, 31]. They are particularly popular for tackling large-scale problems such as statistical estimation [7, 21], matrix and tensor factorization [9], and training deep neural networks [14, 30]. Stochastic gradient methods mainly apply to minimizing the expectation of a stochastic function, i.e.,

$$\min_x \mathbb{E}_\omega[f_\omega(x)],$$

where the expectation is taken over a random variable ω. Note that this problem involves one level of expectation.

In this paper, we study a richer class of stochastic optimization problems, which involve nested expectations over a sequence of random variables. In particular, we

*Received by the editors January 11, 2018; accepted for publication (in revised form) November 14, 2018; published electronically March 5, 2019.
http://www.siam.org/journals/siopt/29-1/M116484.html
Funding: Mengdi Wang was partially supported by NSF CAREER: CMMI-1653435. Ethan X. Fang was partially supported by National Institute on Drug Abuse: P50 DA039838. This paper's contents are solely the responsibility of the authors and do not represent the views of the funding organizations.

†Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027 (sy2614@columbia.edu).
‡Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 (mengdiw@princeton.edu).
§Department of Statistics, Pennsylvania State University, University Park, PA 16802 (xxf13@psu.edu).

consider the *T-level stochastic composition optimization problem*, given by

$$\min_{x\in\mathcal{X}} F(x) = \mathbb{E}_{\omega_1}\Big[f_{\omega_1}^{(1)}\Big(\mathbb{E}_{\omega_2}\big[f_{\omega_2}^{(2)}\big(\cdots\big(\mathbb{E}_{\omega_T}[f_{\omega_T}^{(T)}(x)]\big)\cdots\big)\big]\Big)\Big], \tag{1.1}$$

where $\mathcal{X}$ is a convex and closed set, $f_{\omega_j}^{(j)}(\cdot) : \mathbb{R}^{d_j} \mapsto \mathbb{R}^{d_{j-1}}$ for $j = 1, \ldots, T$ are continuous mappings, and $d_0 = 1$, i.e., $F(x)$ is a real-valued function. The nested composition structure provides a rich modeling tool for data analysis and decision-making applications. For instance, online principal component analysis and policy evaluation in reinforcement learning can be formulated into two-level stochastic composition optimization [16, 28]. In section 4, we illustrate the mean-deviation risk-averse optimization problem with an example from operations research. It can be formulated as a three-level compositional problem [1, 22].

In problem (1.1), for each $f_{\omega_j}^{(j)}$, we use the subscript ω_j to denote a random variable and use the superscript (j) to denote its level. We focus on situations where there exist deterministic functions $f^{(1)}, \ldots, f^{(T)}$ such that

$$f^{(j)}(x_j) = \mathbb{E}[f_{\omega_j}^{(j)}(x_j)|\omega_1, \ldots, \omega_{j-1}]$$

for all $j = 1, \ldots, T$ *with probability* 1 *(w.p.*1*).* We refer to $f^{(1)}, \ldots, f^{(T)}$ as *component functions.* However, these component functions are not explicitly known to us. Note that the multilevel random variables $\omega_1, \ldots, \omega_T$ are not necessarily independent of one another. When we sample from their joint distribution, we may generate a sample path $(\omega_1, \ldots, \omega_T)$ sequentially by sampling each ω_j conditioned on realizations at the previous levels $(\omega_1, \ldots, \omega_{j-1})$. Throughout this paper, we assume that the component functions $f^{(1)}, \ldots, f^{(T)}$ are continuous and that there exists at least one optimal solution x^* to problem (1.1). In some parts of our analysis, we require the overall objective function $F(x)$ be convex, but we *never* require that any individual component function $f_{\omega_j}^{(j)}(\cdot)$ be convex, linear, or monotone. We say that a function f is "smooth" if it has Lipschitz continuous gradients, and that it is "nonsmooth" otherwise.

Our goal is to solve the T-level stochastic composition optimization problem (1.1) by sampling multiple paths of $(\omega_1, \ldots, \omega_T)$. We are interested in scenarios where we do not have explicit knowledge of the expected-value component functions $f^{(j)}$. This often occurs when evaluating $f^{(j)}$ requires making expensive passes over large data sets, and in online learning applications where $f^{(j)}$ cannot be accurately calculated using finitely many samples. Instead of knowing the $f^{(j)}$'s, we suppose that there is a *sample oracle* ($\mathcal{SO}$) such that

- upon each query $(x \in \mathcal{X}, y_1 \in \mathbb{R}^{d_1}, \ldots, y_T \in \mathbb{R}^{d_T})$, the $\mathcal{SO}$ generates a sample path $(\omega_1, \ldots, \omega_T)$ independently from the query,
- the $\mathcal{SO}$ returns a vector $f_{\omega_T}^{(T)}(x) \in \mathbb{R}^{d_{T-1}}$ and a gradient/subgradient

$$\widetilde{\nabla} f_{\omega_T}^{(T)}(x) \in \mathbb{R}^{d_T \times d_{T-1}},$$

- the $\mathcal{SO}$ returns a vector $f_{\omega_j}^{(j)}(y_j) \in \mathbb{R}^{d_j}$ and a gradient $\nabla f_{\omega_j}^{(j)}(y_j) \in \mathbb{R}^{d_j \times d_{j-1}}$,
- the $\mathcal{SO}$ returns a gradient $\nabla f_{\omega_1}^{(1)}(y_1) \in \mathbb{R}^{d_1}$.

In the above, we denote by $\widetilde{\nabla} f_{\omega_T}^{(T)}(x)$ a gradient/subgradient, which is to be specified in context. Let us emphasize that this $\mathcal{SO}$ does *not* return unbiased first-order information regarding the overall objective function. The $\mathcal{SO}$ can be viewed as a *componentwise stochastic first-order oracle* that returns noisy first-order information

for individual component functions $f^{(j)}$. Detailed assumptions on the $\mathcal{SO}$ will be specified later.

One might attempt to apply the sample average approximation (SAA) method to attack the multilevel expectation problem (1.1). However, replacing the nested expectations with empirical averages will not solve the optimization problem. It will reduce one problem with expectations to another one with empirical expectations. However, the two problems share similar structures and the latter problem is not necessarily easier to solve. What we need is an implementable algorithm that computes the optimal solution by iteratively querying the $\mathcal{SO}$ and making efficient updates.

One may alternatively use some version of the gradient method or stochastic gradient method. The stochastic gradient method will not work automatically. The main challenge is that we do not have access to the unbiased sample gradient of F due to the multilevel nested expectations. To see this, let us consider the case when each $f^{(j)}$ is differentiable, and apply the chain rule to get

$$\nabla F(x) = \nabla f^{(T)}(x) \nabla f^{(T-1)}\left(f^T(x)\right) \cdots \nabla f^{(1)}\left(f^{(2)} \circ \cdots \circ f^{(T)}(x)\right).$$

For a given $x \in \mathcal{S}$ and a given sample path $(\omega_1, \ldots, \omega_T)$, one may formulate an unbiased estimate of $\nabla F(x)$ as

$$\nabla f^{(T)}_{\omega_T}(x) \nabla f^{(T-1)}_{\omega_{T-1}}\left(f^{(T)}(x)\right) \cdots \nabla f^{(1)}_{\omega_1}\left(f^{(2)} \circ \cdots \circ f^{(T)}(x)\right),$$

which unfortunately cannot be calculated by calling the $\mathcal{SO}$ once (or even finitely many times). This is because computing the preceding unbiased gradient sample requires the $\mathcal{SO}$ to be queried at values $f^{(T)}(x), f^{(T-1)} \circ f^{(T)}(x), \ldots, f^{(2)} \circ \cdots \circ f^{(T)}(x)$, which are unfortunately not known. As a result, the nested composition structure induces substantial bias in the sample gradients for F as long as $T \geq 2$. In contrast, when $T = 1$, the objective function is linear in the distribution of the random variable ω. For problems with $T \geq 2$, the nonlinear composition between expectations and component functions creates an objective function that is highly nonlinear with respect to the joint probability distribution of $\omega_1, \ldots, \omega_T$. A graphical illustration of the level of difficulty for dealing with multilevel composition optimization is given in Figure 1. We can view the optimization problem (1.1) under the $\mathcal{SO}$ as a form of estimation problem, in which we want to estimate the optimal solution x^* by taking independent sample paths. We can see that the nonlinear composition makes this estimation/optimization problem fundamentally challenging.

Existing work on stochastic composition optimization can be traced back to [8], which considered the two-level problem. In [8, section 6.7], a two-timescale stochastic approximation scheme was proposed, and its almost sure convergence was established without rate analysis. Recently, Wang, Fang, and Liu [27] developed a general class of stochastic compositional gradient descent (SCGD) methods for two-level problems and established convergence rate results under various assumptions. Wang, Liu, and Fang [29] developed an accelerated stochastic compositional proximal gradient (ASC-PG) method for the two-level problem and proved faster convergence in some cases. Lian, Wang, and Liu [17] considered a special case of the two-level problem where each expectation takes the form of a finite sum of loss functions, and developed variance-reduced versions of the compositional gradient methods. As for the general T-level problem, to the best of our knowledge, all existing results only apply to the case when $T = 1, 2$. Multilevel stochastic composition optimization remains largely an open problem.

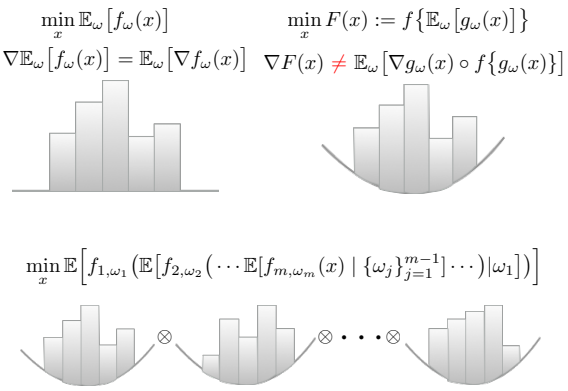

FIG. 1. *In one-level stochastic optimization, the objective function is linear in the probability distribution of ω. In multilevel stochastic composition optimization, the objective is no longer linear in the joint probability distribution of the random variables $(\omega_1, \ldots, \omega_m)$, making the problem fundamentally harder.*

In this paper, we develop sampling-based algorithms and complexity theory for the T-level stochastic compositional problem (1.1). We draw motivation from the optimality conditions of problem (1.1). In particular, we expand the first-order condition into a system of variational equalities and inequalities by introducing auxiliary variables that correspond to a sequence of *value functions* at the optimal solution, i.e., tail compositions of the component functions. Our first attempt is a basic multitimescale stochastic approximation iteration to solve this system.

Furthermore, we develop accelerated multilevel stochastic gradient methods. The accelerated algorithms apply to "smooth" compositional problems and take advantage of the smoothness of individual component functions $f^{(j)}$. An *extrapolation-interpolation* scheme is used to balance the bias-variance tradeoff in approximating each value function. We establish its almost sure convergence using a T-element supermartingale argument for both convex and nonconvex problems. The accelerated updates for the auxiliary variables can be viewed as first-order running approximations of the true values, while the basic method without acceleration uses zeroth-order running approximations. As a result, the accelerated updates are more accurate, and thus the overall convergence rate is improved. In the case when all component functions are smooth, we improve the convergence rate to $\mathcal{O}(n^{-4/(7+T)})$ for convex objective functions and $\mathcal{O}(n^{-4/(3+T)})$ for strongly convex ones. We also obtain convergence and rate of convergence results for nonconvex problems. Table 1 summarizes our results and compares them with the best known ones for the single- and two-level stochastic composition optimization problems [10, 20, 24, 27, 29]. We also provide numerical experiments with a risk-averse regression problem. The numerical results validate our theory.

To the best of our knowledge, this paper is the first to propose the use of multilevel stochastic gradient methods for the composition optimization problem (1.1), where we establish almost sure convergence results and obtain fast convergence rates. For the case when $T = 1$, our results match the best known sample complexity upper and lower bounds. For the case when $T = 2$, our results improve the convergence rate from $\mathcal{O}(n^{-2/9})$ of the a-SCGD in [27] to $\mathcal{O}(n^{-2/5})$. With the additional assumption that the inner-level function $f^{(T)}$ in (1.1) has Lipschitz continuous gradients, we obtain a convergence rate $\mathcal{O}(n^{-4/9})$ for two-level problems, which matches the state-of-art

TABLE 1

Best-known n-sample error bounds for solving multilevel stochastic composition optimization. These bounds are achieved by stochastic gradient-type methods, so they are also n-iteration error bounds. We say the compositional problem is "smooth" if all the component functions have Lipschitz continuous gradients. We use [∗] to denote the current paper.

		Nonconvex	Convex	Strongly convex
1-level		$\mathcal{O}(n^{-1/2})$ [10]	$\mathcal{O}(n^{-1/2})$ [24]	$\mathcal{O}(n^{-1})$ [20]
2-level	Smooth	$\mathcal{O}(n^{-4/9})$ [29]	$\mathcal{O}(n^{-4/9})$ [29]	$\mathcal{O}(n^{-4/5})$ [29]
	Nonsmooth	$\mathcal{O}(n^{-1/4})$ [27]	$\mathcal{O}(n^{-1/4})$ [27]	NA
3-level	Smooth	$\mathcal{O}(n^{-2/5})$ [∗]	$\mathcal{O}(n^{-2/5})$ [∗]	$\mathcal{O}(n^{-2/3})$ [∗]
T-level	Smooth	$\mathcal{O}(n^{-4/(7+T)})$ [∗]	$\mathcal{O}(n^{-4/(7+T)})$ [∗]	$\mathcal{O}(n^{-4/(3+T)})$ [∗]

result achieved by ASC-PG in [29]. A natural further question is how big the hidden constants in these error bounds are. In the case when $T = 1$, the hidden constant is merely determined by the variance of stochastic gradients and the condition number, which can be derived in straightforward way by analyzing a telescoping sum [10, 20]. However, when $T > 1$, the hidden constants depend on a tedious formula involving sums and products of multilevel variances and Lipschitz continuity constants. Within the scope of the current paper, we focus on the dominating order of the error bounds, leaving the constants unspecified. For the cases when $T \geq 3$, our results fill the open gaps and provide the first few sample complexity benchmarks.

Our proposed methods, being optimization algorithms, can be viewed as updating an online estimator by drawing samples from a data stream. Let us evaluate their performance from a statistical perspective. For comparison, the most related result is given by [6], which uses an sample average approximation approach to solve the T-level compositional problem where the multilevel random variables are independently identically distributed random variables. For this case, Dentcheva, Penev, and Ruszczyński [6] proved that the batch method achieves an error rate of $\mathcal{O}(1/\sqrt{n})$, which is obviously statistically nonimprovable. More remarkably, the error bound obtained in [6] is independent of T. In this paper, our result for the smooth convex case is $\mathcal{O}(n^{-4/(7+T)})$, which deteriorates as the number of levels T increases. There are two possible explanations. First, the problem considered in this paper is slightly more general than that of [6] because we do not assume the independence of random variables at different levels. Second, the proposed algorithms use multitimescale updates so that certain random samples are given less weight than others, while the batch approach treats all samples equally. The use of multitimescale updates, which is critical for the proposed online method, may have resulted in inefficient use of data and slowed down the convergence. It remains open whether there exists an online algorithm that can achieve the same rate of convergence as the batch method. We hope the developments of this paper will pave the way toward a more complete understanding of the complexity of multilevel composition optimization.

Paper organization. Section 2 gives a basic algorithm based on multitimescale stochastic approximation and establishes its convergence. Section 3 develops accelerated versions of the algorithm and shows that they achieve faster convergence for smooth problems. Section 4 illustrates one motivating application in operations research and gives numerical experiments.

Notation and definitions. For $x \in \mathbb{R}^n$, we denote by x' its transpose, and by $\|x\|$ its Euclidean norm (i.e., $\|x\| = \sqrt{x'x}$). For two sequences $\{x_k\}$ and $\{y_k\}$, we write

$x_k = \mathcal{O}(y_k)$ if there exists a constant $c > 0$ such that $\|x_k\| \le c\|y_k\|$ for each k. We denote by $\mathbb{I}^{value}_{condition}$ the indicator function, which returns "value" if the "condition" is satisfied, and 0 otherwise. We denote by F^* the optimal objective function value for (1.1), and denote by $\mathcal{X}^*$ the set of optimal solutions. For a set $\mathcal{X} \subset \mathbb{R}^n$ and a vector $y \in \mathbb{R}^n$, we denote by $\Pi_{\mathcal{X}}\{y\} = \operatorname{argmin}_{x\in\mathcal{X}} \|y - x\|^2$ the Euclidean projection of y on $\mathcal{X}$, where the minimization is always uniquely attained if $\mathcal{X}$ is nonempty, convex, and closed. For a function $f(x)$, we denote by $\nabla f(x)$ its gradient at x if f is differentiable, denote by $\partial f(x)$ its subdifferential at x, and denote by $\widetilde{\nabla} f(x)$ some noisy estimate of the gradient/subgradient of f at x. We denote "with probability 1" by "w.p.1."

2. A basic algorithm based on multitimescale stochastic approximation. We start by writing down the optimality condition of problem (1.1) (assuming that the problem is convex):

$$\nabla F(x^*)'(x - x^*) \ge 0 \qquad \forall x \in \mathcal{X},$$

where

$$\nabla F(x) = \nabla f^{(T)}(x) \cdot \nabla f^{(T-1)}\big(f^T(x)\big) \cdots \nabla f^{(1)}\left(f^{(2)} \circ \cdots \circ f^{(T)}(x)\right).$$

However, this optimality condition is not easy to work with. As we have discussed in section 1, the chain rule makes obtaining unbiased samples of $\nabla F(x)$ difficult. Let us rewrite the optimality condition as follows:

$$\begin{aligned}
&\left(\nabla f^{(T)}(x)\nabla f^{(T-1)}(y^{T-1}) \cdots \nabla f^{(1)}(y^{(1)})\right)' (x - x^*) \ge 0 \qquad \forall x \in \mathcal{X},\\
&y^{(T-1)} = f^{(T)}(x),\\
&y^{(T-2)} = f^{(T-1)}(y^{(T-1)}) = f^{(T-1)} \circ f^{(T)}(x),\\
&y^{(1)} = f^{(2)}(y^{(2)}) = f^{(2)} \circ \cdots \circ f^{(T)}(x).
\end{aligned}$$

We refer to $f^{(j)} \circ \cdots \circ f^{(T)}(x)$, $j = 1, \ldots, T-1$, as the *value functions*, i.e., tail compositions of multilevel component functions. By introducing the auxiliary variables $y^{(j)}$'s to represent the value functions, we can decouple the chain product. Now, for a given $(x, y^{(1)}, \ldots, y^{(T-1)})$, our sampling oracle allows us to get unbiased estimates for all the quantities in the preceding system of optimality conditions.

2.1. A *T*-level stochastic gradient method. Motivated by the system of optimality conditions, we develop our first algorithm—a multitimescale approximation iteration. It is also a generalization of the basic-SCGD in [27], which applies only to two-level problems. Our algorithm runs iteratively. Denote by k the iteration counter. A key ingredient of our algorithm the introduction of auxiliary variables $y_k^{(j)}$, defined recursively, as running estimates for the value functions

$$\mathbb{E}_{\omega_{j,k}}[f^{(j)}_{\omega_{j,k}}(y_k^{(j+1)})|\omega_{1,k}, \ldots, \omega_{j-1,k}],$$

where $j = 1, \ldots, T-1$, and $x_k = y_k^{(T)}$. At the kth iteration, we update the current solution x_k by using a quasi-stochastic gradient step given by

$$x_{k+1} = \Pi_{\mathcal{X}}\Big\{x_k - \alpha_k \widetilde{\nabla} f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)}) \cdots \nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})\Big\}.$$

Algorithm 1 Basic stochastic compositional gradient descent (T-SCGD).

Input: $x_0 \in \mathbb{R}^{d_T}$, $y_0^{(j)} \in \mathbb{R}^{d_j}$ for $j = T-1, \ldots, 1$, $\mathcal{SO}$, K, step sizes $\{\alpha_k\}_{k=0}^K$, $\{\beta_{j,k}\}_{k=0}^K$ for $j = T-1, \ldots, 1$.
Output: The sequence $\{x_k\}_{k=0}^K$.
for $k = 0, 1, 2, \ldots, K$ **do**
  Query the $\mathcal{SO}$ for the sample values of $f^{(T)}, \ldots, f^{(1)}$ at $(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})$; obtain the sample gradients/subgradients $\widetilde{\nabla} f^{(T)}_{\omega_{T,k}}(x_k), \nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)}), \ldots, \nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})$.
  Update the main iterate by

$$x_{k+1} = \Pi_{\mathcal{X}}\Big\{x_k - \alpha_k \widetilde{\nabla} f^{(T)}_{\omega_{T,k}}(x_k) \nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)}) \cdots \nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})\Big\}.$$

  Query the $\mathcal{SO}$ for the sample value of $f^{(T)}(\cdot)$ at x_{k+1}; obtain $f^{(T)}_{\omega_{T,k+1}}(x_{k+1})$.
  Update $y_k^{(T-1)}$ by

$$y_{k+1}^{(T-1)} = (1 - \beta_{T-1,k}) y_k^{(T-1)} + \beta_{T-1,k} f^{(T)}_{\omega_{T,k+1}}(x_{k+1}).$$

  for $j = T-2, \ldots, 1$ **do**
    Query the $\mathcal{SO}$ for the sample value of $f^{(j)}$ at $y_{k+1}^{(j)}$; obtain $f^{(j)}_{\omega_{j,k+1}}(y_{k+1}^{(j)})$.
    Update

$$y_{k+1}^{(j)} = (1 - \beta_{j,k}) y_k^{(j)} + \beta_{j,k} f^{(j+1)}_{\omega_{j+1,k+1}}(y_{k+1}^{(j+1)}).$$

  end for
end for

Then, we update the auxiliary variables $y_k^{(j)}$ by taking a weighted average between the previous values and the new samples returned by the $\mathcal{SO}$, i.e., for $j = T-1$, $T-2, \ldots, 1$,

$$y_{k+1}^{(j)} = (1 - \beta_{j,k}) y_k^{(j)} + \beta_{j,k} f^{(j+1)}_{\omega_{j+1,k+1}}(y_{k+1}^{(j+1)}), \tag{2.1}$$

where $\omega_{j,k}$ denotes the realization of the jth level random variable at the kth iteration, and the $\beta_{j,k}$'s are prespecified step sizes. We refer to this update for $y_k^{(j)}$ as a *basic update step*. Letting $y_k^{(T)} = x_k$ and $\alpha_k = \beta_{T,k}$ to simplify the notation, we refer to the preceding iteration as the basic T-level stochastic compositional gradient descent (T-SCGD) method and summarize it in Algorithm 1. Note that we choose the step sizes such that $\beta_{j+1,k}/\beta_{j,k} \to 0$ as $k \to \infty$ for all j, in order to control and balance the convergence speed for each auxiliary variable.

To analyze the convergence of the algorithm, we impose the following assumptions on the smoothness and bounded second-order moments for the stochastic component functions.

Assumption 2.1. Let $C_1, C_2, \ldots, C_T$, $V_1, \ldots, V_T$, and $L_2, L_3, \ldots, L_T$ be positive scalars.

(i) The outer functions $f^{(T-1)}, f^{(T-2)}, \ldots, f^{(1)}$ are continuously differentiable, the inner function $f^{(T)}$ is continuous, the feasible set $\mathcal{X}$ is closed and convex, and there exists at least one optimal solution x^* to problem (1.1).

(ii) The sample paths $(\omega_{1,0}, \omega_{2,0}, \ldots, \omega_{T,0})$, $(\omega_{1,1}, \omega_{2,1}, \ldots, \omega_{T,1}), \ldots, (\omega_{1,k}, \omega_{2,k}, \ldots, \omega_{T,k})$ are independent across k and satisfy, w.p.1,

$$\mathbb{E}[f^{(j)}_{\omega_{j,0}}(x_j) | \omega_{1,0}, \ldots, \omega_{j-1,0}] = f^{(j)}(x_j) \qquad \forall x_j \in \mathbb{R}^{d_j} \text{ for } j = 1, \ldots, T,$$

and

$$\mathbb{E}[\widetilde{\nabla} F_{\omega_0}(x)] \in \partial F(x)$$

for all $x \in \mathcal{X}$, where

$$\widetilde{\nabla} F_{\omega_0}(x) \equiv \widetilde{\nabla} f^{(T)}_{\omega_{T,0}}(x) \nabla f^{(T-1)}_{\omega_{T-1,0}}(f^{(T)}(x)) \cdots \nabla f^{(1)}_{\omega_{1,0}}(f^{(2)} \circ \cdots \circ f^{(T)}(x)).$$

(iii) The function $f^{(T)}(\cdot)$ is Lipschitz continuous with parameter C_T, and the samples $f^{(T)}_{\omega_{T,0}}(\cdot)$, $\widetilde{\nabla} f^{(T)}_{\omega_{T,0}}(\cdot)$ have bounded second-order moments such that, w.p.1,

$$\mathbb{E}\big[\|\widetilde{\nabla} f^{(T)}_{\omega_{T,0}}(x)\|^2 | \omega_{T-1,0}, \ldots, \omega_{1,0}\big] \le C_T,$$
$$\mathbb{E}\big[\|f^{(T)}_{\omega_{T,0}}(x) - f^{(T)}(x)\|^2 | \omega_{T-1,0}, \ldots, \omega_{1,0}\big] \le V_T$$

for all $x \in \mathcal{X}$.

(iv) For $j = 1, \ldots, T-1$, the functions $f^{(j)}(\cdot)$ and $f^{(j)}_{\omega_{j,0}}(\cdot)$ have L_j-Lipschitz continuous gradients such that, w.p.1,

$$\mathbb{E}\big[\|\nabla f^{(j)}_{\omega_{j,0}}(x_j)\|^2 | \omega_{j-1,0}, \ldots, \omega_{1,0}\big] \le C_j,$$
$$\mathbb{E}\big[\|f^{(j)}_{\omega_{j,0}}(x_j) - f^{(j)}(x_j)\|^2 | \omega_{j-1,0}, \ldots, \omega_{1,0}\big] \le V_j,$$

and

$$\|\nabla f^{(j)}_{\omega_{j,0}}(x_j) - \nabla f^{(j)}_{\omega_{j,0}}(\bar{x}_j)\| \le L_j \|x_j - \bar{x}_j\|$$

for all $x_j, \bar{x}_j \in \mathbb{R}^{d_j}$.

In some parts of the analysis, we also assume that the overall objective is sufficiently smooth, as follows.

Assumption 2.2. The function $F(x)$ has Lipschitz continuous gradient, i.e., there exists $L_F > 0$ such that

$$F(z) - F(x) \le \langle \nabla F(x), z - x \rangle + \frac{L_F}{2}\|z - x\|^2 \qquad \forall x, z.$$

Note that in Assumption 2.1 we require the functions $f^{(1)}(\cdot), \ldots, f^{(T-1)}(\cdot)$ to have Lipschitz continuous gradients, and we do not impose such assumptions on $f^{(T)}(\cdot)$. Hence, we cannot guarantee that $F(x)$ has a Lipschitz continuous gradient, which means Assumption 2.1 does not imply Assumption 2.2.

Although Assumptions 2.1 and 2.2 may seem complicated, they are actually quite mild. They essentially require that the component function be sufficiently smooth and that the samples have bounded second moments. The conditions on smoothness can be easily satisfied when the component functions are polynomial functions. The conditions on second-moment boundedness can be satisfied when the random variables are drawn from a finite set or have sub-Gaussian distributions, which are typically satisfied in big data applications; see the numerical example in section 4.

2.2. Convergence results for *T*-SCGD. Theoretical analysis of Algorithm 1 is challenging due to the nested levels of expectations over a path of random variables. These need to be carefully estimated and balanced to ensure convergence of the algorithm.

We first show the almost sure convergence of the algorithm as long as the step sizes are properly chosen and diminishing under Assumption 2.1. For convex problems, we show that the algorithm generates a sequence of solutions that converges to an optimal solution to problem (1.1) w.p.1. For nonconvex problems with smooth objective, we show that all limiting points of the sequence generated by this algorithm are stationary points w.p.1 under mild assumptions.

Next, we analyze the convergence rate of Algorithm 1. Specifically, we derive the rate by taking the averaged iterates $\widehat{x}_n = \frac{1}{N_n}\sum_{k=n-N_n+1}^{n} x_k$, where $N_n = \lceil n/2 \rceil$. Note that similar results still hold if we let $N_n = n/C$, where $C > 1$ is a positive integer. Clearly, the rate of convergence is closely related to the step sizes α_k and $\beta_{j,k}$. We consider step sizes of the form

$$\alpha_k = k^{-a} \qquad \text{and} \qquad \beta_{j,k} = k^{-b_j} \qquad \forall j = T-1, \ldots, 1, \tag{2.2}$$

where a and the b_j's are real numbers, and obtain the convergence rate by optimizing over a and the b_j's.

Furthermore, we consider multilevel compositional problems with an *optimally strongly convex* objective. Algorithm 1 achieves a much faster convergence rate for such problems. In particular, denote by $\mathcal{X}^*$ the set of optimal solutions x^* to problem (1.1). We say that the objective function F is optimally strongly convex with parameter $\lambda > 0$ if

$$F(x) - F\big(\Pi_{\mathcal{X}^*}(x)\big) \geq \lambda \|x - \Pi_{\mathcal{X}^*}(x)\|^2 \qquad \forall x \in \mathcal{X}. \tag{2.3}$$

Clearly, the class of optimally strongly convex functions strictly contains all strongly convex functions, and is thus more general.

THEOREM 2.1 (convergence of T-SCGD). *Let Assumption* 2.1 *hold, and let* $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ *be the sequence generated by Algorithm* 1 *starting with an arbitrary initial point* $(x_0, y_0^{(T-1)}, \ldots, y_0^{(1)})$.

(a) *Let the step sizes* $\{\alpha_{1,k}\}, \{\beta_{2,k}\}, \ldots, \{\beta_{T,k}\}$ *be such that*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \sum_{k=0}^{\infty} \beta_{j,k} = \infty \qquad \forall j = T-1, \ldots, 1,$$

and

$$\sum_{k=0}^{\infty} \Big(\alpha_k^2 + \beta_{T-1,k}^2 + \cdots + \beta_{1,k}^2 + \frac{\alpha_k^2}{\beta_{2,k}} + \frac{\alpha_k^2}{\beta_{3,k}} + \cdots + \frac{\alpha_k^2}{\beta_{T-1,k}} + \frac{\beta_{T-1,k}^2}{\beta_{T-2,k}} + \cdots + \frac{\beta_{2,k}^2}{\beta_{1,k}}\Big) < \infty.$$

(i) *If* F *is convex,* $\{x_k\}$ *converges almost surely to a random point in the set of optimal solutions to problem* (1.1).

(ii) *Suppose Assumption* 2.2 *holds and, in addition,* $\mathcal{X} = \mathbb{R}^{d_T}$. *Then any limiting point of the sequence* $\{x_k\}_{k=0}^{\infty}$ *is a stationary point of problem* (1.1) *almost surely.*

(b) *If* F *is convex, let* $D_x > 0$ *be such that* $\sup_{x \in \mathcal{X}} \|x - x^*\| \leq D_x$ *and set the step sizes* $\alpha_k = k^{-a}$ *and* $\beta_{j,k} = k^{-b_j}$ *for* $j = T-1, \ldots, 1$, *where* $(a, b_{T-1}, b_{T-2}, \ldots, b_1) \in (0,1)$. *Then we obtain*

$$\mathbb{E}[F(\widehat{x}_n) - F^*] \leq \mathcal{O}(n^{-1/2^T}),$$

by choosing $a = 1 - \frac{1}{2^T}, b_{T-1} = 1 - \frac{1}{2^{T-1}}, \ldots, b_1 = 1 - \frac{1}{2}$.

(c) *Suppose that Assumption* 2.2 *holds, and* F *is optimally strongly convex with some parameter* $\lambda > 0$ *defined in* (2.3). *Letting* $\alpha_k = \frac{1}{\lambda}k^{-a}$, $\beta_{j,k} = k^{-b_j}$ *for* $j = T-1, \ldots, 1$, *we obtain*

$$\mathbb{E}[\|x_n - \Pi_{\mathcal{X}^*}(x_n)\|^2] \leq \mathcal{O}(n^{-2/(T+1)}),$$

by choosing $a = 1$ *and* $b_j = \frac{j+1}{T+1}$ *for* $j = T-1, T-2, \ldots, 1$.

This result characterizes the conditions under which Algorithm 1 converges almost surely. It also provides a sample complexity upper bound for the multilevel stochastic composition optimization problems. In the case when $T = 2$, this result guarantees a convergence rate of $\mathcal{O}(n^{-1/4})$ for convex problems, and of $\mathcal{O}(n^{-2/3})$ for strongly convex problems, which match the convergence rates of convex and strongly convex basic-SCGD given in [27], respectively.

When dealing with nonconvex objectives (e.g., part (a) (ii)), we assume the condition $\mathcal{X} = \mathbb{R}^{d_T}$. We note that it is possible to extend the result to the case when this condition is replaced by a milder condition such as "$\mathcal{X}$ is closed and bounded." Such an extension would require a more sophisticated update rule and more complex analysis to deal with the constraint, which is beyond the scope of this paper. In this paper, we choose to present the most succinct result for nonconvex problems with feasible region $\mathcal{X} = \mathbb{R}^{d_T}$.

The detailed proof of Theorem 2.1 can be derived similarly to the proofs of Theorems 3.1, 3.2, and 3.3. In this paper, to avoid repetition, we omit this proof, which can be found in our online supplementary materials.

3. Accelerated multilevel stochastic gradient algorithm. In the previous section, we established an $\mathcal{O}(n^{-1/2^T})$ rate of convergence for the T-level stochastic composition optimization problem. A key question is whether and when we can better utilize noisy gradients of component functions and improve the overall convergence rate.

Throughout this section, in addition to Assumption 2.1, we impose the following assumption.

Assumption 3.1. Let $C_1, C_2, \ldots, C_T$, $V_1, \ldots, V_T$ be positive scalars.

(i) The samples $f^{(j)}_{\omega_{T,k}}(\cdot), \widetilde{\nabla} f^{(j)}_{\omega_{T,k}}(\cdot)$ have bounded fourth-order moments such that, w.p.1,

$$\mathbb{E}[\|\widetilde{\nabla} f^{(T)}_{\omega_{T,0}}(x)\|^4 | \omega_{1,0}, \ldots, \omega_{T-1,0}] \leq C_T^2,$$
$$\text{and } \mathbb{E}[\|f^{(T)}_{\omega_{T,0}}(x) - f^{(T)}(x)\|^4 | \omega_{1,0}, \ldots, \omega_{T-1,0}] \leq V_T^2 \qquad \forall x \in \mathcal{X}.$$

(ii) The samples $f^{(j)}_{\omega_{j,k}}(\cdot)$ and $\nabla f^{(j)}_{\omega_{j,k}}(\cdot)$ have bounded fourth-order moments such that, w.p.1,

$$\mathbb{E}[\|\nabla f^{(j)}_{\omega_{j,0}}(x_j)\|^4 | \omega_{1,0}, \ldots, \omega_{j-1,0}] \leq C_j^2,$$
$$\mathbb{E}[\|f^{(j)}_{\omega_{j,0}}(x_j) - f^{(j)}(x_j)\|^4 | \omega_{1,0}, \ldots, \omega_{j-1,0}] \leq V_j^2 \ \forall x_j \in \mathbb{R}^{d_j},\ j = T-1, \ldots, 1.$$

We also consider the case when the first inner-level function $f^{(T)}$ also has Lipschitz continuous gradients. In some parts of our subsequent analysis, we make the following assumption.

Assumption 3.2. The function $f^{(T)}$ has Lipschitz continuous gradient such that

$$\|\nabla f^{(T)}(x) - \nabla f^{(T)}(\bar{x})\| \leq L_T \|x - \bar{x}\|$$

for all $x, \bar{x} \in \mathcal{X}$.

In what follows, we propose an accelerated algorithm to better utilize these smoothness properties and achieve improved convergence rates.

3.1. An extrapolation-interpolation scheme for acceleration. The basic idea of acceleration is to refine the running estimates of the value functions by using additional extrapolations. The same idea has been used for the case when $T = 2$. Specifically, in [27], with an additional bounded fourth moments assumption, the authors developed an accelerated SCGD (a-SCGD) algorithm and achieved a faster convergence rate using an extra extrapolation step per iteration.

Now we develop a new accelerated algorithm for the multilevel problem that runs as follows: at the kth iteration, we first update the main iterate solution x_{k+1} by the chain rule,

$$x_{k+1} = \Pi_{\mathcal{X}}\Big\{x_k - \alpha_k \widetilde{\nabla} f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y^{(T-1)}_k)\cdots\nabla f^{(1)}_{\omega_{1,k}}(y^{(1)}_k)\Big\}.$$

We then update the running estimate $y^{(T-1)}_k$ for $\mathbb{E}_{\omega_{T,k}}[f^{(T)}_{\omega_{T,k}}(x_k)|\omega_{1,k},\ldots,\omega_{T-1,k}]$ by taking the weighted average of the new sample and the previous estimate. Specifically, we update $y^{(T-1)}_k$ by letting

$$y^{(T-1)}_{k+1} = (1-\beta_{T-1,k})y^{(T-1)}_k + \beta_{T-1,k} f^{(T)}_{\omega_{T,k+1}}(x_{k+1}).$$

Next, we conduct extrapolation steps for acceleration. The intuition is that we can use sample gradients of individual component functions more efficiently when these functions are smooth, which allows us to obtain better estimates of the $f^{(j)}$'s. In particular, our accelerated updates for the auxiliary variables perform first-order running approximations of the true values. In comparison, the corresponding updates used in T-SCGD can be viewed as zeroth-order running approximations. Specifically, at the kth iteration, we refine our estimate $y^{(j)}_{k+1}$ by taking an additional extrapolation step and obtaining a new auxiliary variable $\widehat{y}^{(j)}_{k+1}$:

$$\widehat{y}^{(j)}_{k+1} = (1-1/\beta_{j,k})y^{(j+1)}_k + y^{(j+1)}_{k+1}/\beta_{j,k}.$$

Then, when we update $y^{(j)}_{k+1}$, we plug in this auxiliary variable, aiming for a better estimate:

$$y^{(j)}_{k+1} = (1-\beta_{j,k})y^{(j)}_k + \beta_{j,k}\cdot f^{(j+1)}_{\omega_{j+1,k+1}}(\widehat{y}^{(j)}_{k+1}).$$

We point out that this is essentially a weighted smoothing scheme, where the $\widehat{y}^{(j)}_k$'s are obtained through extrapolation steps to further utilize the smoothness in order to improve the convergence rate. Roughly speaking, this further extrapolation step helps us achieve estimators $y^{(j)}_{k+1}$ for $f^{(j+1)}(y^{(j+1)}_{k+1})$'s accurate up to second-order terms if we take Taylor expansions of the $f^{(j)}$'s. In comparison, without the extrapolation, if we directly plug in the $y^{(j+1)}_{k+1}$'s instead, the estimators are only accurate up to the first-order terms. We call this an *accelerating update step*. Note that here we do not assume $f^{(T)}$ has a Lipschitz continuous gradient, as in some applications where $f^{(T)}$ includes some sparsity-inducing regularization terms and is not continuously differentiable.

When Assumption 3.2 holds, we update the main iteration by the chain rule, and then apply extrapolation to this level to better utilize the smoothness. That is, we refine our estimate $y^{(T-1)}_{k+1}$ with an additional extrapolation step and an auxiliary variable $\widehat{y}^{(T-1)}_{k+1}$ as

$$\widehat{y}^{(T-1)}_{k+1} = (1-1/\beta_{T-1,k})x_k + x_{k+1}/\beta_{T-1,k}.$$

Next, we update $y_{k+1}^{(T-1)}$ by this auxiliary variable such that

$$y_{k+1}^{(T-1)} = (1-\beta_{T-1,k})y_k^{(T-1)} + \beta_{T-1,k} f_{\omega_{T,k+1}}^{(T)}(\widehat{y}_{k+1}^{(T-1)}).$$

For the remaining levels, we apply the same procedure as in the accelerating update steps described above. We summarize these two slightly different accelerated algorithms in Algorithm 2.

Algorithm 2 Accelerated T-level stochastic compositional gradient descent (a-TSCGD).

Input: $x_0 \in \mathbb{R}^{d_T}$, $y_0^{(j)} \in \mathbb{R}^{d_j}$ for $j = T-1, \ldots, 1$, $\mathcal{SO}$, K, step sizes $\{\alpha_k\}_{k=0}^K$, $\{\beta_{j,k}\}_{k=0}^K$ for $j = T-1, \ldots, 1$.
Output: The sequence $\{x_k\}_{k=0}^K$.
for $k = 0, 1, 2, \ldots, K$ **do**
Query the $\mathcal{SO}$ for the sample values of $f^{(T)}, \ldots, f^{(1)}$ at $x_k, y_k^{(T-1)}, \ldots, y_k^{(1)}$; obtain $\widetilde{\nabla} f_{\omega_{T,k}}^{(T)}(x_k)$, $\nabla f_{\omega_{T-1,k+1}}^{(T-1)}(y_k^{(T-1)}), \ldots, \nabla f_{\omega_{1,k}}^{(1)}(y_k^{(1)})$.
Update the main iterate by

$$x_{k+1} = \Pi_{\mathcal{X}}\Big\{x_k - \alpha_k \widetilde{\nabla} f_{\omega_{T,k}}^{(T)}(x_k) \nabla f_{\omega_{T-1,k}}^{(T-1)}(y_k^{(T-1)}) \cdots \nabla f_{\omega_{1,k}}^{(1)}(y_k^{(1)})\Big\}.$$

if Assumption 3.2 is known to hold **then**
Update the auxiliary variable $\widehat{y}_{k+1}^{(T-1)}$ by

$$\widehat{y}_{k+1}^{(T-1)} = (1 - 1/\beta_{T-1,k})x_k + x_{k+1}/\beta_{T-1,k}.$$

Query the $\mathcal{SO}$ for the sample value of $f^{(T)}$ at $\widehat{y}_{k+1}^{(T-1)}$; obtain $f_{\omega_{T,k+1}}^{(T)}(\widehat{y}_{k+1}^{(T-1)})$.
Update

$$y_{k+1}^{(T-1)} = (1-\beta_{T-1,k})y_k^{(T-1)} + \beta_{T-1,k} f_{\omega_{T,k+1}}^{(T)}(\widehat{y}_{k+1}^{(T-1)}).$$

else if Assumption 3.2 is *not* known to hold **then**
Query the $\mathcal{SO}$ for the sample values of $f^{(T)}$ at x_{k+1}; obtain $f_{\omega_{T,k+1}}^{(T)}(x_{k+1})$.
Update $y^{(T-1)}$ by

$$y_{k+1}^{(T-1)} = (1-\beta_{T-1,k})y_k^{(T-1)} + \beta_{T-1,k} f_{\omega_{T,k+1}}^{(T)}(x_{k+1}).$$

end if
for $j = T-1, \ldots, 2$ **do**
Update the auxiliary variable $\widehat{y}_{k+1}^{(j-1)}$ by

$$\widehat{y}_{k+1}^{(j-1)} = \Big(1 - \frac{1}{\beta_{j-1,k}}\Big)y_k^{(j)} + \frac{1}{\beta_{j-1,k}} y_{k+1}^{(j)}.$$

Query the $\mathcal{SO}$ for the sample value of $f^{(j)}$ at $z_{k+1}^{(j-1)}$; obtain $f_{\omega_{j,k+1}}^{(j)}(z_{k+1}^{(j-1)})$.
Update $y^{(j)}$ by

$$y_{k+1}^{(j-1)} = (1-\beta_{j-1,k})y_k^{(j-1)} + \beta_{j-1,k} f_{\omega_{j,k+1}}^{(j)}(\widehat{y}_{k+1}^{(j-1)}).$$

end for
end for

In the remaining part of this section, we provide theoretical guarantees for this accelerated algorithm. We first provide the almost sure convergence result that our algorithm almost surely converges to an optimal solution when the problem is convex, and any limiting point of the generated solution path is a stationary point. Next, we obtain an improved convergence rate for our algorithm for general nonconvex objective functions. Furthermore, we investigate the case when the objective function is strongly convex, and show that one can achieve faster convergence. For all results,

we provide outlines and key lemmas in the main text, and defer the detailed proofs to Appendixes A, B, and C.

3.2. Almost sure convergence of *a*-TSCGD. We first investigate whether and under what conditions the algorithm converges almost surely. In particular, we provide sufficient conditions of the step sizes, such that when the problem is convex the algorithm converges to an optimal solution almost surely, and when the problem is nonconvex all limiting points of the solution path generated by the algorithm are almost surely stationary points when $F(x)$ has a Lipschitz continuous gradient.

THEOREM 3.1 (almost sure convergence for a-TSCGD). *Let Assumptions* 2.1 *and* 3.1 *hold, and let the step sizes* $\{\alpha_k\}, \{\beta_{T-1,k}\}, \ldots, \{\beta_{1,k}\}$ *be such that*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \ \sum_{k=0}^{\infty} \beta_{T,k} = \infty, \ \ldots, \ \sum_{k=0}^{\infty} \beta_{1,k} = \infty,$$

$$\sum_{k=0}^{\infty} \left(\alpha_k^2 + \beta_{T-1,k}^2 + \cdots + \beta_{1,k}^2 + \frac{\alpha_k^2}{\beta_{T-1,k}} + \cdots + \frac{\alpha_k^2}{\beta_{1,k}} \right) < \infty,$$

and

$$\sum_{k=0}^{\infty} \left(\frac{\beta_{T-1,k}^4}{\beta_{T-2,k}^3} + \cdots + \frac{\beta_{2,k}^4}{\beta_{1,k}^3} \right) < \infty.$$

Let $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ *be the sequence generated by Algorithm* 2 *starting with an arbitrary initial point* $(x_0, y_0^{(T-1)}, \ldots, y_0^{(1)})$. *Then,*

(a) *if* F *is convex, the sequence* $\{x_k\}_{k=0}^{\infty}$ *converges almost surely to a random point in the set of optimal solutions to problem* (1.1)*;*

(b) *supposing in addition that Assumption* 2.2 *holds,* $\mathcal{X} = \mathbb{R}^{d_T}$*, then any limiting point of the sequence* $\{x_k\}_{k=0}^{\infty}$ *is a stationary point of problem* (1.1) *almost surely.*

Furthermore, if Assumption 3.2 *also holds, i.e.,* $f^{(T)}$ *has a Lipschitz continuous gradient, then if the step sizes also satisfy*

$$\sum_{k=0}^{\infty} \frac{\alpha_k^4}{\beta_k^3} < \infty,$$

the assertions in (a) *and* (b) *also hold.*

Proof outline. We provide the proof outline here for the case when the first inner-level function $f^{(T)}$ is nonsmooth. The analysis for problems with a smooth first inner-level function could be derived from the nonsmooth case, and we present the details for both cases in Appendix A.

We denote by $\mathbb{F}_k$ the collection of random variables up to the kth iteration to help us better analyze the convergence properties:

$$\mathbb{F}_k = \Big\{ \{x_i\}_{i=0}^{k}, \{y_i^{(T-1)}\}_{i=0}^{k-1}, \ldots, \{y_i^{(1)}\}_{i=0}^{k-1}, \{\widehat{y}_i^{(T-2)}\}_{i=0}^{k-1}, \ldots, \{\widehat{y}_i^{(1)}\}_{i=0}^{k-1}, \{\omega_{T,i}\}_{i=1}^{k-1}, \ldots, \{\omega_{1,i}\}_{i=1}^{k-1} \Big\}.$$

To derive the almost sure convergence of Algorithm 2, we construct two different T-element supermartingales for the convex and nonconvex objectives, respectively.

First, for problems with convex objective F, in the kth iteration, we use the following lemma to analyze the improvement from $\|x_k - x^*\|$ to $\|x_{k+1} - x^*\|$ by $\|y_k^{(T-1)} - f^{(T)}(x_k)\|$, $\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|$, ..., and $\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|$.

LEMMA 3.1. *Let Assumption* 2.1 *hold, and let* $F = f^{(1)} \circ f^{(2)} \circ \cdots \circ f^{(T)}$ *be convex. Then Algorithm* 2 *generates a sequence* $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ *such that there exist a constant* $C_0 > 0$ *and an optimal solution* $x^* \in \mathcal{X}^*$, *for all* k, *w.p.*1*:*

(3.1)
$$\begin{aligned}
&\mathbb{E}[\|x_{k+1} - x^*\|^2 | \mathbb{F}_k] \\
&\le \left(1 + \left[\frac{\alpha_k^2}{\beta_{T-1,k}} + \cdots + \frac{\alpha_k^2}{\beta_{1,k}}\right] C_0\right) \|x_k - x^*\|^2 + \alpha_k^2 C_1 C_2 \cdots C_T - 2\alpha_k \big(F(x_k) - F^*\big) \\
&\quad + (T-1)\beta_{T-1,k} \mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2 | \mathbb{F}_k] \\
&\quad + (T-2)\beta_{T-2,k} \mathbb{E}[\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|^2 | \mathbb{F}_k] \\
&\quad + \cdots + \beta_{1,k} \mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2 | \mathbb{F}_k].
\end{aligned}$$

Lemma 3.1 states that, for a T-level SCGD with convex objective function F, the optimality error $\|x_{k+1} - x^*\|$ can be bounded by $\|x_k - x^*\|$, $\|y_k^{(T-1)} - f^{(T)}(x_k)\|$, $\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|, \ldots$, and $\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|$ in a supermartingale form.

Next, we present a lemma used in the analysis in part (b).

LEMMA 3.2. *Suppose that Assumptions* 2.1 *and* 2.2 *hold, and* $\mathcal{X} = \mathbb{R}^{d_T}$. *Let* $F^* = \min_{x \in \mathcal{X}} F(x)$. *Then Algorithm* 2 *generates a sequence* $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ *such that*

$$\begin{aligned}
&\mathbb{E}[F(x_{k+1}) - F^* | \mathbb{F}_k] \\
&\le F(x_k) - F^* - \frac{\alpha_k}{2} \|\nabla F(x_k)\|^2 + \frac{1}{2} \alpha_k^2 L_F C_1 C_2 \cdots C_T \\
&\quad + (T-1)\beta_{T-1,k} \mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2 | \mathbb{F}_k] \\
&\quad + (T-2)\beta_{T-2,k} \mathbb{E}[\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|^2 | \mathbb{F}_k] + \cdots \\
&\quad + \beta_{1,k} \mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2 | \mathbb{F}_k],
\end{aligned}$$

for k *sufficiently large, w.p.*1.

This lemma tells us that, for a T-level SCGD with general nonconvex objective function F, $(F(x_{k+1}) - F^*)$ can be bounded by $(F(x_k) - F^*)$, $\|y_k^{(T-1)} - f^{(T)}(x_k)\|$, $\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|, \ldots$, and $\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|$ in a supermartingale form. Similarly to the proof of Lemma 3.1, we shall construct the supermartingales for $\|y_k^{(T-1)} - f^{(T)}(x_k)\|$ and $\|y_k^{(j)} - f^{(j+1)}(y_k^{(j+1)})\|$ for $j = T-2, \ldots, 1$, respectively, and then use Lemma 3.6 to show the almost sure convergence of $(F(x_k) - F^*)$ for a T-level SCGD with nonconvex objective F. With further analysis, we show that any limiting point of the sequence $\{x_k\}_{k=0}^{\infty}$ is a stationary point w.p.1, which proves part (b) of Theorem 3.1.

Next, we analyze the term $\|y_k^{(j)} - f^{(j+1)}(y_k^{(j+1)})\|$ for $j = T-1, \ldots, 1$, respectively, and construct the proper supermartingales for them.

Essentially, we construct a T-element supermartingale to derive the almost sure convergence of the algorithm. For the first inner level, since $f^{(T)}$ is nonsmooth, we construct the supermartingale for this level as follows.

LEMMA 3.3. *Let Assumption* 2.1 *hold, and let* $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ *be the sequence generated by Algorithm* 2. *Suppose* $\mathbb{E}[\|x_{k+1}-x_k\|^2] \le \mathcal{O}(\alpha_k^2)$ *for all* k. *Then we have the following.*

(a) *For all* k, *w.p.*1,

$$\begin{aligned}&\mathbb{E}[\|y_{k+1}^{(T-1)} - f^{(T)}(x_{k+1})\|^2|\mathbb{F}_{k+1}]\\ &\le (1-\beta_{T-1,k})\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2\\ &\qquad + \beta_{T-1}^{-1}C_T\mathbb{E}[\|x_{k+1}-x_k\|^2|\mathbb{F}_{k+1}] + 2V_T\beta_{T-1,k}^2.\end{aligned} \tag{3.2}$$

(b) *If* $\sum_{k=1}^{\infty}\alpha_k^2/\beta_{T-1,k} < \infty$, *then*

$$\sum_{k=1}^{\infty}\beta_{T-1,k}^{-1}\mathbb{E}[\|x_{k+1}-x_k\|^2|\mathbb{F}_{k+1}] < \infty, \qquad \textit{w.p.1.}$$

(c) *There exists a constant* $D_{T-1} \ge 0$ *such that* $\mathbb{E}[\|y_{k+1}^{(T-1)} - f^{(T)}(x_{k+1})\|^2] \le D_{T-1}$ *for all* k.

(d) $\mathbb{E}[\|y_{k+1}^{(T-1)} - y_k^{(T-1)}\|^2] \le \mathcal{O}(\beta_{T-1,k}^2)$ *for all* k.

With the additional finite fourth-moment assumption (Assumption 3.1), we can derive a stronger result in the following lemma.

LEMMA 3.4. *Let Assumptions* 2.1 *and* 3.1 *hold, and let* $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ *be the sequence generated by Algorithm* 2. *Suppose* $\mathbb{E}[\|x_{k+1}-x_k\|^4] \le \mathcal{O}(\alpha_k^4)$ *for all* k *and* $\alpha_k/\beta_{T-1,k} \to 0$ *as* $k \to 0$. *In addition to Lemma* 3.3(a), (b), (c), *and* (d), *we have the following:*

(a) *there exists a constant* $S_{T-1} > 0$ *such that* $\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^4] \le S_{T-1}$ *for all* k;

(b) $\mathbb{E}[\|y_{k+1}^{(T-1)} - y_k^{(T-1)}\|^4] \le \mathcal{O}(\beta_{T-1,k}^4)$ *for all* k.

Note that here we use $y_k^{(T)} = x_k$ and $\beta_{T,k} = \alpha_k$ for ease of notation. This lemma constructs supermartingales of $\{\|y_k^{(j)} - f^{(j+1)}(y_k^{(j+1)})\|\}_{k=1}^{\infty}$ for $j = T-1, \ldots, 1$, respectively, and it also shows that under proper assumptions the tail part for the supermartingale, $\beta_j^{-1}C_{j+1}\mathbb{E}[\|y_{k+1}^{(j)} - y_k^{(j)}\|^2|\mathbb{F}_k] + 2V_{j+1}\beta_{j,k}^2$, converges almost surely.

Next, to construct the supermartingale for the accelerating update steps, we present the following lemma.

LEMMA 3.5. *Let Assumptions* 2.1 *and* 3.1 *hold, and let* $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=1}^{\infty}$ *be the sequence generated by Algorithm* 2. *For* $j = T-2, \ldots, 1$, *suppose*

$$\mathbb{E}[\|y_{k+1}^{(j+1)} - y_k^{(j+1)}\|^4] \le \mathcal{O}(\beta_{j+1,k}^4)$$

for all k *and* $\beta_{j+1,k}/\beta_{j,k} \to 0$ *as* $k \to 0$. *Then there exists a random variable* $e_k^{(j)} \in \mathbb{F}_{k+1}$ *for all* k *satisfying* $\|y_k^{(j)} - f^{(j+1)}(y_k^{(j+1)})\| \le e_k^{(j)}$ *such that*

(a) *for all* k, *w.p.*1,

$$\begin{aligned}&\mathbb{E}[[e_{k+1}^{(j)}]^2|\mathbb{F}_{k+1}]\\ &\quad\le \left(1 - \frac{\beta_{j,k}}{2}\right)[e_k^{(j)}]^2 + 2\beta_{j,k}^2V_{j+1} + \mathcal{O}\left(\frac{\mathbb{E}[\|y_{k+1}^{(j+1)} - y_k^{(j+1)}\|^4|\mathbb{F}_{k+1}]}{\beta_{j,k}^3}\right),\end{aligned}$$

(b) *if* $\sum_{k=1}^{\infty} \beta_{j+1,k}^4/\beta_{j,k}^3 < \infty$, *we have*

$$\sum_{k=1}^{\infty} \frac{\mathbb{E}[\|y_{k+1}^{(j+1)} - y_k^{(j+1)}\|^4 | \mathbb{F}_{k+1}]}{\beta_{j,k}^3} < \infty, \qquad w.p.1,$$

(c) *there exists a constant* $D_j \geq 0$ *such that* $\mathbb{E}[e_k^{(j)}]^2 \leq D_j$ *for all* k,
(d) *there exists a constant* $S_j \geq 0$ *such that* $\mathbb{E}[\|y_k^{(j)} - f^{(j+1)}(y_k^{(j+1)})\|^4] \leq S_j$ *for all* k,
(e) $\mathbb{E}[\|y_{k+1}^{(j)} - y_k^{(j)}\|^4] \leq \mathcal{O}(\beta_{j,k}^4)$ *for all* k.

The above lemmas provide the basic building blocks for a T-element supermartingale. We now provide the T-element supermartingale convergent lemma to establish the convergence property of $\{x_k - x^*\}$.

LEMMA 3.6 (T-element supermartingale convergence). *Let* $\{X_k\}$, $\{Y_k^{(T-1)}\}, \ldots, \{Y_k^{(1)}\}$, $\{\eta_k\}$, *and* $\{u_k^{(j)}\}$, $\{\mu_k^{(j)}\}$, $\{\theta_k^{(j)}\}$, *for* $j = 1, \ldots, T$, *be sequences of nonnegative random variables such that*

$$\mathbb{E}[X_{k+1}|\mathbb{G}_k] \leq (1 + \eta_k)X_k - u_k^{(T)} + \sum_{j=1}^{T-1} c_j \theta_k^{(j)} Y_k^{(j)} + \mu_k^{(T)}$$

and

$$\mathbb{E}[Y_{k+1}^{(T-1)}|\mathbb{G}_k] \leq (1 - \theta_k^{(j)})Y_k^{(j)} - u_k^{(j)} + \mu_k^{(j)} \quad \text{for } j = T-1, \ldots, 1,$$

for all k, *where* $\mathbb{G}_k$ *is the collection of random variables*

$$\Big\{\{X_i\}_{i=0}^k, \{Y_i^{(T-1)}\}_{i=0}^k, \ldots, \{Y_i^{(1)}\}_{i=0}^k, \{\eta_i\}_{i=0}^k, \{u_i^{(j)}\}_{i=0}^k, \{\mu_i^{(j)}\}_{i=0}^k, \{\theta_i^{(j)}\}_{i=0}^k \quad \text{for } j = 1, \ldots, T\Big\},$$

and $c_{T-1}, c_{T-2}, \ldots, c_1$ *are positive scalars. Assume that*

$$\sum_{k=0}^{\infty} \eta_k < \infty, \ \sum_{k=0}^{\infty} \mu_k^{(j)} < \infty \quad \text{for } j = 1, \ldots, T.$$

Then $\{X_k\}, \{Y_k^{(1)}\}, \{Y_k^{(2)}\}, \ldots, \{Y_k^{(T-1)}\}$ *converge almost surely to* T *nonnegative random variables, respectively, and we have*

$$\sum_{j=1}^{T} \sum_{k=0}^{\infty} u_k^{(j)} < \infty, \ \sum_{k=0}^{\infty} \sum_{j=1}^{T-1} c_j \theta_k^{(j)} Y_k^{(j)} < \infty, \qquad w.p.1.$$

By Lemmas 3.1, 3.3, 3.4, and 3.5, we construct the T-element supermartingale and show its convergence by letting

$$X_k = \|x_k - x^*\|^2, \qquad Y_k^{(T-1)} = \mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2 | \mathbb{F}_k],$$
$$Y_k^{(T-2)} = \mathbb{E}[[e_k^{(T-2)}]^2 | \mathbb{F}_k], \ldots, Y_k^{(1)} = \mathbb{E}[[e_k^{(1)}]^2 | \mathbb{F}_k],$$

$$\eta_k = \left[\frac{\alpha_k^2}{\beta_{T-1,k}} + \cdots + \frac{\alpha_k^2}{\beta_{1,k}}\right] C_0, \qquad u_k^{(T)} = 2\alpha_k(F(x_k) - F^*),$$
$$u_k^{(1)} = u_k^{(2)} = \cdots = u_k^{(T-1)} = 0, \qquad c_1 = 2,\ \ldots,\ c_{T-2} = 2(T-2),\ c_{T-1} = T-1,$$
$$\mu_k^{(T-1)} = C_T \beta_{T-1,k}^{-1} \mathbb{E}[\|x_{k+1} - x_k\|^2 | \mathbb{F}_k] + 2V_T \beta_{T-1,k}^2,$$
$$\mu_k^{(T-2)} = 2\beta_{T-2,k}^2 V_{T-1} + \mathcal{O}\left(\frac{\mathbb{E}[\|y_{k+1}^{(T-1)} - y_k^{(T-1)}\|^4 | \mathbb{F}_k]}{\beta_{T-2,k}^3}\right), \ldots,$$
$$\mu_k^{(1)} = 2\beta_{1,k}^2 V_1 + \mathcal{O}\left(\frac{\mathbb{E}[\|y_{k+1}^{(2)} - y_k^{(2)}\|^4 | \mathbb{F}_k]}{\beta_{1,k}^3}\right),$$
$$\mu_k^{(T)} = \alpha_k^2 C_1 C_2 \cdots C_T,$$
$$\theta_k^{(1)} = \beta_{1,k}/2,\ \ldots,\ \theta_k^{(T-2)} = \beta_{T-2,k}/2,\ \theta_k^{(T-1)} = \beta_{T-1,k}.$$

Under the conditions in Theorem 3.1, we obtain that the T-element supermartingale converges almost surely to T random variables by Lemma 3.6; thus, $\|x_k - x^*\|$ converges almost surely, and

$$\sum_{k=0}^{\infty} \alpha_k (F(x_k) - F^*) < \infty, \qquad \text{w.p.1},$$

which further implies that

$$\liminf_{k\to\infty} F(x_k) = F^*, \qquad \text{w.p.1}.$$

Finally, the following lemma shows that the sequence $\{x_k\}_{k=0}^{\infty}$ converges almost surely to an optimal solution to problem (1.1), which completes the proof of part (a).

LEMMA 3.7. *Let $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ be the sequence generated by Algorithm* 2. *Let $F^* = F(x^*)$, where x^* is an optimal solution to problem* (1.1). *Suppose*

$$\liminf_{k\to\infty} F(x_k) = F^*, \qquad \textit{w.p.1.}$$

Then $\{x_k\}$ converges almost surely to a random point in the set of optimal solutions to problem (1.1).

For part (b), by Lemmas 3.2 and 3.3, we construct the T-element supermartingale for general nonconvex functions, and show $\{F(x_k) - F^*\}$ converges almost surely by Lemma 3.6, which further implies $\sum_{k=0}^{\infty} \alpha_k \|\nabla F(x_k)\|^2 < \infty$ w.p.1. Then, we have the following lemma, concluding that any limiting point of the sequence $\{x_k\}$ is a stationary point of $F(x)$ w.p.1.

LEMMA 3.8. *Let $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ be the sequence generated by Algorithm* 2. *Suppose $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k \|\nabla F(x_k)\|^2 < \infty$ w.p.*1, *then any limiting point of the sequence $\{x_k\}$ is a stationary point of $F(x)$ w.p.*1.

This concludes the proof for part (b). ∎

3.3. Convergence rate results for *a*-TSCGD. In this subsection, we study the rate of convergence of the algorithm. We consider step sizes of the form

$$\alpha_k = k^{-a}, \quad \beta_{T-1,k} = k^{-b_{T-1}}, \quad \text{and} \quad \beta_{j,k} = 2k^{-b_j} \quad \forall j = T-2, \ldots, 1,$$

where a and the b_j's are real numbers if the first inner-level function $f^{(T)}$ is nonsmooth, and we choose the step sizes to be

$$\alpha_k = k^{-a} \quad \text{and} \quad \beta_{j,k} = 2k^{-b_j} \quad \forall j = T-1, \ldots, 1,$$

if $f^{(T)}$ is smooth. After optimizing the rate over all a and b_j's, we get the following result for both convex and nonconvex $F(x)$.

THEOREM 3.2 (convergence rate of a-TSCGD). *Suppose that Assumptions* 2.1*,* 2.2*, and* 3.1 *hold and* $\mathcal{X} = \mathbb{R}^{d_T}$*. Let the step sizes be* $\alpha_k = k^{-a}$*,* $\beta_{T-1,k} = k^{-b_{T-1}}$*, and* $\beta_{j,k} = 2k^{-b_j}$ *for* $j = T-2, \ldots, 1$*, where* $a, b_{T-1}, \ldots, b_1 \in (0,1)$*. If we choose the step sizes as* $a = \frac{4+T}{8+T}$ *and* $b_j = \frac{j+3}{8+T}$ *for* $j = T-2, \ldots, 1$*, letting* $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ *be the sequence generated by a-TSCGD Algorithm* 2*, we obtain*

$$\frac{\sum_{k=1}^{n} \mathbb{E}[\|\nabla F(x_k)\|^2]}{n} \leq \mathcal{O}(n^{-4/(8+T)}).$$

Furthermore, if Assumption 3.2 *also holds, Algorithm* 2 *yields*

$$\frac{\sum_{k=1}^{n} \mathbb{E}[\|\nabla F(x_k)\|^2]}{n} \leq \mathcal{O}(n^{-4/(7+T)}),$$

with $\alpha_k = k^{-a}$ *and* $\beta_{j,k} = 2k^{-b_j}$*, where* $a = \frac{3+T}{7+T}$ *and* $b_j = \frac{j+3}{7+T}$ *for* $j = T-1, \ldots, 1$*.*

Proof outline. We present the outline of the proof here and defer the detailed analysis to Appendix B.

We first derive the convergence rate of $\|y_{k+1}^{(j)} - f^{(j+1)}(y_{k+1}^{(j+1)})\|$ and $\|y_{k+1}^{(j)} - y_k^{(j)}\|$ for $j = T-1, \ldots, 1$. By Lemma 3.3 and by Lemma B.1 in Appendix B, we have the following lemma characterizing the corresponding convergence rates.

LEMMA 3.9. *Let Assumption* 2.1 *hold, and let* $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ *be the sequence generated by Algorithm* 2*. Considering the basic update step for the first inner level, we have*

$$\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2] \leq \mathcal{O}(k^{-2a+2b_{T-1}}) + \mathcal{O}(k^{-b_{T-1}}) \quad \forall k.$$

For the accelerating update steps, by Lemma 3.5 and by Lemma B.1 in Appendix B, we have the following result.

LEMMA 3.10. *Let Assumptions* 2.1 *and* 3.1 *hold and let* $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ *be the sequence generated by Algorithm* 2*. Then, for any accelerated update step, we have, for all* k*,*

$$\mathbb{E}[\|y_{k+1}^{(j)} - f^{(j+1)}(y_{k+1}^{(j+1)})\|^2] \leq \mathcal{O}(k^{4(b_j - b_{j+1})}) + \mathcal{O}(k^{-b_j}), \quad j = T-2, \ldots, 1.$$

Under an additional assumption (Assumption 2.2) that F has a Lipschitz gradient, we obtain the following result.

LEMMA 3.11. *Let Assumptions* 2.1*,* 2.2*, and* 3.1 *hold, and let*

$$\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$$

be the sequence generated by Algorithm 2. *Then we have, for all* k,

$$\begin{aligned}
&\mathbb{E}[\|\nabla F(x_k)\|^2] \\
&\le 2\alpha_k^{-1}\mathbb{E}[F(x_k)] - 2\alpha_k^{-1}\mathbb{E}[F(x_{k+1})] + \mathcal{O}\left(\mathbb{E}[\|y_{k+1}^{(T-1)} - f^{(T)}(x_k)\|^2]\right) \\
&\quad + \mathcal{O}\left(\mathbb{E}[\|y_{k+1}^{(T-2)} - f^{(T-1)}(y_{k+1}^{(T-1)})\|^2]\right) + \cdots \\
&\quad + \mathcal{O}\left(\mathbb{E}[\|y_{k+1}^{(1)} - f^{(2)}(y_{k+1}^{(2)})\|^2]\right) + \mathcal{O}(\alpha_k).
\end{aligned}$$

Summing up the inequalities in the previous lemma from $k = 0$ to n, by Lemmas 3.9 and 3.10, we obtain

$$\begin{aligned}
\frac{\sum_{k=1}^n \mathbb{E}[\|\nabla F(x_k)\|^2]}{n} &\le \mathcal{O}(n^{a-1} + n^{-2a+2b_{T-1}}\mathbb{I}^{\log n}_{2(a-b_{T-1})=1} + n^{-b_{T-1}} + n^{-a}) \\
&\quad + \mathcal{O}\left(\sum_{j=1}^{T-2}[n^{4(b_j-b_{j+1})}\mathbb{I}^{\log n}_{4(b_{j+1}-b_j)=1} + n^{-b_j}]\right) \\
&\le \mathcal{O}(n^{-4/(8+T)}),
\end{aligned}$$

by choosing $a = \frac{4+T}{8+T}$ and $b_j = \frac{3+j}{8+T}$ for $j = T-1, \ldots, 1$.

Furthermore, if Assumption 3.2 also holds, i.e., the first inner-level function $f^{(T)}$ has a Lipschitz continuous gradient, then the first inner level could also be updated by the accelerating update rule. By similar analysis as in Lemma 3.10, we have, for all k,

$$\mathbb{E}[\|y_{k+1}^{(T-1)} - f^{(T)}(x_{k+1})\|^2] \le \mathcal{O}(k^{4(b_{T-1}-a)}) + \mathcal{O}(k^{-b_{T-1}}).$$

Combine this inequality with Lemmas 3.10 and 3.11, by choosing $a = \frac{3+T}{7+T}$ and $b_j = \frac{3+j}{7+T}$ for $j = T-1, \ldots, 1$, we obtain

$$\frac{\sum_{k=1}^n \mathbb{E}[\|\nabla F(x_k)\|^2]}{n} \le \mathcal{O}(n^{-4/(7+T)}),$$

which completes the proof. ∎

This result shows that one can solve the multilevel compositional problem using few calls to the sampling oracle when individual component functions are smooth. In the special case when $T = 2$, when the first inner level is smooth, our result strictly improves the convergence rate of the a-SCGD in [27], from $\mathcal{O}(n^{-2/7})$ to $\mathcal{O}(n^{-4/9})$. In this case our result matches the convergence rate by ASC-PG in [29]. To the best of our knowledge, our results for the T-level problem strictly improve and generalize the existing results for the case when $T = 2$.

Next we investigate the convergence rate of Algorithm 2 for an optimally strongly convex objective defined in (2.3). In the next theorem, we prove that, for an optimally strongly convex objective, our algorithm converges faster. We defer the detailed proof to Appendix C.

THEOREM 3.3 (convergence rate of a-TSCGD for strongly convex problems). *Let Assumptions* 2.1, 2.2, *and* 3.1 *hold. Suppose that the objective function* $F(x)$ *in* (1.1) *is optimally strongly convex with some parameter* $\lambda > 0$ *defined in* (2.3). *Set* $\alpha_k = \frac{1}{\lambda}k^{-a}$, $\beta_{T-1,k} = k^{-b_{T-1}}$, *and* $\beta_{j,k} = 2k^{-b_j}$ *for* $j = T-2, \ldots, 1$. *Let*

$\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ be the sequence generated by the a-TSCGD algorithm (Algorithm 2). Then

$$\mathbb{E}[\|x_n - \Pi_{\mathcal{X}^*}(x_n)\|^2] \le \mathcal{O}\left(n^{-a} + n^{-2(a-b_{T-1})} + n^{-b_{T-1}} + \sum_{j=1}^{T-2}[n^{-4(b_{j+1}-b_j)} + n^{-b_j}]\right).$$

With the choice of $a = 1$, $b_{T-1} = \frac{2+T}{4+T}, b_{T-2} = \frac{1+T}{4+T}, \ldots, b_1 = \frac{4}{4+T}$, we have

$$\mathbb{E}[\|x_n - \Pi_{\mathcal{X}^*}(x_n)\|^2] \le \mathcal{O}(n^{-4/(4+T)}).$$

Furthermore, if Assumption 3.2 also holds, Algorithm 2 yields

$$\mathbb{E}[\|x_n - \Pi_{\mathcal{X}^*}(x_n)\|^2] \le \mathcal{O}(n^{-4/(3+T)}),$$

with the step sizes being $\alpha_k = \frac{1}{\lambda}k^{-a}$ and $\beta_{j,k} = 2k^{-b_j}$, where $a = 1$ and $b_j = \frac{3+j}{3+T}$ for $j = T-1, \ldots, 1$.

This result shows that our algorithm achieves a faster convergence for those problems of optimally strongly convexity in the objective functions. For the special case $T = 1$ with a smooth strongly convex function, this result achieves a convergence rate of $\mathcal{O}(n^{-1})$, which meets the convergence rate of the single-level strongly convex stochastic optimization. In addition, for a special case $T = 2$ with a smooth first inner-level function, this result achieves a convergence rate of $\mathcal{O}(n^{-4/5})$, which matches the convergence rate ASC-PG in [29] for optimally strongly convex problems.

4. Example and numerical experiments. In this section, we provide a practical example of the T-level stochastic composition optimization problem (1.1), the risk-averse stochastic optimization, and conduct numerical experiments. Risk-averse stochastic optimization finds wide applications in many fields, such as risk management [4] and government planning [3]. Among the different formulations of risk-averse stochastic optimization problems, one particularly important problem is the mean-deviation risk-averse optimization problem:

$$\min_x \rho(U(x,w)) := \min_x \left\{\mathbb{E}_\omega\left[U(x,\omega)\right] + \lambda\mathbb{E}\Big[\left(\mathbb{E}\left[U(x,\omega)\right] - U(x,\omega)\right)_+^p\Big]^{1/p}\right\}. \tag{4.1}$$

Here the objective ρ is the composition of three expected-value functions. It is also a law-invariant coherent risk measure. See [22, 1] for more detailed discussions.

This problem falls into the problem class (1.1) as a three-level stochastic composition optimization problem. In particular, the problem is equivalent to

$$\min_x (f^{(1)} \circ f^{(2)} \circ f^{(3)})(x),$$

where

$$f^{(1)}\big((y_1, y_2)\big) = y_1 - y_2^{1/p},$$
$$f^{(2)}\big(z, x\big) = \big(z, \mathbb{E}_\omega\left[(z - U(x,\omega))_+^p\right]\big),$$

and

$$f^{(3)}(x) = \left(\mathbb{E}_\omega\left[U(x,\omega)\right], x\right).$$

Note that this problem involves only two random variables (in the nested inner functions $f^{(2)}$ and $f^{(3)}$), yet it is a three-level compositional problem due to the outer function $f^{(1)}$. We remark that stochastic composition optimization is challenging due to the bias induced by using the chain rule to calculate stochastic gradients. In three-level problems, the bias is caused by the inner two levels, so the current problem is as hard as any three-level problem, even though its most outer level is deterministic. As a result, it cannot be solved using existing methods for the two-level problem. Using methods developed in this paper, we can now solve it using a three-level SCGD algorithm.

Next, we conduct numerical experiments. We consider the risk-averse stochastic optimization in a regression setting. In particular, consider a linear model $Y = X\beta^* + \varepsilon$, where we assume all samples of X and ε are independently and identically distributed. Our goal is to estimate β^*, and we consider a risk-averse formulation. Consider the risk-averse optimization problem (4.1). Denoting the ith sample by $\omega_i = \{x_i, y_i\}$, we take

$$U(\beta, \omega_i) = (y_i - x_i^T \beta)^2,$$

and we set $p = 2$. To the best of our knowledge, our algorithm is the first gradient-based method which can be adopted to solve this three-level stochastic optimization problem. We point out that this risk-averse regression approach tends to provide "stable" solutions. This defines a general notion of stability in statistics given in [18, 25], where the stability is usually defined as variance, and the "good" cases are penalized when the empirical error is smaller than its expectation. In comparison, in our approach, we do not penalize these "good" cases.

Let the dimension of the covariate x_i be d. We consider three setups to generate the data.

- Setup 1: $X \sim N(0, I_d)$.
- Setup 2: $X \sim N(0, \Sigma)$, where $\Sigma_{jj} = 1$ and $\Sigma_{jk} = 0.5$ for $j, k = 1, \ldots, d$ and $j \neq k$.
- Setup 3: $X \sim N(0, \Sigma)$, where $\Sigma_{jk} = 0.5e^{-\frac{|j-k|}{d}}$.

Since our problem is convex, by our theoretical analysis, the generated sequence of solutions converges to the optimal solution. As the true optimal solution is unknown (note that β^* is not necessarily the optimal solution), we take the solution after 500,000 iterations as the optimal solution. Meanwhile, in all setups, we draw the random variables $\varepsilon \sim N(0, 0.2)$, and generate each component of $\beta^* \in \mathbb{R}^d$ independently from a standard normal distribution. We set the step sizes to be $\alpha_k = k^{-3/5}$, $\beta_{2,k} = 2k^{-1/2}$, and $\beta_{1,k} = 2k^{-2/5}$. The samples of X are generated independently by the distribution specified in the corresponding setup. In each iteration of the algorithm, we draw a new sample of (X, Y), and update the solution using Algorithm 2.

We have experimented with the proposed algorithm using multiple values of the dimension, i.e., $d \in \{50, 100, 150, 200\}$. For each instance, we ran 100 replications and plotted the averaged differences between the solution at the kth iteration β_k and the optimal solution $\widehat{\beta}$ in Figures 2, 3, 4, and 5. Let us study the performance of the algorithm when d varies. When $d = 200$ we note that it takes approximately twice as many iterations as the $d = 50$ case to reach the same accuracy. This is mainly because the variance of the stochastic gradient increases as the dimension grows, due to the fact that the noise in our experiment is Gaussian with unit variance per entry. As a result, the hidden constant inside the error bound, involving sums and products of multilevel variances and Lipschitz constants, also grows polynomially as d grows.

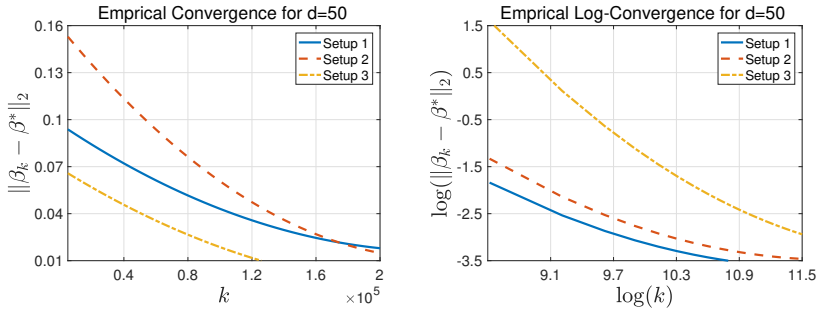

FIG. 2. *Averaged difference between the generated solution and the optimal solution and empirical convergence rate when* $d = 50$.

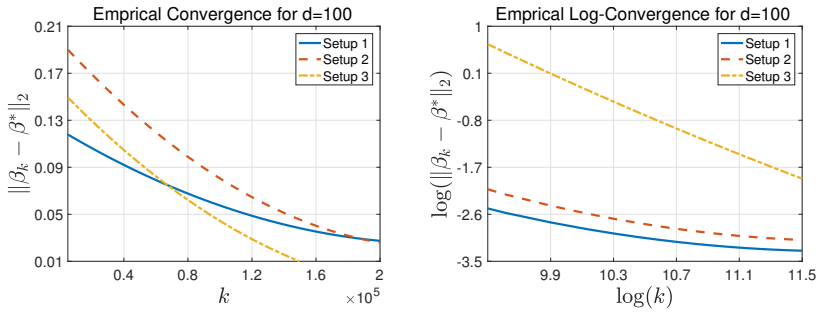

FIG. 3. *Averaged difference between the generated solution and the optimal solution and empirical convergence rate when* $d = 100$.

Therefore, it takes more iterations for the algorithm to converge to the same accuracy level when the overall variance increases with growing dimensions.

To further investigate empirical rates of convergence under all the different settings, we plot the averaged $\log(k)$ against $\log(\|\beta_k - \widehat{\beta}\|)$ after 100 replications, where $\widehat{\beta}$ is the optimal solution. As we can see from the figures, the slopes remain the same regardless of the dimension d. We find that for all cases the slopes of the lines are close to $-2/5$, which matches our theoretical analysis that our algorithm converges at a rate of $\mathcal{O}(k^{-2/5})$ for three-level problems.

5. Conclusion. In this paper, we propose what we believe to be the first gradient-type algorithms for a class of multilevel stochastic composition optimization problems. We provide strong theoretical guarantees for our algorithms. In particular, we prove almost sure convergence results: when the problem is convex, our algorithm converges to an optimal solution, and when the problem is nonconvex, every limiting point of the sequence of solutions is an stationary point. Under various assumptions, we further characterize the convergence rates of our algorithms. In the case when $T = 2$, our convergence rate result matches and strictly generalizes the best known result in [29]. In the case when $T \geq 3$, our results provide the first few benchmarks on the sample complexity for solving multilevel stochastic optimization problems.

There are several interesting future research problems. First, our convergence rate result requires that the inner-level functions $f^{(2)}, \ldots, f^{(T)}$ be smooth. It is unclear how

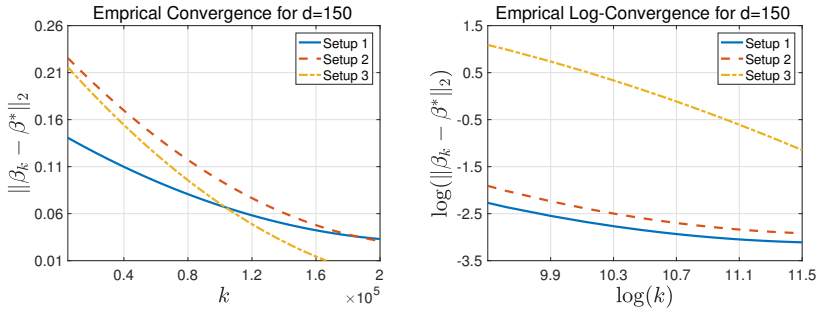

FIG. 4. *Averaged difference between the generated solution and the optimal solution and empirical convergence rate when* $d = 150$.

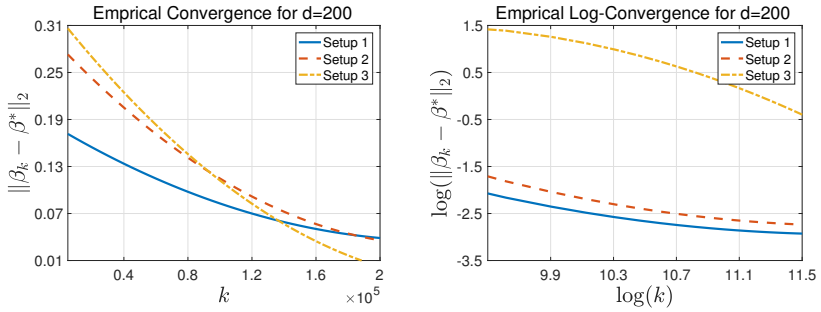

FIG. 5. *Averaged difference between the generated solution and the optimal solution and empirical convergence rate when* $d = 200$.

to achieve fast convergence when some of these functions are nonsmooth. Second, it is not clear whether the convergence rate can be improved or not. We are not aware of any sample complexity lower bound for the multilevel stochastic optimization problem. Third, it is of practical interest to consider the special case when all expectations are finite sums. In this case, one may conjecture that variance reduction can be used to further improve the algorithms' efficiency.

Appendix A. Proof of Theorem 3.1. In this section, we present the detailed proof for Theorem 3.1.

A.1. Proof of Lemma 3.1. Before presenting the detailed proof of Lemma 3.1, we present a lemma that is used in proving it.

LEMMA A.1. *Suppose Assumption* 2.1 *holds, and let* $\{(x_k, y_k^{(T-1)}, \ldots, y_k^{(1)})\}_{k=0}^{\infty}$ *be the sequence generated by Algorithm* 1. *Define*

$$\widetilde{\nabla} F_{\omega_k}(x_k) \equiv \widetilde{\nabla} f^{(T)}_{\omega_{T,k}}(x_k) \nabla f^{(T-1)}_{\omega_{T-1,k}}(f^{(T)}(x_k)) \cdots \nabla f^{(1)}_{\omega_{1,k}}(f^{(2)} \circ \cdots \circ f^{(T)}(x_k)),$$

and let $X_k \in \mathbb{F}_k$ *be a vector of random variables, where* $\mathbb{F}_k$ *is the collection of random variables*

$$\{\{x_i\}_{i=0}^{k}, \{y_i^{(T-1)}\}_{i=0}^{k-1}, \ldots, \{y_i^{(1)}\}_{i=0}^{k-1}, \{\omega_{T,i}\}_{i=0}^{k-1}, \ldots, \{\omega_{1,i}\}_{i=0}^{k-1}\}.$$

*Then there exists a constant $C_0 > 0$ dependent only on the number of levels T such that for all k, w.p.*1,

$$\begin{aligned}
&X_k'\mathbb{E}[\widetilde{\nabla}F_{\omega_k}(x_k) - \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})|\mathbb{F}_k]\\
&\le (T-1)\beta_{T-1,k}\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2|\mathbb{F}_k]\\
&\qquad + (T-2)\beta_{T-2,k}\mathbb{E}[\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|^2|\mathbb{F}_k]\\
&\qquad + \cdots + \beta_{1,k}\mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2|\mathbb{F}_k] + C_0\left(\frac{1}{\beta_{T-1,k}} + \cdots + \frac{1}{\beta_{1,k}}\right)\|X_k\|^2.
\end{aligned}$$

Proof. We begin our analysis by the chain rule as follows:

$$\begin{aligned}
&\widetilde{\nabla}F_{\omega_k}(x_k) - \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})\\
&= \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(f^{(T)}(x_k))\nabla f^{(T-2)}_{\omega_{T-2,k}}(f^{(T-1)}(f^{(T)}(x_k)))\\
&\qquad \cdots\nabla f^{(1)}_{\omega_{1,k}}(f^{(2)}\circ\cdots\circ f^{(T)}(x_k))\\
&\qquad - \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\nabla f^{(T-2)}_{\omega_{T-2,k}}(f^{(T-1)}(f^{(T)}(x_k)))\\
&\qquad \cdots\nabla f^{(1)}_{\omega_{1,k}}(f^{(2)}\circ\cdots\circ f^{(T)}(x_k))\\
&\vdots\\
&\qquad + \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\nabla f^{(T-2)}_{\omega_{T-2,k}}(y_k^{(T-2)})\\
&\qquad \cdots\nabla f^{(2)}_{\omega_{2,k}}(y_k^{(2)})\nabla f^{(1)}_{\omega_{1,k}}(f^{(2)}\circ\cdots\circ f^{(T)}(x_k))\\
&\qquad - \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\nabla f^{(T-2)}_{\omega_{T-2,k}}(y_k^{(T-2)})\cdots\nabla f^{(2)}_{\omega_{2,k}}(y_k^{(2)})\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)}).
\end{aligned}$$

Define

$$\begin{aligned}
S_m = \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k^{(T)})\cdots\nabla f^{(m)}_{\omega_{m,k}}(y_k^{(m)})\nabla f^{(m-1)}_{\omega_{m-1,k}}(f^{(m)}\circ\cdots\circ f^{(T)}(x_k))\\
\cdots\nabla f^{(1)}_{\omega_{1,k}}(f^{(2)}\circ\cdots\circ f^{(T)}(x_k)).
\end{aligned}$$

Clearly, $S_T = \widetilde{\nabla}F_{\omega_k}(x_k)$ and $S_1 = \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})$, and we have

$$\begin{aligned}
&\widetilde{\nabla}F_{\omega_k}(x_k) - \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})\\
&= (S_T - S_{T-1}) + (S_{T-1} - S_{T-2}) + \cdots + (S_2 - S_1).
\end{aligned} \tag{A.1}$$

Considering $S_m - S_{m-1}$, by definition, we obtain

(A.2)

$$\begin{aligned}
\|S_m - S_{m-1}\| &= \|\widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(y_k^{(T)})\cdots\nabla f^{(m)}_{\omega_{m,k}}(y_k^{(m)})P_m\nabla f^{(m-2)}_{\omega_{m-2,k}}(f^{(m-1)}\circ\cdots\circ f^{(T)}(x_k))\\
&\qquad \cdots\nabla f^{(1)}_{\omega_{1,k}}(f^{(2)}\circ\cdots\circ f^{(T)}(x_k))\|\\
&\le M_m\|P_m\|,
\end{aligned}$$

where $P_m = \nabla f^{(m-1)}_{\omega_{m-1,k}}(f^{(m)}\circ\cdots\circ f^{(T)}(x_k)) - \nabla f^{(m-1)}_{\omega_{m-1,k}}(y_k^{(m-1)})$ and

$$\begin{aligned}
M_m = \|\widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(y_k^{(T)})\|\cdots\|\nabla f^{(m)}_{\omega_{m,k}}(y_k^{(m)})\|\|\nabla f^{(m-2)}_{\omega_{m-2,k}}(f^{(m-1)}\circ\cdots\circ f^{(T)}(x_k))\|\\
\cdots\|\nabla f^{(1)}_{\omega_{1,k}}(f^{(2)}\circ\cdots\circ f^{(T)}(x_k))\|.
\end{aligned}$$

By Assumptions 2.1(iii) and (iv), we have

$$\mathbb{E}[M_m^2|\mathbb{F}_k] \le C_T C_{T-1} \cdots C_m C_{m-2} \cdots C_1. \tag{A.3}$$

Consider P_m,

$$\begin{aligned} \|P_m\| &= \|\nabla f^{(m-1)}_{\omega_{m-1,k}}(f^{(m)} \circ \cdots \circ f^{(T)}(x_k)) - \nabla f^{(m-1)}_{\omega_{m-1,k}}(y_k^{(m-1)})\| \\ &\le L_{m-1}\sqrt{C_m \cdots C_{T-1}}\|f^{(T)}(x_k) - y_k^{(T-1)}\| \\ &\quad + L_{m-1}\sqrt{C_m \cdots C_{T-2}}\|f^{(T-1)}(y_k^{(T-1)}) - y_k^{(T-2)}\| \\ &\quad + \cdots + L_{m-1}\|f^{(m)}(y_k^{(m)}) - y_k^{(m-1)}\|, \end{aligned} \tag{A.4}$$

where the first inequality holds by Assumption 2.1(iv) as $f^{(m-1)}_{\omega_{m-1,k}}$ has Lipschitz continuous gradient with parameter L_j, and the last inequality holds by Assumptions 2.1(iii) and (iv), i.e., that $f^{(j)}$ is Lipschitz continuous with parameter C_j for all j.

Substituting (A.4) into (A.2) yields

$$\begin{aligned} &\|X_k\|\|S_m - S_{m-1}\| \\ &\le M_m\|X_k\|\|P_m\| \\ &\le M_m\|X_k\|(L_{m-1}\sqrt{C_m \cdots C_{T-1}}\|f^{(T)}(x_k) - y_k^{(T-1)}\| \\ &\qquad + L_{m-1}\sqrt{C_m \cdots C_{T-2}}\|f^{(T-1)}(y_k^{(T-1)}) - y_k^{(T-2)}\| \\ &\qquad + \cdots + L_{m-1}\|f^{(m)}(y_k^{(m)}) - y_k^{(m-1)}\|) \\ &\le \beta_{T-1,k}\|f^{(T)}(x_k) - y_k^{(T-1)}\|^2 + \beta_{T-2,k}\|f^{(T-1)}(y_k^{(T-1)}) - y_k^{(T-2)}\|^2 \\ &\quad + \cdots + \beta_{m-1,k}\|f^{(m)}(y_k^{(m)}) - y_k^{(m-1)}\|^2 \\ &\quad + M_k^2\|X_k\|^2\left(\frac{L_{m-1}^2 C_m \cdots C_{T-1}}{4\beta_{T-1,k}} + \frac{L_{m-1}^2 C_m \cdots C_{T-2}}{4\beta_{T-2,k}} + \cdots + \frac{L_{m-1}^2}{4\beta_{m-1,k}}\right), \end{aligned}$$

where the last inequality holds by the fact that $2xy \le ax^2 + \frac{1}{a}y^2$ for any $x, y \in \mathbb{R}$ and $a > 0$. Since $X_k \in \mathbb{F}_k$, taking expectation on both sides of the previous inequality and combining it with (A.3), there exists a constant $R_m > 0$ such that almost surely

$$\begin{aligned} &\mathbb{E}[\|X_k\|\|S_m - S_{m-1}\||\mathbb{F}_k] \\ &\le \beta_{T-1,k}\mathbb{E}[\|f^{(T)}(x_k) - y_k^{(T-1)}\|^2|\mathbb{F}_k] + \beta_{T-2,k}\mathbb{E}[\|f^{(T-1)}(y_k^{(T-1)}) - y_k^{(T-2)}\|^2|\mathbb{F}_k] \\ &\quad + \cdots + \beta_{m-1,k}\mathbb{E}[\|f^{(m)}(y_k^{(m)}) - y_k^{(m-1)}\|^2|\mathbb{F}_k] \\ &\quad + R_m\left(\frac{1}{\beta_{T-1,k}} + \cdots + \frac{1}{\beta_{m-1,k}}\right)\|X_k\|^2. \end{aligned} \tag{A.5}$$

Meanwhile, we have

$$\begin{aligned} &X_k'\mathbb{E}[\widetilde{\nabla}F_{\omega_k}(x_k) - \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)}) \cdots \nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})|\mathbb{F}_k] \\ &\le \|X_k\|\mathbb{E}[\|S_T - S_{T-1}\| + \|S_{T-1} - S_{T-2}\| + \cdots + \|S_2 - S_1\||\mathbb{F}_k]. \end{aligned} \tag{A.6}$$

Substituting (A.5) into (A.6) and summing from $m = 2$ to $m = T$, with some algebraic manipulation, we conclude that there exists a constant $C_0 > 0$ dependent only on the

number of levels T such that, w.p.1,

$$\begin{aligned}
&X_k'\mathbb{E}[\widetilde{\nabla}F_{\omega_k}(x_k) - \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})|\mathbb{F}_k]\\
&\le (T-1)\beta_{T-1,k}\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2|\mathbb{F}_k]\\
&\quad + (T-2)\beta_{T-2,k}\mathbb{E}[\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|^2|\mathbb{F}_k]\\
&\quad + \cdots + \beta_{1,k}\mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2|\mathbb{F}_k] + C_0\left(\frac{1}{\beta_{T-1,k}} + \cdots + \frac{1}{\beta_{1,k}}\right)\|X_k\|^2,
\end{aligned}$$

which completes the proof. ∎

Next, we present the proof of Lemma 3.1.

Proof of Lemma 3.1. We have

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\le \|x_k - x^* - \alpha_k\widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})\|^2\\
&= \|x_k - x^*\|^2 - 2\alpha_k(x_k - x^*)'\widetilde{\nabla}F_{\omega_k}(x_k) + u_k\\
&\quad + \alpha_k^2\|\widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})\|^2,
\end{aligned} \tag{A.7}$$

where

$$\widetilde{\nabla}F_{\omega_k}(x_k) = \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(f^{(T)}(x_k))\cdots\nabla f^{(1)}_{\omega_{1,k}}(f^{(2)}\circ\cdots\circ f^{(T)}(x_k)),$$

as defined in the main text, and

$$u_k = 2\alpha_k(x_k - x^*)'[\widetilde{\nabla}F_{\omega_k}(x_k) - \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})].$$

By Assumptions 2.1(iii) and (iv), the $\widetilde{\nabla}f^{(j)}_{\omega_{j,k}}$'s have bounded second-order moments. Thus, with probability 1,

$$\mathbb{E}[\|\widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})\|^2|\mathbb{F}_k] \le C_1\cdots C_T. \tag{A.8}$$

Taking expectation on both sides of (A.7), and conditioning on $\mathbb{F}_k$, we have

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^*\|^2|\mathbb{F}_k] \le \|x_k - x^*\|^2 + \alpha_k^2 C_1C_2\cdots C_T\\
+ \mathbb{E}[u_k|\mathbb{F}_k] - 2\alpha_k(x_k - x^*)'\mathbb{E}[\widetilde{\nabla}F_{\omega_k}(x_k)|\mathbb{F}_k].
\end{aligned}$$

By the convexity of $F(x)$ and Assumption 2.1(ii), we obtain

$$\mathbb{E}[\|x_{k+1}-x^*\|^2|\mathbb{F}_k] \le \|x_k-x^*\|^2+\alpha_k^2C_1C_2\cdots C_T-2\alpha_k(F(x_k)-F^*)+\mathbb{E}[u_k|\mathbb{F}_k]. \tag{A.9}$$

For the term u_k, there exists a constant $C_0 > 0$ such that

$$\begin{aligned}
&\mathbb{E}[u_k|\mathbb{F}_k]\\
&\le (T-1)\beta_{T-1,k}\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2|\mathbb{F}_k]\\
&\quad + (T-2)\beta_{T-2,k}\mathbb{E}[\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|^2|\mathbb{F}_k]\\
&\quad + \cdots + \beta_{1,k}\mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2|\mathbb{F}_k] + C_0\left(\frac{\alpha_k^2}{\beta_{T-1,k}} + \cdots + \frac{\alpha_k^2}{\beta_{1,k}}\right)\|x_k - x^*\|^2,
\end{aligned} \tag{A.10}$$

where the last inequality comes from Lemma A.1 by letting $X_k = 2\alpha_k(x_k - x^*) \in \mathbb{F}_k$. Substituting (A.10) into (A.9), we get

$$\begin{aligned}
&\mathbb{E}[\|x_{k+1} - x^*\|^2|\mathbb{F}_k] \\
&\le \left(1 + C_0\left(\frac{\alpha_k^2}{\beta_{T-1,k}} + \frac{\alpha_k^2}{\beta_{T-2,k}} + \cdots + \frac{\alpha_k^2}{\beta_{1,k}}\right)\right)\|x_k - x^*\|^2 \\
&\quad + \alpha_k^2 C_1 C_2 \cdots C_T - 2\alpha_k(F(x_k) - F^*) \\
&\quad + (T-1)\beta_{T-1,k}\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2|\mathbb{F}_k] \\
&\quad + \cdots + \beta_{1,k}\mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2|\mathbb{F}_k],
\end{aligned}$$

which completes the proof. ∎

A.2. Proof of Lemma 3.2. By the assumptions in part (b), F has Lipschitz continuous gradient with parameter L_F and $\mathcal{X} = \mathbb{R}^{d_T}$. We denote by $\nabla F(x)$ the gradient of $F(x)$, and obtain

$$\begin{aligned}
&F(x_{k+1}) - F(x_k) \\
&\le \langle \nabla F(x_k), x_{k+1} - x_k\rangle + \frac{L_F}{2}\|x_{k+1} - x_k\|^2 \\
&= -\alpha_k\|\nabla F(x_k)\|^2 \\
&\quad + \alpha_k \nabla F(x_k)'[\nabla F(x_k) - \widetilde{\nabla} f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)}) \cdots \nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})] \\
&\quad + \frac{L_F}{2}\|x_{k+1} - x_k\|^2.
\end{aligned} \tag{A.11}$$

As defined in the main text,

$$\widetilde{\nabla} F_{\omega_k}(x_k) = \widetilde{\nabla} f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(f^{(T)}(x_k)) \cdots \nabla f^{(1)}_{\omega_{1,k}}(f^{(2)} \circ \cdots \circ f^{(T)}(x_k)).$$

By Assumption 2.1(ii), we have

$$\mathbb{E}[\widetilde{\nabla} F_{\omega_k}(x_k)|\mathbb{F}_k] = \nabla F(x_k).$$

We obtain that w.p.1 there exists a constant $C_0 > 0$ such that

$$\begin{aligned}
&\alpha_k\mathbb{E}[\nabla F(x_k)'(\nabla F(x_k) - \widetilde{\nabla} f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)}) \cdots \nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)}))|\mathbb{F}_k] \\
&\le (T-1)\beta_{T-1,k}\mathbb{E}[\|y_{k+1}^{(T-1)} - f^{(T)}(x_k)\|^2|\mathbb{F}_k] \\
&\quad + (T-2)\beta_{T-2,k}\mathbb{E}[\|y_{k+1}^{(T-2)} - f^{(T-1)}(y_{k+1}^{(T-1)})\|^2|\mathbb{F}_k] \\
&\quad + \cdots + \beta_{1,k}\mathbb{E}[\|y_{k+1}^{(1)} - f^{(2)}(y_{k+1}^{(2)})\|^2|\mathbb{F}_k] + C_0\left(\frac{\alpha_k^2}{\beta_{T-1,k}} + \cdots + \frac{\alpha_k^2}{\beta_{1,k}}\right)\|\nabla F(x_k)\|^2,
\end{aligned}$$

where the last inequality comes from Lemma A.1 by letting $X_k = \alpha_k \nabla F(x_k) \in \mathbb{F}_k$. Also note that

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x_k\|^2|\mathbb{F}_k] &= \alpha_k^2\mathbb{E}[\|\widetilde{\nabla} f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)}) \cdots \nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})\|^2|\mathbb{F}_k] \\
&\le \alpha_k^2 C_1 \cdots C_T.
\end{aligned} \tag{A.12}$$

Combining the results above, we obtain

$$\begin{aligned}\mathbb{E}[F(x_{k+1}) - F^*|\mathbb{F}_k] \le F(x_k) - F^* &+ \frac{1}{2}\alpha_k^2 L_F C_1 C_2 \cdots C_T \\ &- \alpha_k\left(1 - \left(\frac{\alpha_k}{\beta_{T-1,k}} - \cdots - \frac{\alpha_k}{\beta_{1,k}}\right)C_0\right)\|\nabla F(x_k)\|^2 \\ &+ (T-1)\beta_{T-1,k}\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2|\mathbb{F}_k] \\ &+ (T-2)\beta_{T-2,k}\mathbb{E}[\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|^2|\mathbb{F}_k] \\ &+ \cdots + \beta_{1,k}\mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2|\mathbb{F}_k].\end{aligned}$$

Besides, we have $1/2 \le 1 - (\frac{\alpha_k}{\beta_{T-1,k}} - \cdots - \frac{\alpha_k}{\beta_{T-1,k}})C_0$ for k sufficiently large since $\alpha_k/\beta_{j,k} \to 0$ as $k \to \infty$ for all j. Finally, we conclude

$$\begin{aligned}\mathbb{E}[F(x_{k+1}) - F^*|\mathbb{F}_k] \le F(x_k) - F^* &- \frac{\alpha_k}{2}\|\nabla F(x_k)\|^2 + \frac{1}{2}\alpha_k^2 L_F C_1 C_2 \cdots C_T \\ &+ (T-1)\beta_{T-1,k}\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2|\mathbb{F}_k] \\ &+ \cdots + \beta_{1,k}\mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2|\mathbb{F}_k],\end{aligned}$$

for k sufficiently large, which completes our proof. ∎

A.3. Proof of Lemma 3.3. For ease of presentation, we denote $y_{k+1}^{(T-1)}$ by z_{k+1}, $\beta_{T-1,k}$ by γ_k, and $f_{\omega_{T,k+1}^{(T)}}^{(T)}(x_k)$ by $h_{u_{k+1}}(x_k)$. The corresponding update step can be written as

$$z_{k+1} = (1-\gamma_k)z_k + \gamma_k h_{u_{k+1}}(x_{k+1}).$$

Proof of Lemma 3.3. Lemma 3.3 analyzes the behavior of a basic update step in the a-TSCGD. Note that parts (a)–(c) of this lemma can be deduced from [27, Lemma 2]; we skip these proofs to avoid repetition, and present the detailed proof of part (d) below.

By the definition of z_{k+1}, we have $z_{k+1} = (1-\gamma_k)z_k + \gamma_k h_{u_{k+1}}(x_{k+1})$ and

$$\begin{aligned}(1-\gamma_k)^2\|z_{k+1} - z_k\|^2 &= \gamma_k^2\|h_{u_{k+1}}(x_{k+1}) - z_{k+1}\|^2 \\ &\le 2\gamma_k^2\|h_{u_{k+1}}(x_{k+1}) - h(x_{k+1})\|^2 + 2\gamma_k^2\|z_{k+1} - h(x_{k+1})\|^2.\end{aligned}$$

Then we obtain

$$\mathbb{E}[\|z_{k+1} - z_k\|^2] \le \frac{2\gamma_k^2}{(1-\gamma_k)^2}V_h + \frac{2\gamma_k^2}{(1-\gamma_k)^2}\mathbb{E}[\|z_{k+1} - h(x_{k+1})\|^2]. \tag{A.13}$$

From part (c), we have that there exists $D_z \ge 0$ such that $\mathbb{E}[\|z_{k+1} - h(x_{k+1})\|^2] \le D_z$. Plugging this into (A.13), we conclude that

$$\mathbb{E}[\|z_{k+1} - z_k\|^2] \le \mathcal{O}(\gamma_k^2).$$ ∎

A.4. Proof of Lemma 3.4. Note that here we use the same notation as in the proof of Lemma 3.3. Now we show the detailed proof of Lemma 3.4.

Proof. (a) Under the assumption $\mathbb{E}[\|x_{k+1} - x_k\|^4] \le \mathcal{O}(\alpha_k^4)$, there exists a constant $C_0 > 0$ such that $\mathbb{E}[\|x_{k+1} - x_k\|^4] \le C_0\alpha_k^4$. By Lemma 3.3, there exists a constant $D_z > 0$ such that $\mathbb{E}[\|z_{k+1} - h(x_{k+1})\|^2] \le D_z$, and we also have $\mathbb{E}[\|z_{k+1} - h(x_{k+1})\|] \le \sqrt{\mathbb{E}[\|z_{k+1} - h(x_{k+1})\|^2]} \le \sqrt{D_z}$.

Let $e_{k+1} = (1-\gamma_k)(h(x_{k+1}) - h(x_k))$. Together with the definition of z_{k+1}, we get

(A.14)
$$z_{k+1} - h(x_{k+1}) + e_{k+1} - e_{k+1} = (1-\gamma_k)(z_k - h(x_k)) + \gamma_k(h_{u_{k+1}}(x_{k+1}) - h(x_{k+1})) - e_{k+1}.$$

By the Lipschitz continuity of h, we obtain

$$\|e_{k+1}\| \le (1-\gamma_k)\sqrt{C_h}\|x_{k+1} - x_k\|.$$

Meanwhile, we have

$$\|z_{k+1} - h(x_{k+1})\| \le \|(1-\gamma_k)(z_k - h(x_k)) + \gamma_k(h_{u_{k+1}}(x_{k+1}) - h(x_{k+1}))\| + \|e_{k+1}\|.$$

Let $P_k = \|(1-\gamma_k)(z_k - h(x_k)) + \gamma_k(h_{u_{k+1}}(x_{k+1}) - h(x_{k+1}))\|$. Considering the fourth moment, we get

$$\begin{aligned} P_k^4 &= \|(1-\gamma_k)(z_k - h(x_k)) + \gamma_k(h_{u_{k+1}}(x_{k+1}) - h(x_{k+1}))\|^4 \\ &\le (1-\gamma_k)^4\|z_k - h(x_k)\|^4 + 4(1-\gamma_k)^3\gamma_k(z_k - h(x_k))^3(h_{u_{k+1}}(x_{k+1}) - h(x_{k+1})) \\ &\quad + 6(1-\gamma_k)^2\gamma_k^2\|z_k - h(x_k)\|^2\|h_{u_{k+1}}(x_{k+1}) - h(x_{k+1})\|^2 \\ &\quad + 4(1-\gamma_k)\gamma_k^3\|z_k - h(x_k)\|\|h_{u_{k+1}}(x_{k+1}) - h(x_{k+1})\|^3 \\ &\quad + \gamma_k^4\|h_{u_{k+1}}(x_{k+1}) - h(x_{k+1})\|^4. \end{aligned}$$

So we obtain

$$\mathbb{E}[P_k^4] \le (1-\gamma_k)^4\mathbb{E}[\|z_k - h(x_k)\|^4] + 6(1-\gamma_k)^2\gamma_k^2 D V_h + 4(1-\gamma_k)\gamma_k^3\sqrt{D}V_h^{3/2} + \gamma_k^4 V_h^2.$$

Using the fact that $\|a+b\|^2 \le (1+\varepsilon)\|a\|^2 + (1+1/\varepsilon)\|b\|^2$ and $\|a+b\|^4 \le (1+\varepsilon)^3\|a\|^4 + (1+1/\varepsilon)^3\|b\|^4$ for any $\varepsilon > 0$, we have

$$\|z_{k+1} - h(x_{k+1})\|^4 \le (1+\gamma_k)^3 P_k^4 + (1+1/\gamma_k)^3\|e_{k+1}\|^4, \tag{A.15}$$

and

(A.16)
$$\begin{aligned} &\mathbb{E}[\|z_{k+1} - h(x_{k+1})\|^4] \\ &\le (1-\gamma_k)\mathbb{E}[\|z_k - h(x_k)\|^4] + 12\gamma_k^2 D_z V_h + 16\gamma_k^3\sqrt{D_z}V_h^{3/2} + 8\gamma_k^4 V_h^2 + \frac{\alpha_k^4}{\gamma_k^3}C_h^2 C_0. \end{aligned}$$

Finally, we complete the proof by induction. Since $\alpha_k/\gamma_k \to 0$, there exists a constant $M > 0$ such that $\alpha_k/\gamma_k \le M$ for all k. Let $S_z = \|z_0 - h(x_0)\|^4 + 12D_z V_h + 16\sqrt{D_z}V_h^{3/2} + 8M^4 V_h^2 + C_h^2 C_0$. Then $\|z_0 - h(x_0)\|^4 \le S_z$, and $S_z - 12D_z\gamma_k V_h - 16\gamma_k^2\sqrt{D_z}V_h^{3/2} - 8\gamma_k^4 V_h^2 - (\alpha_k^4/\gamma_k^4)C_h^2 C_0 \ge 0$ for all k. Suppose the claim is true for $0, 1, \ldots, k$. Then

$$\begin{aligned} &\mathbb{E}[\|z_{k+1} - h(x_{k+1})\|^4] \\ &\le (1-\gamma_k)S_z + 12\gamma_k^2 D_z V_h + 16\gamma_k^3\sqrt{D_z}V_h^{3/2} + 8\gamma_k^4 V_h^2 + \frac{\alpha_k^4}{\gamma_k^3}C_h^2 C_0 \le S_z, \end{aligned}$$

which completes the proof.

(b) By the definition of z_{k+1}, we have $z_{k+1} = (1-\gamma_k)z_k + \gamma_k h_{u_{k+1}}(x_{k+1})$, and

$$(1-\gamma_k)^4\|z_{k+1} - z_k\|^4 \le 8\gamma_k^4\|h_{u_{k+1}}(x_{k+1}) - h(x_{k+1})\|^4 + 8\gamma_k^4\|z_{k+1} - h(x_{k+1})\|^4,$$

where the inequality follows from $(a+b)^4 \le (2a^2+2b^2)^2 \le 8a^4+8b^4$. Thus

$$\mathbb{E}[\|z_{k+1}-z_k\|^4] \le \frac{8\gamma_k^4}{(1-\gamma_k)^4}V_h^2 + \frac{8\gamma_k^4}{(1-\gamma_k)^4}\mathbb{E}[\|z_{k+1}-h(x_{k+1})\|^4].$$

By part (a) that there exists a constant $S_z \ge 0$ such that $\mathbb{E}[\|z_{k+1}-h(x_{k+1})\|^4] \le S_z$, we obtain

$$\mathbb{E}[\|z_{k+1}-z_k\|^4] \le \mathcal{O}(\gamma_k^4),$$

which completes the proof. ∎

A.5. Proof of Lemma 3.5. Letting Assumptions 2.1 and 3.1 hold, we apply the basic update rule to the first inner level and the accelerated update rule to the remaining levels. We consider the analysis for the second inner level, updated as

$$\begin{aligned}\widehat{y}_{k+1}^{(T-2)} &= (1-1/\beta_{T-2,k})y_k^{(T-1)} + y_{k+1}^{(T-1)}/\beta_{T-2,k},\\ y_{k+1}^{(T-2)} &= (1-\beta_{T-2,k})y_k^{(T-2)} + \beta_{T-2,k}\cdot f_{\omega_{T-1,k+1}}^{(T-1)}(\widehat{y}_{k+1}^{(T-2)}).\end{aligned}$$

For ease of notation, we denote $y_k^{(T-1)}$ by z_k, $\beta_{T-2,k}$ by β_k, $\widehat{y}_k^{(T-2)}$ by $\widehat{y}_{k+1}$, $y_k^{(T-2)}$ by y_k, and $f_{\omega_{T-1,k+1}}^{(T-1)}(\cdot)$ by $g_{w_{k+1}}(\cdot)$. The corresponding update step can be written as

$$\widehat{y}_{k+1} = \left(1-\frac{1}{\beta_k}\right)z_k + \frac{1}{\beta_k}z_{k+1}, \quad y_{k+1} = (1-\beta_k)y_k + \beta_k g_{w_{k+1}}(\widehat{y}_{k+1}).$$

Next, we prove Lemma 3.5 to construct the supermartingale with respect to an upper bound of $\{y_k - g(z_k)\}$.

Proof of Lemma 3.5. Lemma 3.5 analyzes the behavior of an accelerated update step in the a-TSCGD. Note that part (a) of this lemma can be deduced from [27, Lemma D.3]; we skip the proof to avoid repetition, and present the detailed proofs of parts (b)–(e) here.

(b) Under the condition $\sum_{k=1}^{\infty}\gamma_k^4\beta_k^{-3} < \infty$, we have

$$\sum_{k=1}^{\infty}\beta_{j,k}^{-3}\{\mathbb{E}[\|z_{k+1}-z_k\|^4|\mathbb{F}_k]\} = \sum_{k=1}^{\infty}\beta_{j,k}^{-3}\mathbb{E}[\|z_{k+1}-z_k\|]^4 \le \sum_{k=1}^{\infty}\mathcal{O}\left(\frac{\gamma_k^4}{\beta_k^3}\right) < \infty,$$

where the inequality holds by the assumption $\mathbb{E}[\|z_{k+1}-z_k\|^4] \le \mathcal{O}(\gamma_k^4)$. By the monotone convergence theorem, we obtain that $\sum_{k=1}^{n}\beta_k^{-3}\mathcal{O}(\mathbb{E}[\|z_{k+1}-z_k\|^4|\mathbb{F}_k])$ converges almost surely to some random variable with finite expectation as $n\to\infty$. Therefore, the limit $\sum_{k=1}^{\infty}\beta_k^{-3}\mathcal{O}(\mathbb{E}[\|z_k-z_{k-1}\|^4|\mathbb{F}_k])$ exists and is finite w.p.1.

(c) By part (a), we have that there exists a constant $C \ge 0$ such that

$$\mathbb{E}[e_{k+1}^2] \le (1-\beta_k/2)\mathbb{E}[e_k^2] + 2\beta_k^2 V_g + C\frac{\gamma_k^4}{\beta_k^3}. \tag{A.17}$$

Since $\gamma_k/\beta_k \to 0$, there exists an $M > 0$ such that $\gamma_k \le M\beta_k$ for all k. Letting $D_y = \mathbb{E}[e_1^2] + 4V_g + 2M^4C$, by $\gamma_k \le M\beta_k$ and $\beta_k \le 1$, we have $D_y \ge 4V_g\beta_k + 2\beta_k^{-4}\gamma_k^4 C$ for all k.

We prove by induction that $\mathbb{E}[e_k^2] \le D_y$ for all k. Clearly, the claim holds for $k=1$. Suppose the claim holds for $1,2,\ldots,k$. We have, by (A.17),

$$\mathbb{E}[e_{k+1}^2] \le \left(1-\frac{\beta_k}{2}\right)\mathbb{E}[e_k^2] + 2\beta_k^2V_g + C\frac{\gamma_k^4}{\beta_k^3} \le D_y - \frac{\beta_k}{2}(D_y - 4\beta_k V_g - 2\beta_k^{-4}\gamma_k^4 C) \le D_y,$$

where the last inequality uses the fact that $D_y - 4\beta_k V_g - 2\beta_k^{-4}\gamma_k^4 C \geq 0$ for all k. The claim thus holds as desired.

(d) By the assumption that $\mathbb{E}[\|z_{k+1} - z_k\|^4] \leq \mathcal{O}(\gamma_k^4)$, there exists a constant $C_0 > 0$ such that $\mathbb{E}[\|z_{k+1} - z_k\|^4] \leq \gamma_k^4 C_0$. By part (c), we have that there exists a constant $D_y > 0$ such that $\mathbb{E}[\|y_{k+1} - g(z_{k+1})\|^2] \leq \mathbb{E}[e_{k+1}^2] \leq D_y$ for all k. Thus, $\mathbb{E}[\|y_{k+1} - g(z_{k+1})\|] \leq \sqrt{\mathbb{E}[\|y_{k+1} - g(z_{k+1})\|^2]} \leq \sqrt{D_y}$. Now we begin our analysis to show the finiteness of the fourth moment, $\mathbb{E}[\|y_{k+1} - g(z_{k+1})\|^4]$.

Let $e_{k+1} = (1-\beta_k)(g(z_{k+1}) - g(z_k))$. By the Lipschitz continuity of g, we have

$$\|e_{k+1}\| \leq (1-\beta_k)\sqrt{C_g}\|z_{k+1} - z_k\|.$$

By the definition of y_{k+1}, we have

$$\begin{aligned} &y_{k+1} - g(z_{k+1}) + e_{k+1} - e_{k+1} \\ &= (1-\beta_k)(y_k - g(z_k)) \\ &\quad + \beta_k(g_{w_{k+1}}(\widehat{y}_{k+1}) - g(\widehat{y}_{k+1})) + \beta_k(g(\widehat{y}_{k+1}) - g(z_{k+1})) - e_{k+1}. \end{aligned}$$

Again, by the Lipschitz continuity of g, we get

$$\beta_k\|g(\widehat{y}_{k+1}) - g(z_{k+1})\| \leq \beta_k\sqrt{C_g}\|\widehat{y}_{k+1} - z_{k+1}\| = (1-\beta_k)\sqrt{C_g}\|z_{k+1} - z_k\|.$$

So we obtain

$$\begin{aligned} \|y_{k+1} - g(z_{k+1})\| \leq \|(1-\beta_k)(y_k - g(z_k)) + \beta_k(g_{w_{k+1}}(\widehat{y}_{k+1}) - g(\widehat{y}_{k+1}))\| \\ + 2(1-\beta_k)\sqrt{C_g}\|z_{k+1} - z_k\|. \end{aligned}$$

Let $P_k = \|(1-\beta_k)(y_k - g(z_k)) + \beta_k(g_{w_{k+1}}(\widehat{y}_{k+1}) - g(\widehat{y}_{k+1}))\|$, then we have

$$\begin{aligned} P_k^4 &= \|(1-\beta_k)(y_k - g(z_k)) + \beta_k(g_{w_{k+1}}(\widehat{y}_{k+1}) - g(\widehat{y}_{k+1}))\|^4 \\ &\leq (1-\beta_k)^4\|y_k - g(z_k)\|^4 + 4(1-\beta_k)^3\beta_k(y_k - g(z_k))^3(g_{w_{k+1}}(\widehat{y}_{k+1}) - g(\widehat{y}_{k+1})) \\ &\quad + 6(1-\beta_k)^2\beta_k^2\|y_k - g(z_k)\|^2\|g_{w_{k+1}}(\widehat{y}_{k+1}) - g(\widehat{y}_{k+1})\|^2 \\ &\quad + 4(1-\beta_k)\beta_k^3\|(y_k - g(z_k))\|\|(g_{w_{k+1}}(\widehat{y}_{k+1}) - g(\widehat{y}_{k+1}))\|^3 \\ &\quad + \beta_k^4\|g_{w_{k+1}}(\widehat{y}_{k+1}) - g(\widehat{y}_{k+1})\|^4, \end{aligned}$$

which implies

$$\mathbb{E}[P_k^4] \leq (1-\beta_k)^4\mathbb{E}[\|y_k - g(z_k)\|^4] + 6(1-\beta_k)^2\beta_k^2 D_y V_g + 4\beta_k^3\sqrt{D_y}V_g^{3/2} + \beta_k^4 V_g^2.$$

Using the fact that $\|a+b\|^2 \leq (1+\varepsilon)\|a\|^2 + (1+1/\varepsilon)\|b\|^2$ and $\|a+b\|^4 \leq (1+\varepsilon)^3\|a\|^4 + (1+1/\varepsilon)^3\|b\|^4$ for $\varepsilon > 0$, we obtain

$$\|y_{k+1} - g(z_{k+1})\|^4 \leq (1+\beta_k)^3 P_k^4 + 16(1+1/\beta_k)^3(1-\beta_k)^4 C_g^2\|z_{k+1} - z_k\|^4,$$

which implies

$$\begin{aligned} \mathbb{E}[\|y_{k+1} - g(z_{k+1})\|^4] \leq (1-\beta_k)\mathbb{E}[\|y_k - g(z_k)\|^4] \\ + 12\beta_k^2 D_y V_g + 16\beta_k^3\sqrt{D_y}V_g^{3/2} + 8\beta_k^4 V_g^2 + \frac{16\gamma_k^4}{\beta_k^3}C_g^2 C_0. \end{aligned}$$

Since $\gamma_k/\beta_k \to 0$, there exists a constant $M > 0$ such that $\gamma_k \le \beta_k M$ for all k. We set $S_y = \|y_0 - g(z_0)\|^4 + 12D_yV_g + 16\sqrt{D_y}V_g^{3/2} + 8V_g^2 + 16M^4C_g^2C_0$, and prove the claim by induction on k. The rest of the analysis follows the same argument as in the proof of Lemma 3.4(a). We omit the details to avoid repetition. We conclude that there exists a constant $S_y > 0$ such that $\mathbb{E}[\|y_{k+1} - g(z_{k+1})\|^4] \le S_y$ for all k.

(e) By the definition of $y_k, \widehat{y}_k, z_k$, we have

$$\begin{aligned}||y_{k+1} - y_k|| &= ||(1-\beta_{k+1})y_k + \beta_k g_{w_{k+1}}(\widehat{y}_{k+1}) - y_k|| \\ &\le \beta_k[||g_{w_{k+1}}(\widehat{y}_{k+1}) - g(\widehat{y}_{k+1})|| + L_g||\widehat{y}_{k+1} - z_k|| + ||g(z_k) - y_k||] \\ &= \beta_k||g_{w_{k+1}}(\widehat{y}_{k+1}) - g(\widehat{y}_{k+1})|| + L_g||z_{k+1} - z_k|| + \beta_k||g(z_k) - y_k||.\end{aligned}$$

Using the fact that $(a+b+c)^4 \le [2(a+b)^2+2c^2]^2 \le 8(a+b)^4+8c^4 \le 64a^4+64b^4+8c^4$, it is easy to see that

$$||y_{k+1}-y_k||^4 \le 64\beta_k^4||g_{w_{k+1}}(\widehat{y}_{k+1})-g(\widehat{y}_{k+1})||^4+64L_g^4||z_{k+1}-z_k||^4+8\beta_k^4||g(z_k)-y_k||^4.$$

Then we get

$$\mathbb{E}[||y_{k+1} - y_k||^2] \le 64\beta_k^4V_g^2 + 64L_g^2\mathbb{E}[||z_{k+1} - z_k||^4] + 8\beta_k^4\mathbb{E}[||g(z_k) - y_k||^4], \tag{A.18}$$

By part (d), we have that $\mathbb{E}[\|y_k - g(z_k)\|^4] \le S_y$. Since $\mathbb{E}[\|z_{k+1} - z_k\|^4] \le \mathcal{O}(\gamma_k^4)$ and $\gamma_k/\beta_k \to 0$, we obtain

$$\mathbb{E}[\|y_{k+1} - y_k\|^4] \le \mathcal{O}(\beta_k^4) + \mathcal{O}(\gamma_k^4) \le \mathcal{O}(\beta_k^4),$$

which concludes the proof. ∎

A.6. Proof of Lemma 3.6. This lemma proves the T-element supermartingale convergent lemma to establish convergence property of $\{x_{k+1} - x^*\}$.

Proof. Let J_k be the random variable $J_k \equiv X_k + c_{T-1}Y_k^{(T-1)} + c_{T-2}Y_k^{(T-2)} + \cdots + c_1Y_k^{(1)}$. We have

$$\begin{aligned}\mathbb{E}[J_{k+1}|\mathbb{F}_k] &= \mathbb{E}[X_{k+1}|\mathbb{F}_k] + c_{T-1}\mathbb{E}[Y_{k+1}^{(T-1)}|\mathbb{F}_k] + \cdots + c_1\mathbb{E}[Y_{k+1}^{(1)}|\mathbb{F}_k] \\ &\le (1+\eta_k)J_k + \mu_k^{(T)} + c_{T-1}\mu_k^{(T-1)} + \cdots + c_1\mu_k^{(1)}.\end{aligned}$$

Since $\sum_{k=0}^{\infty}\mu_k^{(T)} + c_{T-1}\mu_k^{(T-1)} + \cdots + c_1\mu_k^{(1)} < \infty$ and $\sum_{k=0}^{\infty}\eta_k < \infty$, we obtain that J_k converges almost surely to a random variable by [26, Theorem 1], and J_k is bounded by a constant w.p.1.

By the definition above, we have that $X_k \le J_k$, $Y_k^{(2)} \le \frac{1}{c_2}J_k, \ldots, Y_k^{(T)} \le \frac{1}{c_T}J_k$. Then $X_k, Y_k^{(T-1)}, Y_k^{(T-2)}, \ldots, Y_k^{(1)}$ are also bounded w.p.1. Since $\sum_{k=0}^{\infty}\mu_k^{(j)} < \infty$ for $j = 1, \ldots, T$, and $\theta_k^{(T-1)}, \ldots, \theta_k^{(1)}$ are nonnegative, we have

$$\mathbb{E}[Y_{k+1}^{(T-1)}|\mathbb{F}_k] \le (1-\theta_k^{(j)})Y_k^{(j)} + \mu_k^{(j)} \quad \text{for } j = T-1, \ldots, 1.$$

Again, by [26, Theorem 1], we obtain that $Y_k^{(T-1)}, \ldots, Y_k^{(1)}$ converge almost surely to $T-1$ random variables, and $\sum_{k=1}^{\infty}\theta_k Y_k^{(j)} \le \infty$, w.p.1 for $j = 1, \ldots, T-1$. Since $Y_k^{(1)}, \ldots, Y_k^{(T-1)}$ and $J_k = X_k + c_{T-1}Y_k^{(T-1)} + c_{T-2}Y_k^{(T-2)} + \cdots + c_1Y_k^{(1)}$ are almost

surely convergent, X_k must converge almost surely to a random variable. Together with the conditions $\sum_{k=0}^{\infty}\eta_k<\infty$, $\sum_{k=0}^{\infty}\mu_k^{(j)}<\infty$ for $j=1,\dots,T$, we have

$$\sum_{j=1}^{T}\sum_{k=0}^{\infty}u_k^{(j)}<\infty, \qquad \text{w.p.1.}$$

So this completes the proof. ∎

A.7. Proof of Lemma 3.7. Lemma 3.7 can be deduced from the proof of [27, Theorem 1(a)]; we omit the details here to avoid repetition. ∎

A.8. Proof of Lemma 3.8. We focus on a single trajectory $\{x_k(\omega)\}$ generated by the algorithm such that the preceding inequality, $\sum_{k=0}^{\infty}\alpha_k\|\nabla F(x_k)\|^2<\infty$, holds, since this event happens w.p.1. Denote by A the event that there exists a limiting point that is not a stationary point. For simplicity, we omit the notation (ω) in the rest of the proof.

Letting $\varepsilon>0$ be arbitrary, we note that $\|\nabla F(x_k)\|\le\varepsilon$ holds for infinitely many k. Otherwise, we would have, for some $\bar{k}>0$, that $\sum_{k=0}^{\infty}\alpha_k\|\nabla F(x_k)\|^2\ge\sum_{k=\bar{k}}^{\infty}\alpha_k\varepsilon^2=\infty$, yielding a contradiction. Consequently, there exists a closed set $\bar{N}$ (e.g., a closed union of neighborhoods of all ε-stationary x_k's) such that $\{x_k\}$ visits infinitely often, and

$$\|\nabla F(x)\|\begin{cases}\le\varepsilon & \text{if } x\in\bar{N},\\ >\varepsilon & \text{if } x\notin\bar{N},\ x\in\{x_k\}.\end{cases}$$

Under event A, there exists a limit point $\widetilde{x}$ such that $\|\nabla F(\widetilde{x})\|>2\varepsilon$. Then there exists a closed set $\widetilde{N}$ (e.g., a union of neighborhoods of all x_k such that $\|\nabla F(x_k)\|>2\varepsilon$) such that $\{x_k\}$ visits infinitely often, and

$$\|\nabla F(x)\|\begin{cases}\ge 2\varepsilon & \text{if } x\in\widetilde{N},\\ <2\varepsilon & \text{if } x\notin\widetilde{N},\ x\in\{x_k\}.\end{cases}$$

By the continuity of ∇F and $\varepsilon>0$, we obtain that the sets $\widetilde{N}$ and $\bar{N}$ are disjoint, i.e., $\operatorname{dist}(\bar{N},\widetilde{N})>0$. Since the sequence $\{x_k\}$ enters both sets $\bar{N}$ and $\widetilde{N}$ infinitely often, there exists a subsequence

$$\{x_k\}_{k\in\mathcal{K}}=\{\{x_k\}_{k=s_i}^{t_i-1}\}_{i=1}^{\infty}$$

that crosses the two sets infinitely often, with $x_{s_i}\in\widetilde{N}$, $x_{t_i}\in\widetilde{N}$ for all i. In other words, we have for every i that

$$\|\nabla F(x_{s_i})\|\ge 2\varepsilon>\|\nabla F(x_k)\|>\varepsilon\ge\|\nabla F(x_{t_i})\| \qquad \forall k=s_i+1,\dots,t_i-1.$$

By using the triangle inequality, we have

$$\begin{aligned}\sum_{k\in\mathcal{K}}\|x_{k+1}-x_k\|&=\sum_{i=1}^{\infty}\sum_{k=s_i}^{t_i-1}\|x_{k+1}-x_k\|\ge\sum_{i=1}^{\infty}\|x_{t_i}-x_{s_i}\|\\&\ge\sum_{i=1}^{\infty}\operatorname{dist}(\bar{N},\widetilde{N})=\infty.\end{aligned}$$

Denote by A_1 the event $\{\sum_{k\in\mathcal{K}} \|x_{k+1}-x_k\| = \infty\}$. Clearly, A implies A_1 and $A \subset A_1$. On the other hand, we have

$$\infty > \sum_{k=0}^{\infty} \alpha_k \|\nabla f(x_k)\|^2 \geq \sum_{k\in\mathcal{K}} \alpha_k \|\nabla f(x_k)^2\| \geq \varepsilon^2 \sum_{k\in\mathcal{K}} \alpha_k.$$

However, we can further obtain that $\sum_{k\in\mathcal{K}} \mathbb{E}[\|x_{k+1}-x_k\|] = \mathcal{O}(\sum_{k\in\mathcal{K}} \alpha_k) < \infty$. This, together with the monotone convergence theorem, implies that, on this subsequence $\mathcal{K}$, we have

$$\sum_{k\in\mathcal{K}} \|x_{k+1} - x_k\| < \infty, \qquad \text{w.p.1.}$$

Denote A_2 as the event $\{\sum_{k\in\mathcal{K}} \|x_{k+1}-x_k\| < \infty\}$. Then A implies A_2 w.p.1. However, as A_1 and A_2 are disjoint, we have $\mathbb{P}(A) = 0$.

Since ε can be arbitrarily small, we conclude that any limiting point of $\{x_k\}$ is a stationary point w.p.1. ∎

A.9. Proof of Theorem 3.1. (a) Let x^* be an arbitrary optimal solution to problem (1.1), and let $F^* = F(x^*)$. By the same argument as in the proof of Lemma 3.1, we obtain that there exists $C_0 > 0$ such that

(A.19)
$$\begin{aligned}\mathbb{E}[\|x_{k+1} - x^*\|^2|\mathbb{F}_k] \leq{}& \left(1 + C_0\left(\frac{\alpha_k^2}{\beta_{T-1,k}} + \frac{\alpha_k^2}{\beta_{T-2,k}} + \cdots + \frac{\alpha_k^2}{\beta_{1,k}}\right)\right)\|x_k - x^*\|^2 \\ &+ \alpha_k^2 C_1 C_2 \cdots C_T - 2\alpha_k(F(x_k) - F^*) \\ &+ (T-1)\beta_{T-1,k}\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\||\mathbb{F}_k] \\ &+ (T-2)\beta_{T-2,k}\mathbb{E}[\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|^2|\mathbb{F}_k] \\ &+ \cdots + \beta_{1,k}\mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2|\mathbb{F}_k].\end{aligned}$$

First, we consider the case when the first inner-level function $f^{(T)}$ does not have Lipschitz continuous gradients. In this case, Algorithm 2 runs with the basic update step for the first inner level and accelerated update steps for all other levels. By Assumption 3.1, we have

$$\mathbb{E}[\|x_{k+1} - x_k\|^4] \leq \alpha_k^4 C_1^2 C_2^2 \cdots C_T^2,$$

which is the sufficient condition for Lemma 3.4 to be true. Applying Lemmas 3.3 and 3.4 to the first update step, we have

(A.20)
$$\begin{aligned}&\mathbb{E}[\mathbb{E}[\|y_{k+1}^{(T-1)} - f^{(T)}(x_{k+1})\|^2|\mathbb{F}_{k+1}]|\mathbb{F}_k] \\ &\leq (1-\beta_{T-1,k})\mathbb{E}[\|y_{k-1}^{(T-1)} - f^{(T)}(x_{k-1})\|^2|\mathbb{F}_k] \\ &\quad + \beta_{T-1}^{-1} C_T \mathbb{E}[\|x_k - x_{k-1}\|^2|\mathbb{F}_k] + 2V_T\beta_{T-1,k}^2,\end{aligned}$$

and $\mathbb{E}[\|y_{k+1}^{(T-1)} - y_k^{(T-1)}\|^4] \leq \mathcal{O}(\beta_{T-1,k}^4)$, which serves as the sufficient condition for level $(T-1)$ in Lemma 3.5 to be true so that $\mathbb{E}[\|y_{k+1}^{(T-2)} - y_k^{(T-2)}\|^4] \leq \mathcal{O}(\beta_{T-2,k}^4)$.

Applying Lemma 3.5 recursively for the accelerated update from $j = T-2$ to 1, we have that, for all j and all k, there exists an $e_k^{(j)} \in \mathbb{F}_{k+1}$ such that almost surely

$$\begin{aligned}
&\mathbb{E}[\|y_k^{(j)} - f^{(j+1)}(y_k^{(j+1)})\|^2|\mathbb{F}_k] \le \mathbb{E}[[e_k^{(j)}]^2|\mathbb{F}_k],\\
&\mathbb{E}[\mathbb{E}[[e_{k+1}^{(j)}]^2|\mathbb{F}_{k+1}]|\mathbb{F}_k] \le (1-\beta_{j,k}/2)\mathbb{E}[[e_k^{(j)}]^2|\mathbb{F}_k] + 2\beta_{j,k}^2 V_{j+1}\\
&\qquad\qquad + \mathcal{O}\Big(\frac{\mathbb{E}[\|y_{k+1}^{(j+1)} - y_k^{(j+1)}\|^4|\mathbb{F}_k]}{\beta_{j,k}^3}\Big),
\end{aligned} \tag{A.21}$$

and $\mathbb{E}[\|y_{k+1}^{(j)} - y_k^{(j)}\|^4] \le \mathcal{O}(\beta_{j,k}^4)$.

By Lemma 3.5(b), under the condition that $\sum_{k=1}^{\infty} \frac{\beta_{j+1,k}^4}{\beta_{j,k}^3} < \infty$, we have

$$\sum_{k=1}^{\infty} \frac{\mathbb{E}[\|y_{k+1}^{(j+1)} - y_k^{(j+1)}\|^4|\mathbb{F}_k]}{\beta_{j,k}^3} < \infty,$$

w.p.1. Together with the condition $\sum_{k=1}^{\infty} \beta_{j,k}^2 < \infty$, the sum of the tail part of this supermartingale,

$$2\beta_{j,k}^2 V_{j+1} + \mathcal{O}\Big(\frac{\mathbb{E}[\|y_{k+1}^{(j+1)} - y_k^{(j+1)}\|^4|\mathbb{F}_k]}{\beta_{j,k}^3}\Big),$$

converges almost surely.

Similarly, by Lemma 3.3(b), under the condition $\sum_{k=1}^{\infty} \frac{\alpha_k^2}{\beta_{T-1,k}} < \infty$, we have that, w.p.1,

$$\sum_{k=1}^{\infty} \frac{\mathbb{E}[\|x_{k+1} - x_k\|^2|\mathbb{F}_k]}{\beta_{T-1,k}} < \infty.$$

Together with the condition $\sum_{k=1}^{\infty} \beta_{T-1,k}^2 < \infty$, the sum of the tail part of the supermartingale for (A.20), $2V_T\beta_k^2 + \frac{C_T\mathbb{E}[\|x_{k+1}-x_k\|^2|\mathbb{F}_k]}{\beta_{T-1,k}}$, converges almost surely.

Now we apply the T-element supermartingale convergent lemma to (A.20), (A.19), and (A.21). By letting

$$\begin{aligned}
&X_k = \|x_k - x^*\|^2, \qquad Y_k^{(T-1)} = \mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2|\mathbb{F}_k],\\
&Y_k^{(T-2)} = \mathbb{E}[[e_k^{(T-2)}]^2|\mathbb{F}_k],\ \ldots,\ Y_k^{(1)} = \mathbb{E}[[e_k^{(1)}]^2|\mathbb{F}_k],\\
&\eta_k = \Big[\frac{\alpha_k^2}{\beta_{T-1,k}} + \cdots + \frac{\alpha_k^2}{\beta_{1,k}}\Big]C_0,\ u_k^{(T)} = 2\alpha_k(F(x_k) - F^*),\\
&u_k^{(1)} = u_k^{(2)} = \cdots = u_k^{(T-1)} = 0,\ c_1 = 2,\ \ldots,\ c_{T-2} = 2(T-2),\ c_{T-1} = T-1,\\
&\mu_k^{(1)} = 2\beta_{1,k}^2 V_1 + \mathcal{O}\Big(\frac{\mathbb{E}[\|y_{k+1}^{(2)} - y_k^{(2)}\|^4|\mathbb{F}_k]}{\beta_{1,k}^3}\Big),\ \ldots,\\
&\mu_k^{(T-2)} = 2\beta_{T-2,k}^2 V_{T-1} + \mathcal{O}\Big(\frac{\mathbb{E}[\|y_{k+1}^{(T-1)} - y_k^{(T-1)}\|^4|\mathbb{F}_k]}{\beta_{T-2,k}^3}\Big),\\
&\mu_k^{(T-1)} = C_T\beta_{T-1,k}^{-1}\mathbb{E}[\|x_{k+1} - x_k\|^2|\mathbb{F}_k] + 2V_T\beta_{T-1,k}^2,\\
&\mu_k^{(T)} = \alpha_k^2 C_1 C_2 \cdots C_T,\\
&\theta_k^{(1)} = \beta_{1,k}/2,\ \ldots,\ \theta_k^{(T-2)} = \beta_{T-2,k}/2,\ \theta_k^{(T-1)} = \beta_{T-1,k},
\end{aligned}$$

we obtain that $\|x_{k+1}-x^*\|$ converges almost surely to a nonnegative random variable, and

$$\sum_{k=0}^{\infty}\alpha_k(F(x_k)-F^*)<\infty,$$

which further implies

$$\liminf_{k\to\infty}F(x_k)=F^*, \qquad \text{w.p.1.}$$

Using Lemma 3.7, we conclude that the sequence $\{x_k\}$ converges almost surely to a random point in the set of optimal solutions to problem (1.1).

Next, we study the case when Assumption 3.2 also holds, i.e., $f^{(T)}$ has Lipschitz continuous gradient. In this case, Algorithm 2 runs with accelerating update steps for all levels. The only difference is that we use the accelerated update rule for the first inner level instead of the basic one, so Lemmas 3.1 and 3.5 still hold. Consider the first inner level. Since it is also updated by the accelerated update rule, we apply a similar analysis to that in Lemma 3.5 for this level. We have that there exists $e_k^{(T-1)}\in\mathbb{F}_{k+1}$ such that, w.p.1,

$$\mathbb{E}[\|y_k^{(T-1)}-f^{(T)}(x_k)\|^2|\mathbb{F}_k]\le\mathbb{E}[[e_k^{(T-1)}]^2|\mathbb{F}_k],$$

and

$$\begin{aligned}&\mathbb{E}[\mathbb{E}[[e_{k+1}^{(T-1)}]^2|\mathbb{F}_{k+1}]|\mathbb{F}_k]\\&\le\left(1-\frac{\beta_{j,k}}{2}\right)\mathbb{E}[[e_k^{(T-1)}]^2|\mathbb{F}_k]+2\beta_{j,k}^2V_{j+1}+\mathcal{O}\left(\frac{\mathbb{E}[\|x_{k+1}-x_k\|^4|\mathbb{F}_k]}{\beta_{j,k}^3}\right).\end{aligned}\tag{A.22}$$

By similar argument as in Lemma 3.5(b), under the condition $\sum_{k=1}^{\infty}\frac{\alpha_k^4}{\beta_{T-1,k}^3}<\infty$, we have, w.p.1,

$$\sum_{k=1}^{\infty}\frac{\mathbb{E}[\|x_{k+1}-x_k\|^4|\mathbb{F}_k]}{\beta_{j,k}^3}<\infty.$$

Now we apply the T-element supermartingale convergent lemma to (A.19), (A.21), and (A.22). The remaining part follows the same lines as the case when $f^{(T)}$ does not have Lipschitz continuous gradient. We conclude that $\{x_k\}$ converges almost surely to a random optimal solution.

(b) First, consider the case when $f^{(T)}$ does not have Lipschitz continuous gradient. Since problem (1.1) has at least one optimal solution, the function F is bounded from below, and we denote by F^* the optimal value of $F(x)$ over $\mathcal{X}$. As a result, we can treat $\{F(x_k)-F^*\}$ as nonnegative random variables. The x_k update steps of Algorithms 2 and 1 are the same. Thus, Lemma 3.2 holds for Algorithm 2 as well. It follows that, for sufficiently large k,

$$\begin{aligned}\mathbb{E}[F(x_{k+1})-F^*|\mathbb{F}_k]\le F(x_k)-F^*-\frac{\alpha_k}{2}\|\nabla F(x_k)\|^2+\frac{1}{2}\alpha_k^2L_FC_1C_2\cdots C_T\\+(T-1)\beta_{T-1,k}\mathbb{E}[\|y_k^{(T-1)}-f^{(T)}(x_k)\|^2|\mathbb{F}_k]\\+\cdots+\beta_{1,k}\mathbb{E}[\|y_k^{(1)}-f^{(2)}(y_k^{(2)})\|^2|\mathbb{F}_k].\end{aligned}\tag{A.23}$$

Using a similar argument to that in part (a), we apply the T-element supermartingale

convergence lemma to (A.21), (A.22), and (A.23). By letting

$$X_k = F(x_k) - F^*, \qquad Y_k^{(T-1)} = \mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2|\mathbb{F}_k],$$
$$Y_k^{(T-2)} = \mathbb{E}[[e_k^{(T-2)}]^2|\mathbb{F}_k], \ \ldots, \ Y_k^{(1)} = \mathbb{E}[[e_k^{(1)}]^2|\mathbb{F}_k],$$
$$\eta_k = 0, \qquad u_k^{(T)} = \frac{1}{2}\alpha_k\|\nabla F(x)\|^2,$$
$$u_k^{(1)} = u_k^{(2)} = \cdots = u_k^{(T-1)} = 0, \qquad c_1 = 2, \ \ldots, \ c_{T-2} = 2(T-2), \ c_{T-1} = T-1,$$
$$\mu_k^{(T-1)} = C_T\beta_{T-1,k}^{-1}\mathbb{E}[\|x_{k+1} - x_k\|^2|\mathbb{F}_k] + 2V_T\beta_{T-1,k}^2,$$
$$\mu_k^{(T-2)} = 2\beta_{T-2,k}^2 V_{T-1} + \mathcal{O}\bigg(\frac{\mathbb{E}[\|y_{k+1}^{(T-1)} - y_k^{(T-1)}\|^4|\mathbb{F}_k]}{\beta_{T-2,k}^3}\bigg), \ \ldots,$$
$$\mu_k^{(1)} = 2\beta_{1,k}^2 V_1 + \mathcal{O}\bigg(\frac{\mathbb{E}[\|y_{k+1}^{(2)} - y_k^{(2)}\|^4|\mathbb{F}_k]}{\beta_{1,k}^3}\bigg),$$
$$\mu_k^{(T)} = \frac{1}{2}\alpha_k^2 L_F C_1 C_2 \cdots C_T,$$
$$\theta_k^{(1)} = \beta_{1,k}/2, \ \ldots, \ \theta_k^{(T-2)} = \beta_{T-2,k}/2, \ \theta_k^{(T-1)} = \beta_{T-1,k},$$

we obtain that $\{F(x_k) - F^*\}$ converges almost surely to a nonnegative random variable, and

$$\sum_{k=0}^{\infty} \alpha_k\|\nabla F(x_k)\|^2 < \infty, \qquad \text{w.p.1.}$$

By Lemma 3.8, we conclude that any limit point of the sequence $\{x_k\}$ is a stationary point w.p.1, which completes the proof.

When $f^{(T)}$ has Lipschitz continuous gradient, we apply the T-element supermartingale convergent lemma to (A.21), (A.22), and (A.23). The rest of the proof follows the same lines as in the case when $f^{(T)}$ is nonsmooth. ∎

Appendix B. Proof of Theorem 3.2. In this section, we present the detailed proof for Theorem 3.2.

Note that we let the step sizes be $\alpha_k = k^{-a}$, $\beta_{T-1,k} = k^{-b_{T-1}}, \beta_{T-2,k} = k^{-b_{T-2}}$, $\ldots, \beta_{1,k} = k^{-b_1}$. We slightly modify [29, Lemma 5] to help us derive the convergence rates.

LEMMA B.1. *Given a sequence of positive real numbers $\{w_k\}_{k=1}^{\infty}$ satisfying*

$$w_{k+1} \le (1 - C_1\beta_k)w_k + C_2k^{-a},$$

where $C_1 \ge 0$, $C_2 \ge 0$, and $a \ge 0$, and choosing $c = a - b$ and β_k to be $\beta_k = C_3k^{-b}$, where $b \in (0, 1]$ and $C_3 > c/C_1$, the sequence can be bounded by $w_k \le Ck^{-c}$, where C is defined as

$$C := w_0 + \frac{C_2}{C_1C_3 - c}.$$

In other words, we have

$$w_k \le \mathcal{O}(k^{-a+b}).$$

Proof. We prove the lemma by induction. Clearly, the claim holds for $k = 0$. Next, supposing the claim holds for "k," we prove it is also true for "$k+1$." That is, given $w_k \le Ck^{-c}$, we need to prove $w_{k+1} \le C(k+1)^{-c}$. We have

$$w_{k+1} \le (1 - C_1C_3k^{-b})Ck^{-c} + C_2k^{-a} = Ck^{-c} - CC_1C_3k^{-b-c} + C_2k^{-a}. \tag{B.1}$$

To prove (B.1) is bounded by $C(k+1)^{-c}$, it suffices to show that

$$\Delta := (k+1)^{-c} - k^{-c} + C_1C_3k^{-b-c} > 0 \text{ and } C \ge \frac{C_2k^{-a}}{\Delta}.$$

From the convexity of function $h(t) = t^{-c}$, we have the inequality $(k+1)^{-c} - k^{-c} \ge -ck^{-c-1}$. Therefore, we obtain

$$\Delta \ge -ck^{-c-1} + C_3k^{-b-c} \ge (C_1C_3 - c)k^{-b-c} \ge 0.$$

To verify the second claim, we have

$$\frac{C_2k^{-a}}{\Delta} \le \frac{C_2}{C_1C_3 - 2}k^{-a+b+c} = \frac{C_2}{C_1C_3 - c} \le C,$$

where the equality comes from the definition that $c = a - b$ and the last inequality holds by the definition of C. This completes the proof. ∎

B.1. Proof of Lemma 3.9. By Lemma 3.3, considering the basic update step for the first inner level, we have $\mathbb{E}[\|x_{k+1} - x_k\|^2] \le \mathcal{O}(\alpha_k^2)$, and

$$\begin{aligned}&\mathbb{E}[\|y_{k+1}^{(T-1)} - f^{(T)}(x_{k+1})\|^2|\mathbb{F}_{k+1}]\\ &\le (1 - \beta_{T-1,k})\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2 + \beta_{T-1}^{-1}C_T\mathbb{E}[\|x_{k+1} - x_k\|^2|\mathbb{F}_{k+1}] + 2V_T\beta_{T-1,k}^2,\end{aligned}$$

w.p.1. Thus,

$$\begin{aligned}\mathbb{E}[\|y_{k+1}^{(T-1)} - f^{(T)}(x_{k+1})\|^2] \le (1 - \beta_{T-1,k})\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2]\\ + \beta_{T-1}^{-1}\mathcal{O}(\alpha_k^2) + 2V_T\beta_{T-1,k}^2.\end{aligned} \tag{B.2}$$

Substituting $\alpha_k = k^{-a}$ and $\beta_{T-1,k} = k^{-b_{T-1}}$ into (B.2), we get

$$\begin{aligned}\mathbb{E}[\|y_{k+1}^{(T-1)} - f^{(T)}(x_{k+1})\|^2] \le (1 - k^{-b_{T-1}})\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2]\\ + \mathcal{O}(k^{-2a+b_{T_1}}) + 2V_Tk^{-2b_{T-1}}.\end{aligned}$$

By Lemma B.1, we obtain

$$\mathbb{E}[\|y_{k+1}^{(T-1)} - f^{(T)}(x_{k+1})\|^2] \le \mathcal{O}(k^{-2a+2b_{T-1}}) + \mathcal{O}(k^{-b_{T-1}}).$$ ∎

B.2. Proof of Lemma 3.10. By Lemma 3.5, we have that, for $j = T-2, \dots, 1$, there exist random variables $e_k^{(j)} \in \mathbb{F}_{k+1}$ for all k satisfying $\|y_k^{(j)} - f^{(j+1)}(y_k^{(j+1)})\| \le e_k^{(j)}$ such that

$$\mathbb{E}[[e_{k+1}^{(j)}]^2|\mathbb{F}_k] \le (1 - \beta_{j,k}/2)[e_k^{(j)}]^2 + 2\beta_{j,k}^2V_{j+1} + \mathcal{O}\Big(\frac{\mathbb{E}[\|y_{k+1}^{(j+1)} - y_k^{(j+1)}\|^4|\mathbb{F}_k]}{\beta_{j,k}^3}\Big)$$

almost surely, and

$$\mathbb{E}[\|y_{k+1}^{(j)} - y_k^{(j)}\|^4] \le \mathcal{O}(\beta_{j,k}^4).$$

So we have

$$\mathbb{E}[[e_{k+1}^{(j)}]^2] \le (1-\beta_{j,k}/2)\mathbb{E}[[e_k^{(j)}]^2] + 2\beta_{j,k}^2 V_{j+1} + \mathcal{O}\Big(\frac{\beta_{j+1,k}^4}{\beta_{j,k}^3}\Big). \tag{B.3}$$

Substituting $\beta_{j,k} = 2k^{-b_j}$ into (B.3) and applying Lemma B.1, we obtain

$$\mathbb{E}[\|y_k^{(j)} - f^{(j+1)}(y_k^{(j+1)})\|^2] \le \mathbb{E}[[e_k^{(j)}]^2] \le \mathcal{O}(k^{-4b_{j+1}+4b_j}) + \mathcal{O}(k^{-b_j}),$$

which completes the proof. ∎

B.3. Proof of Lemma 3.11. By the Lipschitz continuous gradient condition in Assumption 2.2, we have

$$\begin{aligned}
&F(x_{k+1}) - F(x_k)\\
&\le \langle \nabla F(x_k), x_{k+1} - x_k\rangle + \frac{L_F}{2}\|x_{k+1} - x_k\|^2\\
&\le -\alpha_k \|\nabla F(x_k)\|^2\\
&\quad + \alpha_k \langle \nabla F(x_k), \nabla F(x_k) - \widetilde{\nabla} f_{\omega_{T,k}}^{(T)}(x_k)\nabla f_{\omega_{T-1,k}}^{(T-1)}(y_k^{(T-1)})\cdots\nabla f_{\omega_{1,k}}^{(1)}(y_k^{(1)})\rangle\\
&\quad + \mathcal{O}(\alpha_k^2).
\end{aligned} \tag{B.4}$$

Letting

$$Q = \langle \nabla F(x_k), \nabla F(x_k) - \widetilde{\nabla} f_{\omega_{T,k}}^{(T)}(x_k)\nabla f_{\omega_{T-1,k}}^{(T-1)}(y_k^{(T-1)})\cdots\nabla f_{\omega_{1,k}}^{(1)}(y_k^{(1)})\rangle,$$

we have

$$\mathbb{E}[Q] = \mathbb{E}[\langle \nabla F(x_k), \widetilde{\nabla} F_{\omega_k}(x_k) - \widetilde{\nabla} f_{\omega_{T,k}}^{(T)}(x_k)\nabla f_{\omega_{T-1,k}}^{(T-1)}(y_k^{(T-1)})\cdots\nabla f_{\omega_{1,k}}^{(1)}(y_k^{(1)})\rangle],$$

where $\widetilde{\nabla} F_{\omega_k}(x_k) = \widetilde{\nabla} f_{\omega_{T,k}}^{(T)}(x_k)\nabla f_{\omega_{T-1,k}}^{(T-1)}(f^{(T)}(x_k))\cdots\nabla f_{\omega_{1,k}}^{(1)}(f^{(2)}\circ\cdots\circ(f^{(T)}(x_k)))$ and the equality comes from Assumption 2.1(ii). Based on the fact that $2ab \le a^2 + b^2$ for all a, b, we obtain

$$\begin{aligned}
&\mathbb{E}[\langle \nabla F(x_k), \widetilde{\nabla} F_{\omega_k}(x_k) - \widetilde{\nabla} f_{\omega_{T,k}}^{(T)}(x_k)\nabla f_{\omega_{T-1,k}}^{(T-1)}(y_k^{(T-1)})\cdots\nabla f_{\omega_{1,k}}^{(1)}(y_k^{(1)})\rangle]\\
&\le \frac{1}{2}\mathbb{E}[\|\nabla F(x_k)\|^2]\\
&\quad + \frac{1}{2}\mathbb{E}[\|\widetilde{\nabla} F_{\omega_k}(x_k) - \widetilde{\nabla} f_{\omega_{T,k}}^{(T)}(x_k)\nabla f_{\omega_{T-1,k}}^{(T-1)}(y_k^{(T-1)})\cdots\nabla f_{\omega_{1,k}}^{(1)}(y_k^{(1)})\|^2].
\end{aligned}$$

Applying the inequality $\|a+b\|^2 \le (\|a\| + \|b\|)^2 \le 2\|a\|^2 + 2\|b\|^2$ to (A.1)–(A.4) in Lemma A.1, we have

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}[\|\widetilde{\nabla} F_{\omega_k}(x_k) - \widetilde{\nabla} f_{\omega_{T,k}}^{(T)}(x_k)\nabla f_{\omega_{T-1,k}}^{(T-1)}(y_k^{(T-1)})\cdots\nabla f_{\omega_{1,k}}^{(1)}(y_k^{(1)})\|^2]\\
&\le \mathcal{O}(\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2]) + \cdots + \mathcal{O}(\mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2]).
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
\mathbb{E}[Q] \le \frac{1}{2}\mathbb{E}[\|\nabla F(x_k)\|^2] &+ \mathcal{O}(\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2])\\
&+ \mathcal{O}(\mathbb{E}[\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|^2])\\
&+ \cdots + \mathcal{O}(\mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2]).
\end{aligned}$$

Taking expectations on both sides of (B.4) and substituting $\mathbb{E}[Q]$ by its upper bound derived above, we have

$$\begin{aligned}\frac{\alpha_k}{2}\|\nabla F(x_k)\|^2 \le \mathbb{E}[F(x_k)] - \mathbb{E}[F(x_{k+1})] &+ \mathcal{O}(\alpha_k \mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2]) \\ &+ \mathcal{O}(\alpha_k \mathbb{E}[\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|^2]) \\ &+ \cdots + \mathcal{O}(\alpha_k \mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2]) + \mathcal{O}(\alpha_k^2).\end{aligned}$$

This implies that

(B.5)
$$\begin{aligned}\mathbb{E}[\|\nabla F(x_k)\|^2] \le 2\alpha_k^{-1}\mathbb{E}[F(x_k)] - 2\alpha_k^{-1}\mathbb{E}[F(x_{k+1})] &+ \mathcal{O}(\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2]) \\ &+ \mathcal{O}(\mathbb{E}[\|y_k^{(T-2)} - f^{(T-1)}(y_k^{(T-1)})\|^2]) \\ &+ \cdots + \mathcal{O}(\mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2]) + \mathcal{O}(\alpha_k),\end{aligned}$$

which completes the proof. ∎

B.4. Proof of Theorem 3.2. First, we consider the case when $f^{(T)}$ does not have Lipschitz continuous gradient. Since Assumptions 2.1 and 2.2 hold, we apply Lemma 3.11 and sum (B.5) from $k = 1$ to n. Then we obtain

(B.6)
$$\begin{aligned}&\frac{\sum_{k=1}^n \mathbb{E}(\|\nabla(x_k)\|^2)}{n} \\ &\le 2n^{-1}F(x_0) + n^{-1}\sum_{k=1}^n ak^{a-1}\mathbb{E}[F(x_k)] + n^{-1}\sum_{k=1}^n \mathcal{O}(\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2]) \\ &\quad + \cdots + n^{-1}\sum_{k=1}^n \mathcal{O}(\mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2]) + n^{-1}\sum_{k=1}^n \mathcal{O}(n^{-a}),\end{aligned}$$

where the second inequality holds by the fact that $(k+1)^a \le k^a + ak^{a-1}$ since $h(t) = t^a$ is a concave function for $0 < a < 1$.

Meanwhile, by Lemmas 3.9 and 3.10, with the choice of $a = \frac{4+T}{8+T}$ and $b_j = \frac{3+j}{8+T}$ for $j = T-1, T-2, \ldots, 1$, we have $\mathbb{E}[\|y_{k+1}^{(T-1)} - f^{(T)}(x_k)\|^2] \le \mathcal{O}(k^{-4/(8+T)})$ and $\mathbb{E}[\|y_{k+1}^{(j)} - f^{(j+1)}(y_{k+1}^{(j+1)})\|^2] \le \mathcal{O}(k^{-4/(8+T)})$ for $j = T-2, \ldots, 1$. Plugging them into (B.6), we have

(B.7)
$$\begin{aligned}\frac{\sum_{k=1}^n \mathbb{E}(\|\nabla F(x_k)\|^2)}{n} &\le \mathcal{O}\Bigg(n^{a-1} + n^{2(b_{T-1}-a)}\mathbb{I}^{\log n}_{2(a-b_{T-1})=1} + n^{-b_{T-1}} \\ &\qquad + \sum_{j=1}^{T-2}[n^{4(b_j - b_{j+1})}\mathbb{I}^{\log n}_{4(b_{j+1}-b_j)=1} + n^{-b_j}] + n^{-a}\Bigg) \\ &\le \mathcal{O}(n^{-4/(8+T)}),\end{aligned}$$

which completes the proof.

Next, when $f^{(T)}$ has Lipschitz continuous gradient, the first inner level is also updated by the accelerating rule. By similar analysis to that in Lemma 3.10, we have

$$\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2] \le \mathcal{O}(k^{-4a+4b_{T-1}}) + \mathcal{O}(k^{-b_{T-1}}).$$

Plugging this convergent rate into (B.7), and using Lemma 3.10, we obtain

$$\begin{aligned}\frac{\sum_{k=1}^{n}\mathbb{E}(\|\nabla F(x_k)\|^2)}{n} &\le \mathcal{O}\Bigg(n^{a-1}+n^{-4a+4b_{T-1}}\mathbb{I}^{\log n}_{4(a-b_{T-1})=1}+n^{-b_{T-1}}\\ &\qquad +\sum_{j=1}^{T-2}[n^{4(b_j-b_{j+1})}\mathbb{I}^{\log n}_{4(b_{j+1}-b_j)}+n^{-b_j}]+n^{-a}\Bigg)\\ &\le \mathcal{O}(n^{-4/(7+T)}),\end{aligned}$$

by choosing $a=\frac{3+T}{7+T}$ and $b_j=\frac{3+j}{7+T}$ for $j=T-1,T-2,\dots,1$, which completes the proof. ∎

Appendix C. Proof of Theorem 3.3. Define $F^*=\min_{x\in\mathcal{X}}F(x)$, noting that $F^*=F(\Pi_{\mathcal{X}^*}(x))$ for all $x\in\mathcal{X}$. When $\mathcal{X}=\mathbb{R}^{d_T}$, based on the definition of x_k, we have

$$x_{k+1}-x_k=-\alpha_k\widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f_{\omega_{1,k}}(y_k^{(1)}).$$

Then, for the term $||x_{k+1}-\Pi_{\mathcal{X}^*}(x_{k+1})||^2$, we have

$$\begin{aligned}&||x_{k+1}-\Pi_{\mathcal{X}^*}(x_{k+1})||^2\\ &\le ||x_{k+1}-x_k+x_k-\Pi_{\mathcal{X}^*}(x_k)||^2\\ &\le ||x_k-\Pi_{\mathcal{X}^*}(x_k)||^2-||x_{k+1}-x_k||^2+2\alpha_k\langle\nabla F(x_k),\Pi_{\mathcal{X}^*}(x_k)-x_{k+1}\rangle\\ &\quad+2\alpha_k\langle\widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})\\ &\qquad\qquad-\nabla F(x_k),\Pi_{\mathcal{X}^*}(x_k)-x_{k+1}\rangle,\end{aligned}\tag{C.1}$$

Let $T_1=\langle\nabla F(x_k),\Pi_{\mathcal{X}^*}(x_k)-x_{k+1}\rangle$ and

$$T_2=\langle\widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})-\nabla F(x_k),\Pi_{\mathcal{X}^*}(x_k)-x_{k+1}\rangle.$$

For the term T_1, we have

$$\begin{aligned}T_1&=\langle\nabla F(x_k),x_k-x_{k+1}\rangle+\langle\nabla F(x_k),\Pi_{\mathcal{X}^*}(x_k)-x_k\rangle\\ &\le F(x_k)-F(x_{k+1})+\frac{L_F}{2}\|x_{k+1}-x_k\|^2+F(\Pi_{\mathcal{X}^*}(x_k))-F(x_k)\\ &\le F^*-F(x_{k+1})+\mathcal{O}(\alpha_k^2).\end{aligned}$$

For the term T_2, we have

$$\begin{aligned}\mathbb{E}[T_2]\le&\underbrace{\mathbb{E}[\langle\widetilde{\nabla}F_{\omega_k}(x_k)-\widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)}),x_k-\Pi_{\mathcal{X}^*}(x_k)\rangle]}_{T_{2,1}}\\ &+\frac{\alpha_k}{2}\underbrace{\mathbb{E}[\|\widetilde{\nabla}F_{\omega_k}(x_k)-\widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})\|^2]}_{T_{2,2}}\\ &+\frac{1}{2\alpha_k}\|x_k-x_{k+1}\|^2,\end{aligned}$$

where the inequality comes from the fact that $\langle a, b\rangle \le \frac{1}{2\alpha_k}\|a\|^2 + \frac{\alpha_k}{2}\|b\|^2$. For $T_{2,1}$, we have

$$\begin{aligned} T_{2,1} &\le \frac{\alpha_k}{2\phi_k}\mathbb{E}[\|\widetilde{\nabla}F_{\omega_k}(x_k) - \widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})\|^2] \\ &\quad + \mathbb{E}\left[\frac{\phi_k}{2\alpha_k}\|x_k - \Pi_{\mathcal{X}^*}(x_k)\|^2\right] \\ &\le \mathcal{O}\left(\frac{\alpha_k}{\phi_k}\right)[\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2] + \cdots + \mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2]] \\ &\quad + \frac{\phi_k}{2\alpha_k}\mathbb{E}[\|x_k - \Pi_{\mathcal{X}^*}(x_k)\|^2], \end{aligned}$$

where ϕ_k is a scalar and will be specified later. By the fact that $\|a-b\|^2 \le 2a^2 + 2b^2$ and by Assumptions 2.1(iii) and (iv), we have

$$\begin{aligned} T_{2,2} &\le 2\mathbb{E}[\|\widetilde{\nabla}F_{\omega_k}(x_k)\|^2] + 2\mathbb{E}[\|\widetilde{\nabla}f^{(T)}_{\omega_{T,k}}(x_k)\nabla f^{(T-1)}_{\omega_{T-1,k}}(y_k^{(T-1)})\cdots\nabla f^{(1)}_{\omega_{1,k}}(y_k^{(1)})\|^2] \\ &\le \mathcal{O}(1). \end{aligned}$$

Taking expectations on both sides of (C.1) and plugging in the upper bounds of T_1 and T_2 derived above, we obtain

$$\begin{aligned} &2\alpha_k(\mathbb{E}[F(x_{k+1})] - F^*) + \mathbb{E}[\|x_{k+1} - \Pi_{\mathcal{X}^*}(x_{k+1})\|^2] \\ &\le (1+\phi_k)\mathbb{E}[\|x_k - \Pi_{\mathcal{X}^*}(x_k)\|^2] + \mathcal{O}(\alpha_k^3) + \mathcal{O}(\alpha_k^2/\phi_k)[\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2] \\ &\quad + \cdots + \mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2]] + \mathcal{O}(\alpha_k^2). \end{aligned}$$

By the definition of optimally strong convexity in (2.3), we have

$$F(x_{k+1}) - F^* \ge \lambda\|x_{k+1} - \Pi_{\mathcal{X}^*}(x_{k+1})\|^2,$$

which further implies

$$\begin{aligned} &(1+2\lambda\alpha_k)\mathbb{E}[\|x_{k+1} - \Pi_{\mathcal{X}^*}(x_{k+1})\|^2] \\ &\le (1+\phi_k)\mathbb{E}[\|x_k - \Pi_{\mathcal{X}^*}(x_k)\|^2] \\ &\quad + \mathcal{O}(\alpha_k^2/\phi_k)(\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2] + \cdots + \mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2]) \\ &\quad + \mathcal{O}(\alpha_k^3) + \mathcal{O}(\alpha_k^2). \end{aligned}$$

By dividing $(1+2\lambda\alpha_k) > 0$ on both sides, it follows that

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - \Pi_{\mathcal{X}^*}(x_{k+1})\|^2] &\le \frac{1+\phi_k}{1+2\lambda\alpha_k}\mathbb{E}[\|x_k - \Pi_{\mathcal{X}^*}(x_k)\|^2] \\ &\quad + \mathcal{O}(\alpha_k^2/\phi_k)(\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2] \\ &\quad + \cdots + \mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2]) + \mathcal{O}(\alpha_k^3) + \mathcal{O}(\alpha_k^2). \end{aligned}$$

Choosing $\phi_k = \lambda\alpha_k - 2\lambda^2\alpha_k^2$ yields that

(C.2)

$$\begin{aligned} &\mathbb{E}[\|x_{k+1} - \Pi_{\mathcal{X}^*}(x_{k+1})\|^2] \\ &\le (1-\lambda\alpha_k)\mathbb{E}[\|x_k - \Pi_{\mathcal{X}^*}(x_k)\|^2] \\ &\quad + \mathcal{O}(\alpha_k^2) + \mathcal{O}\left(\frac{\alpha_k}{\lambda}\right)(\mathbb{E}[\|y_k^{(T-1)} - f^{(T)}(x_k)\|^2] + \cdots + \mathbb{E}[\|y_k^{(1)} - f^{(2)}(y_k^{(2)})\|^2]). \end{aligned}$$

When $f^{(T)}$ has Lipschitz continuous gradient, applying Lemmas 3.9 and 3.10 yields that

$$\begin{aligned}&\mathbb{E}[\|x_{k+1}-\Pi_{\mathcal{X}^*}(x_{k+1})\|^2]\\&\le(1-\lambda\alpha_k)\mathbb{E}[\|x_k-\Pi_{\mathcal{X}^*}(x_k)\|^2]+\mathcal{O}(k^{-2a})\\&\quad+\mathcal{O}\Big(\frac{\alpha_k}{\lambda}\Big)\Big(\mathcal{O}(k^{-2a+2b_{T-1}})+\mathcal{O}(k^{-b_{T-1}})+\sum_{j=1}^{T-2}\mathcal{O}(k^{-4b_j+4b_{j+1}})+\mathcal{O}(k^{-b_j})\Big).\end{aligned}$$

Applying Lemma B.1 to the previous inequality, we have

$$\begin{aligned}\mathbb{E}[\|x_k-\Pi_{\mathcal{X}^*}(x_k)\|^2]&\le\mathcal{O}(k^{-a})+\mathcal{O}(k^{-2a+2b_{T-1}})\\&\quad+\mathcal{O}(k^{-b_{T-1}})+\sum_{j=1}^{T-2}[\mathcal{O}(k^{-4b_j+4b_{j+1}})+\mathcal{O}(k^{-b_j})].\end{aligned}$$

Letting $a=1$ and $b_{T-1}=\frac{2+T}{4+T},\ b_{T-2}=\frac{1+T}{4+T},\ \ldots,\ b_1=\frac{4}{4+T}$, we have

$$\mathbb{E}[\|x_k-\Pi_{\mathcal{X}^*}(x_k)\|^2]\le\mathcal{O}(k^{-4/(4+T)}),$$

which provides the convergence rate result for the optimally strongly convex T-level accelerated SCGD.

Next, when $f^{(T)}$ has Lipschitz continuous gradient, the first inner function is also updated by the accelerated update rule. By a similar analysis to that in Lemma 3.10, we have

$$\mathbb{E}[\|y_k^{(T-1)}-f^{(T)}(x_k)\|^2]\le\mathcal{O}(k^{-4a+4b_{T-1}})+\mathcal{O}(k^{-b_{T-1}}).$$

Plugging this convergence rate into (C.2), we have

$$\begin{aligned}\mathbb{E}[\|x_k-\Pi_{\mathcal{X}^*}(x_k)\|^2]&\le\mathcal{O}(k^{-a})+\mathcal{O}(k^{-4a+4b_{T-1}})+\mathcal{O}(k^{-b_{T-1}})\\&\quad+\sum_{j=1}^{T-2}\big[\mathcal{O}(k^{-4b_j+4b_{j+1}})+\mathcal{O}(k^{-b_j})\big]\\&\le\mathcal{O}(k^{-4/(3+T)})\end{aligned}$$

by choosing $a=1$, $b_{T-1}=\frac{2+T}{3+T},\ b_{T-2}=\frac{1+T}{3+T},\ \ldots,\ b_1=\frac{4}{3+T}$, which completes the proof. ∎

REFERENCES

[1] S. Ahmed, U. Çakmak, and A. Shapiro, *Coherent risk measures in inventory problems*, Eur. J. Oper. Res., 182 (2007), pp. 226–238.

[2] F. Bach and E. Moulines, *Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$*, in Adv. Neural Inf. Process. Syst. 26, Curran Associates, Red Hook, NY, 2013, pp. 773–781.

[3] S. Bruno, S. Ahmed, A. Shapiro, and A. Street, *Risk neutral and risk averse approaches to multistage renewable investment planning under uncertainty*, Eur. J. Oper. Res., 250 (2016), pp. 979–989.

[4] S. Cole, X. Giné, and J. Vickery, *How does risk management influence production decisions? Evidence from a field experiment*, Rev. Financ. Stud, 30 (2017), pp. 1935–1970; available at https://doi.org/10.1093/rfs/hhw080.

[5] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, in Adv. Neural Inf. Process. Syst. 27, Curran Associates, Red Hook, NY, 2014, pp. 1646–1654.

[6] D. Dentcheva, S. Penev, and A. Ruszczyński, *Statistical estimation of composite risk functionals and risk optimization problems*, Ann. Inst. Statist. Math., 69 (2017), pp. 737–760.
[7] J. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, J. Mach. Learn. Res., 12 (2011), pp. 2121–2159.
[8] Y. Ermoliev, *Methods of Stochastic Programming*, Monogr. Optim. Oper. Res., Nauka, Moscow, 1976.
[9] R. Ge, F. Huang, C. Jin, and Y. Yuan, *Escaping from saddle points: Online stochastic gradient for tensor decomposition.*, in Proceedings of the International Conference on Learning Theory, Proc. Mach. Learn. Res. 40, 2015, pp. 797–842; available at http://proceedings.mlr.press/v40/.
[10] S. Ghadimi and G. Lan, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, SIAM J. Optim., 23 (2013), pp. 2341–2368.
[11] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, in Adv. Neural Inf. Process. Syst. 26, Curran Associates, Red Hook, NY, 2013, pp. 315–323.
[12] J. Konečnỳ and P. Richtárik, *Semi-stochastic gradient descent methods*, Front. Appl. Math. Statist., 3 (2017), p. 9.
[13] G. Lan, A. Nemirovski, and A. Shapiro, *Validation analysis of mirror descent stochastic approximation method*, Math. Program., 134 (2012), pp. 425–458.
[14] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature, 521 (2015), pp. 436–444.
[15] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, *Gradient descent only converges to minimizers*, in Proceedings of the International Conference on Learning Theory, Proc. Mach. Learn. Res. 49, 2016, pp. 1246–1257; available at http://proceedings.mlr.press/v49/.
[16] C. J. Li, M. Wang, H. Liu, and T. Zhang, *Near-optimal Stochastic Approximation for Online Principal Component Estimation*, preprint, https://arxiv.org/abs/1603.05305, 2016.
[17] X. Lian, M. Wang, and J. Liu, *Finite-sum Composition Optimization via Variance Reduced Gradient Descent*, preprint, https://arxiv.org/abs/1610.04674, 2016.
[18] C. Lim and B. Yu, *Estimation stability with cross-validation (ESCV)*, J. Comput. Graph. Statist., 25 (2016), pp. 464–492.
[19] D. Needell, R. Ward, and N. Srebro, *Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm*, in Adv. Neural Inf. Process. Syst. 27, Curran Associates, Red Hook, NY, 2014, pp. 1017–1025.
[20] A. Rakhlin, O. Shamir, and K. Sridharan, *Making gradient descent optimal for strongly convex stochastic optimization*, in Proceedings of the 29th International Conference on Machine Learning, Omnipress, Madison, WI, 2012, pp. 1571–1578.
[21] B. Recht and C. Ré, *Parallel stochastic gradient algorithms for large-scale matrix completion*, Math. Program. Comput., 5 (2013), pp. 201–226.
[22] A. Ruszczyński and A. Shapiro, *Optimization of convex risk functions*, Math. Oper. Res., 31 (2006), pp. 433–452.
[23] M. Schmidt, N. Le Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient*, Math. Program., 162 (2017), pp. 83–112.
[24] O. Shamir and T. Zhang, *Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes*, in Proceedings of the International Conference on Machine Learning 2013, Proc. Mach. Learn. Res. 28, 2013, pp. 71–79; available at http://proceedings.mlr.press/v28/.
[25] W. W. Sun, X. Qiao, and G. Cheng, *Stabilized nearest neighbor classifier and its statistical properties*, J. Amer. Statist. Assoc., 111 (2016), pp. 1254–1265.
[26] M. Wang and D. P. Bertsekas, *Stochastic first-order methods with random constraint projection*, SIAM J. Optim., 26 (2016), pp. 681–717.
[27] M. Wang, E. X. Fang, and H. Liu, *Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions*, Math. Program., 161 (2017), pp. 419–449.
[28] M. Wang and J. Liu, *A stochastic compositional gradient method using Markov samples*, in Proceedings of the 2016 Winter Simulation Conference, IEEE Press, Piscataway, NJ, 2016, pp. 702–713.
[29] M. Wang, J. Liu, and E. X. Fang, *Accelerating stochastic composition optimization*, in Adv. Neural Inf. Process. Syst. 29, Curran Associates, Red Hook, NY, 2016, pp. 1714–1722.
[30] S. Wiesler, A. Richard, R. Schluter, and H. Ney, *Mean-normalized stochastic gradient for large-scale deep learning*, in Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE Press, Piscataway, NJ, 2014, pp. 180–184.
[31] L. Xiao and T. Zhang, *A proximal stochastic gradient method with progressive variance reduction*, SIAM J. Optim., 24 (2014), pp. 2057–2075.