Efficient Smooth Non-Convex Stochastic Compositional Optimization via Stochastic Recursive Gradient Descent

Wenqing Hu
Missouri University of Science and Technology
huwen@mst.edu

Chris Junchi Li
Tencent AI Lab
junchi.li.duke@gmail.com

Xiangru Lian
University of Rochester
admin@mail.xrlian.com

Ji Liu
University of Rochester & Kwai Inc.
ji.liu.uwisc@gmail.com

Huizhuo Yuan
Peking University
hzyuan@pku.edu.cn

Abstract

Stochastic compositional optimization arises in many important machine learning tasks such as reinforcement learning and portfolio management. The objective function is the composition of two expectations of stochastic functions, and is more challenging to optimize than vanilla stochastic optimization problems. In this paper, we investigate the stochastic compositional optimization in the general smooth non-convex setting. We employ a recently developed idea of Stochastic Recursive Gradient Descent to design a novel algorithm named SARAH-Compositional, and prove a sharp Incremental First-order Oracle (IFO) complexity upper bound for stochastic compositional optimization: \(O\left((n + m)^{1/2}\varepsilon^{-2}\right)\) in the finite-sum case and \(O(\varepsilon^{-3})\) in the online case. Such a complexity is known to be the best one among IFO complexity results for non-convex stochastic compositional optimization. Numerical experiments validate the superior performance of our algorithm and theory.

1 Introduction

We consider the general smooth, non-convex compositional optimization problem of minimizing the composition of two expectations of stochastic functions:

\[
\min_{x \in \mathbb{R}^d} \{ \Phi(x) \equiv (f \circ g)(x) \},
\]

where the outer and inner functions \( f : \mathbb{R}^l \to \mathbb{R}, g : \mathbb{R}^d \to \mathbb{R}^l \) are defined as \( f(y) := E_v[f_v(y)] \), \( g(x) := E_w[g_w(y)] \) where \( v \) and \( w \) are random variables, and each component \( f_v, g_w \) are smooth but not necessarily convex. Compositional optimization can be used to formulate many important machine learning problems, e.g. reinforcement learning (Sutton and Barto, 1998), risk management (Dentcheva et al., 2017), multi-stage stochastic programming (Shapiro et al., 2009) and deep neural net (Yang et al., 2018), etc. We list two specific application instances that can be written in the stochastic compositional form of (1):

- **Risk management problem**, which is formulated as

\[
\min_{x \in \mathbb{R}^N} - \frac{1}{T} \sum_{t=1}^{T} \langle r_t, x \rangle + \frac{1}{T} \sum_{t=1}^{T} \left( \langle r_t, x \rangle - \frac{1}{T} \sum_{s=1}^{T} \langle r_s, x \rangle \right)^2,
\]

\[\text{2}\]

*Partial work was performed when the author was an intern at Tencent AI Lab.

We use the same notation for the online case, in which case either $n$ or $m$ can be infinite.

A fundamental theoretical question for stochastic compositional optimization is the Incremental First-order Oracle (IFO) (the number of individual gradient and function evaluations; see Definition 1) below) complexity bounds for stochastic compositional optimization. Our new SARAH-Compositional algorithm is developed by integrating the iteration of Stochastic Recursive Gradient
Descent [Nguyen et al., 2017], shortened as SARAH, with the stochastic compositional optimization structure [Wang et al., 2017a]. The motivation of this approach is that SARAH with specific choice of stepsizes is known to be optimal in stochastic optimization and regarded as a cutting-edge variance reduction technique, with significantly reduced oracle access complexities than earlier variance reduction method [Fang et al., 2018]. We prove that SARAH-Compositional can reach an IFO computational complexity of $O(\min((n + m)^{1/2}\varepsilon^{-2}, \varepsilon^{-3}))$, improving the best known result of $O(\min((n + m)^{2/3}\varepsilon^{-2}, \varepsilon^{-3.0}))$ in non-convex compositional optimization. See Table 1 for detailed comparison.

**Related Works** Gradient descent and stochastic gradient descent are basic algorithms in optimization, and their performance are surpassed by numerous recently proposed first-order methods. Nesterov’s acceleration (Nesterov, 2004) is one such example. Ghadimi and Lan (2016) proposed Nesterov-type method and Li and Lin (2015) proposed accelerated proximal gradient for non-convex functions, in which case Nesterov’s accelerated gradient method achieves the same convergence rate with SGD, while obtaining accelerated rate than GD when the objective function is non-stochastic. Another variant of GD named normalized gradient descent (Nesterov, 2004) stabilizes the training process by fixing the step length, which enjoys in the non-convex setting the same IFO complexity as GD.

In recent years, variance reduction methods have received intensive attentions in both convex and non-convex optimization. When the objective is strongly convex, variance reduction techniques can improve (from sub-linear convergence rates by GD) to linear convergence rates. Variance-reduced gradient methods include SVRG (Xiao and Zhang, 2014), SAG/SAGA (Schmidt et al., 2017; Defazio et al., 2014), SDCA (Shalev-Shwartz and Zhang, 2013, 2014), etc. For non-convex objectives, the method of stochastic variance reduced gradient (SVRG) (Xiao and Zhang, 2013) applies in a straightforward fashion (Allen-Zhu and Hazan, 2016; Reddi et al., 2016), as well as an online variant of SVRG (called SCSG) later proposed by (Lei et al., 2017). The IFO complexities above can be further reduced by a recently proposed SPIDER algorithm (Fang et al., 2018), which hybrids the stochastic recursive gradient descent framework (SARAH) (Nguyen et al., 2017) with normalized gradient descent (NGD). It adopt a small stepsize that is proportional to $\varepsilon$ and hence variance can be effectively controlled. For finding $\varepsilon$-accurate solution purposes, Wang et al. (2018) small stepsize restriction in Fang et al. (2018) while achieving the same complexity. The convergence property of both SARAH and SPIDER methods in the non-convex case outperforms that of SVRG, and is state-of-the-art.

It turns out in compositional optimization problem, gradient-descent based methods for optimizing a single objective function can either be non-applicable, or it brings at least $O(m)$ queries to calculate the inner function $g$. To remedy for this problem, Wang et al. (2017a,b) proposed SCGD method to calculate the inner finite-sum more efficiently and achieve a polynomial rate that is independent of $m$. Later on, Lian et al. (2017); Liu et al. (2017); Huo et al. (2018) and Lin et al. (2018) merged SVRG method into the compositional optimization framework to do variance reduction on all three steps of the estimation. Among these, Lin et al. (2018), Huo et al. (2018) and Wang et al. (2017b) consider the case of composition objective with an extra non-smooth proximal term. Due to the lack of space, we satisfy ourselves with the smooth case and leave the non-smooth analysis to the future.

**Contributions** This work makes two contributions as follows. First, we propose a new algorithm for stochastic compositional optimization called SARAH-Compositional, which operates SARAH/SPIDER-type variance reduction to approximate and estimate relevant quantities. Such algorithm is designed to unify both online and finite-sum cases. Second, we conduct theoretical analysis for both online and finite-sum cases. In the finite-sum case, we obtain a complexity of $(n + m)^{1/2}\varepsilon^{-2}$ which improves over the prior best known complexity $(n + m)^{2/3}\varepsilon^{-2}$ achieved by Huo et al. (2018). In the online case we obtain a complexity of $\varepsilon^{-3}$ which improves the prior best known complexity $\varepsilon^{-3.6}$ obtained in Liu et al. (2017).

**Notational Conventions** Throughout the paper, we treat the parameters $L_g, L_f, L_q, M_g, M_f, \Delta$ and $\sigma$ as global constants. Let $\| \bullet \|_2$ denote the Euclidean norm of a vector or the operator norm of a matrix induced by Euclidean norm, and let $\| \bullet \|_F$ denotes the Frobenius norm. For fixed $T \geq t \geq 0$ let $x_t : T$ denote the sequence $\{x_t, \ldots, x_T\}$. Let $E_{\theta}[\bullet]$ denote the conditional expectation $E_{\theta}[\bullet|x_0, x_1, \ldots, x_t]$. Let $|n| = \{1, \ldots, n\}$ and $S$ denote the cardinality of a multi-set $S \subseteq |n|$ of samples (a generic set that permits repeated instances). The averaged sub-sampled stochastic estimator is denoted as $A_{\mathcal{S}} = (1/S) \sum_{i \in S} A_i$ where the summation counts repeated instances. We denote $p_n = O(q_n)$ if
there exist some constants $0 < c < C < \infty$ such that $c q_n \leq p_n \leq C q_n$ as $n$ becomes large. Other notations are explained at their first appearances.

**Organization** The rest of our paper is organized as follows. §2 formally poses our algorithm and assumptions. §3 presents the convergence rate theorem and §4 presents numerical experiments that apply our algorithm to the task of portfolio management. We conclude our paper in §5. Proofs of convergence results for finite-sum and online cases and auxiliary lemmas are deferred to the supplementary material.

## 2 SARAH for Stochastic Compositional Optimization

Recall our goal is to solve the compositional optimization problem (1), i.e. to minimize $\Phi(x) = f(g(x))$ where

$$f(y) := \frac{1}{n} \sum_{i=1}^{n} f_i(y), \quad g(x) := \frac{1}{m} \sum_{j=1}^{m} g_j(x).$$

Here for each $j \in [m]$ and $i \in [n]$ the functions $g_j : \mathbb{R}^d \to \mathbb{R}^l$ and $f_i : \mathbb{R}^l \to \mathbb{R}$. We can formally take a derivative to the function $\Phi(x)$ and obtain (via the chain rule) the gradient descent iteration

$$x_{t+1} = x_t - \eta \partial g(x_t) \nabla f(g(x_t)),$$

where the $\partial$ operator computes the Jacobian matrix of the smooth mapping, and the gradient operator $\nabla$ is only taken with respect to the first-level variable. As discussed in §1 it can be either impossible (online case) or time-consuming (finite-sum case) to estimate the terms $\partial g(x_t) = \frac{1}{m} \sum_{j=1}^{m} \partial g_j(x_t)$ and $g(x_t) = \frac{1}{m} \sum_{j=1}^{m} g_j(x_t)$ in the above iteration scheme. In this paper, we design a novel algorithm (SARAH-Compositional) based on Stochastic Compositional Variance Reduced Gradient method (see Lin et al. [2018]) yet hybriding with the stochastic recursive gradient method Nguyen et al. (2017).

We introduce some definitions and assumptions. First, we assume the algorithm has accesses to the Incremental First-order Oracle (IFO) in our black-box environment (Lin et al., 2018); also see (Agarwal and Bottou, 2015) for vanilla optimization case:

**Definition 1 (IFO).** (Lin et al., 2018) The Incremental First-order Oracle (IFO) returns, when some $x \in \mathbb{R}^d$ and $j \in [m]$ are inputted, the vector-matrix pair $[g_j(x), \partial g_j(x)]$ or when some $y \in \mathbb{R}^l$ and $i \in [n]$ are inputted, the scalar-vector pair $[f_i(y), \nabla f_i(y)]$.

Second, our goal in this work is to find an $\varepsilon$-accurate solution, defined as

**Definition 2 ($\varepsilon$-accurate solution).** We call $x \in \mathbb{R}^d$ an $\varepsilon$-accurate solution to problem (1), if

$$\|\nabla \Phi(x)\| \leq \varepsilon.$$  \hspace{1cm} (8)

It is worth remarking here that the inequality (8) can be modified to $\|\nabla \Phi(x)\| \leq C \varepsilon$ for some global constant $C > 0$ without hurting the complexities.

Let us first make some assumptions regarding to each component of the (compositional) objective function. Analogous to Assumption 1(i) of Fang et al. (2018), we make the following finite gap assumption:

**Assumption 1 (Finite-gap).** We assume that the algorithm is initialized at $x_0 \in \mathbb{R}^d$ with

$$\Delta := \Phi(x_0) - \Phi^* < \infty,$$

where $\Phi^*$ denotes the global minimum value of $\Phi(x)$.

We make the following smoothness and boundedness assumptions, which are standard in recent compositional optimization literatures (e.g. Lian et al. (2017), Huo et al. (2018), Lin et al. (2018)).
whose applicability can be verified via a simple application of the chain rule. We integrate both
we adopt the following typical choice of

\[ L \text{(Boundedness)} \]

Assumption 3 (Smoothness). There exist Lipschitz constants \( L_g, L_f, L_\Phi > 0 \) such that for \( i \in [n] \),
\( j \in [m] \) we have
\[
\| \partial g_j(x) - \partial g_j(x') \|_F \leq L_g \| x - x' \| \quad \text{for } x, x' \in \mathbb{R}^d, \\
\| \nabla f_i(y) - \nabla f_i(y') \| \leq L_f \| y - y' \| \quad \text{for } y, y' \in \mathbb{R}^l, \quad (10)
\]
\[
\| \partial g_j(x) \| \nabla f_i(g(x)) - [\partial g_j(x')] \| \nabla f_i(g(x')) \| \leq L_\Phi \| x - x' \| \quad \text{for } x, x' \in \mathbb{R}^d.
\]

Here for the purpose of using stochastic recursive estimation of \( \partial g(x) \), we slightly strengthen the
smoothness assumption by adopting the Frobenius norm in left hand of the first line of (10).

Assumption 3 (Boundedness). There exist boundedness constants \( M_g, M_f > 0 \) such that for \( i \in [n] \),
\( j \in [m] \) we have
\[
\| \partial g_j(x) \| \leq M_g \quad \text{for } x \in \mathbb{R}^d, \\
\| \nabla f_i(y) \| \leq M_f \quad \text{for } y \in \mathbb{R}^l. \quad (11)
\]

Notice that applying mean-value theorem for vector-valued functions to (11) gives another Lipschitz condition
\[
\| g_j(x) - g_j(x') \| \leq M_g \| x - x' \| \quad \text{for } x, x' \in \mathbb{R}^d, \quad (12)
\]
and analogously for \( f_i(y) \). It turns out that under the above two assumptions, a choice of \( L_\Phi \) in (10)
can be expressed as a polynomial of \( L_f, L_g, M_f, M_g \). For clarity purposes in the rest of this paper,
we adopt the following typical choice of \( L_\Phi \)
\[
L_\Phi \equiv M_f L_g + M_g^2 L_f, \quad (13)
\]
whose applicability can be verified via a simple application of the chain rule. We integrate both
finite-sum and online cases into one algorithm SARAH-Compositional and write it in Algorithm 1.
3 Convergence Rate Analysis

In this section, we aim to justify that our proposed SARAH-Compositional algorithm provides IFO complexities of $O((n + m)^{1/2} \varepsilon^{-2})$ in the finite-sum case and $O(\varepsilon^{-3})$ in the online case, which supersedes the concurrent and comparative algorithms (see more in Table 1).

Let us first analyze the convergence in the finite-sum case. In this case we have $S^1 = [m]$, $S^2 = [m]$, $S^3 = [n]$. Involved analysis leads us to conclude

**Theorem 1** (Finite-sum case). Suppose Assumptions 1, 2 and 3 hold, let $S^1 = S^2 = [m]$, $S^3 = [n]$, $q = (2m + n)/3$, and set the stepsize

$$
\eta = \frac{1}{\sqrt{6(2m + n) \left( M_g^4 L_f^2 + M_f^2 L_g^2 \right)}}.
$$

Then for the finite-sum case, SARAH-Compositional Algorithm 1 outputs an $\bar{x}$ satisfying $E\|\nabla \Phi(\bar{x})\|^2 \leq \varepsilon^2$ in $2m + n + \sqrt{6(2m + n) \left( M_g^4 L_f^2 + M_f^2 L_g^2 \right)} \cdot \sqrt{\frac{216}{\varepsilon^2} [\Phi(x_0) - \Phi^*]}.

Like in Fang et al. (2018), the IFO complexity to achieve an $\varepsilon$-accurate solution is upper bounded by $O((m + n)^{1/2} \varepsilon^{-2})$. We see that in the online case, the IFO complexity to achieve an $\varepsilon$-accurate solution is upper bounded by $O(\varepsilon^{-3})$.

Theorem 1 allows us to achieve an $\varepsilon$-accurate solution, and a simple application of Markov’s inequality allows us to derive high-probability results for achieving $\varepsilon$-accurate solutions. Compared with Fang et al. (2018), one observes that Theorem 1 indicates an IFO complexity upper bound of $O((m + n + (m + n)^{1/2} \varepsilon^{-2})$, sharing a similar form with that of SARAH/SPIDER for non-convex stochastic optimization when $m + n$ is regarded as the number of individual functions. SPIDER-SFO (as a SARAH variant) is optimal in the finite-sum case, in the sense that it matches the theoretical lower bound, which makes it tempting to claim that our proposed SARAH-Compositional as its extension is also optimal. We emphasize that the set of assumptions for compositional optimization is different from vanilla optimization, and claiming optimality of the IFO complexity requires a corresponding lower bound result, left as a future direction to explore.

Let us then analyze the convergence in the online case, where we sample minibatches $S^1, S^2, S^3$ of relevant quantities instead of the ground truth once every $q$ iterates. To characterize the estimation error, we put in one additional finite variance assumption:

**Assumption 4** (Finite Variance). We assume that there exists $H_1, H_2$ and $H_3$ as the upper bounds on the variance of the functions $f(y)$, $\partial g(x)$, and $g(x)$, respectively, such that

$$
\begin{align*}
E\|g_i(x) - g(x)\|^2 & \leq H_1 \quad \text{for } x \in \mathbb{R}^d, \\
E\|\partial g_i(x) - \partial g(x)\|^2 & \leq H_2 \quad \text{for } x \in \mathbb{R}^d, \\
E\|\nabla f_i(y) - \nabla f(y)\|^2 & \leq H_3 \quad \text{for } y \in \mathbb{R}^l.
\end{align*}
$$

From Assumptions 2 and 3 we can easily verify, via triangle inequality and convexity of norm, that $H_2$ as $4M_f^2$ and $H_3$ can be chosen as $4M_f^2$. The choice of $H_1$ cannot be represented as a function of boundedness and smoothness constants.

We know that $\|G_{(t/q)}\| = \|\partial g_S(x_{(t/q)})\| \leq M_g$. From (17) we derive

$$
E\|\left(G_{(t/q)}\right)^T \nabla f_i(g_{(t/q)}) - \left(G_{(t/q)}\right)^T \nabla f(g_{(t/q)})\|^2 \leq H_3 M_g^2.
$$

\footnote{Here and in below, the smoothness and boundedness parameters and $\Phi(x_0) - \Phi^*$ are treated as constants}
Then for the online case, SARAH-Compositional Algorithm 1 outputs an \( \bar{x} \) satisfying
\[
\mathbb{E} \| \nabla \Phi(\bar{x}) \|^2 \leq 2\varepsilon^2 \text{ in } 2\varepsilon^2 \text{ iterates.}
\]
The IFO complexity to achieve an \( \varepsilon \)-accurate solution is bounded by
\[
\frac{D_0}{\varepsilon^2} + \sqrt{\frac{D_0}{\varepsilon^2}} \cdot \sqrt{9 \left( M_g^2 L_f^2 + M_f^2 L_g^2 \right)} \cdot \frac{\sqrt{216} | \Phi(x_0) - \Phi^* |}{\varepsilon^3}.
\]

Due to space limits, the detailed proofs of Theorems 1 and 2 are deferred to §A in the supplementary material.

4 Experiments

In this section, we study performance of our algorithm to (risk-adverse) portfolio management minimization problem and conduct numerical experiments to support our theory. We follow the setups in [Huo et al. 2018; Liu et al. 2017] and compare with existing algorithms for compositional optimization. Readers are referred to [Wang et al. 2017a; b] for more tasks our algorithm can be potentially applied for.

Recall that in Section 1, we formulate our portfolio management minimization problem as a mean-variance optimization problem (2), which can be formulated as a compositional optimization problem (1). For convenience we repeat the display here:
\[
\min_{x \in \mathbb{R}^N} -\frac{1}{T} \sum_{t=1}^{T} \langle r_t, x \rangle + \frac{1}{T} \sum_{t=1}^{T} \left( \langle r_t, x \rangle - \frac{1}{T} \sum_{s=1}^{T} \langle r_s, x \rangle \right)^2,
\]
where \( x = \{ x_1, x_2, \ldots, x_N \} \in \mathbb{R}^N \) denotes the quantities invested at every assets \( i = 1, \ldots, N \). For illustration purpose we set \( T = 2000 \) and \( N = 200 \). Similar to the setup of [Huo et al. 2018], we sample \( r_t \) from the Gaussian distribution and take the absolute value. Each row \( r_t \) in the \( T \times N \) matrix \( [r_1, r_2, \ldots, r_T]^T \) is (independently) generated from a zero-mean \( N \)-dimensional Gaussian distribution with covariance matrix \( \Sigma \in \mathbb{R}^{N \times N} \). We prescribe the conditional number \( \kappa \) of the population covariance \( \Sigma \) as one of our parameters, and tested the cases where \( \kappa(\Sigma) = 4 \) and \( \kappa(\Sigma) = 20 \). When applying SARAH-Compositional we adopt the finite-sum case i.e. \( S_1 = S_2 = S_3 = T \). Our search of stepsize is among \( \{ 5 \times 10^{-3}, 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 2 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}, 2 \times 10^{-1} \} \), and we plot the learning curve of each algorithms corresponding to their individual best stepsize found. The results are shown in Figure 1.

We demonstrate the comparison between our algorithm SARAH-Compositional, basic SCGD (Wang et al. 2017b) for compositional optimization, and VRSC-PG (Huo et al. 2018) which serves as a baseline for variance reduced stochastic compositional optimization methods. We plot the objective function value gap and gradient norm against IFO complexity (measured by counts of gradient

When \( S_3 = \frac{3H_2M_g^2}{\varepsilon^2}, S_2 = \frac{3H_2M_f^2}{\varepsilon^2}, S_1 = \frac{3H_1M_g^2L_f^2}{\varepsilon^2} \), under Assumption 4 \( \omega_1, \omega_2 \) and \( \omega_3 \) can be chosen as \( \frac{e}{3} \cdot \frac{e^2}{3M_f^2L_f^2} \), respectively.

We conclude the following theorem for the online case:

**Theorem 2** (Online case). Suppose Assumptions 1, 2, 3 and 4 in §2 hold, let \( S_1 = \frac{3H_1M_g^2L_f^2}{\varepsilon^2}, S_2 = \frac{3H_2M_f^2}{\varepsilon^2}, S_3 = \frac{3H_3M_f^2}{\varepsilon^2} \), let \( q = \frac{D_0}{3\varepsilon^2} \) where we denote the noise-relevant parameter
\[
D_0 := 3 \left( H_1M_g^2L_f^2 + H_2M_f^2 + H_3M_f^2 \right),
\]
and set the stepsize
\[
\eta = \frac{\varepsilon}{\sqrt{6 \left( M_g^2 L_f^2 + M_f^2 L_g^2 \right) D_0}}.
\]

Then for the online case, SARAH-Compositional Algorithm 1 outputs an \( \bar{x} \) satisfying \( \mathbb{E} \| \nabla \Phi(\bar{x}) \|^2 \leq 2\varepsilon^2 \) in
\[
\sqrt{6 \left( M_g^2 L_f^2 + M_f^2 L_g^2 \right)} \cdot 2|\Phi(x_0) - \Phi^*| \varepsilon^3.
\]
(a) The risk matrix are generated by a Gaussian distribution with covariance matrix $\Sigma$, $\kappa(\Sigma) = 4$

(b) The risk matrix are generated by a Gaussian distribution with covariance matrix $\Sigma$, $\kappa(\Sigma) = 20$

Figure 1: Experiment on the portfolio management. The $x$-axis is the number of gradients calculations, the $y$-axis is the function value gap and the norm of gradient respectively.

calculation) for all three algorithms in each of the two covariance settings. We observe that SARAH-Compositional outperforms the other two algorithms SCGD and VRSC-PG in our experiments.

The toy experiment justifies that our proposed SARAH-based compositional optimization on portfolio management achieves state-of-the-art performance. Moreover, we note that due to the small size of the batches, basic SCGD achieves a less satisfactory result, a phenomenon also shown by [Huo et al. (2018); Lian et al. (2017)]. In near future, we will provide more simulation results and release our code on Github.

5 Conclusion

In this paper, we propose a novel algorithm called SARAH-Compositional for solving stochastic compositional optimization problems using the idea of a recently proposed variance reduced gradient method. Our algorithm achieves both outstanding theoretical and experimental results. Theoretically, we show that the SARAH-Compositional algorithm can achieve improved convergence rates and IFO complexities for finding an $\varepsilon$-accurate solution to non-convex compositional problems in both finite-sum and online cases. Experimentally, we compare our new compositional optimization method with a few rival algorithms for the task of portfolio management. Future directions include handling the non-smooth case and the theory of lower bounds for stochastic compositional optimization. We hope this work can provide new perspectives to both optimization and machine learning communities interested in compositional optimization.
References


