Non-convex stochastic compositional optimization under heavy-tailed noise

Chunhao Han* Xiao Wang† Pengxiang Xu‡ Jin Zhang§

January 21, 2026

Abstract

This paper investigates non-convex stochastic compositional optimization under heavy-tailed noise, where the stochastic noise exhibits bounded pth moment with $p \in (1, 2]$. The main challenges arise from biased gradient estimates of the objective and the violation of the standard bounded-variance assumption. To address these issues, we propose a generic algorithm framework of Normalized Stochastic Compositional Gradient methods (NSCG) and explore two specific variance-reduced methods within this framework: NSCG-M and NSCG-S. Considering both scenarios with and without prior knowledge of p, we analyze the sample complexity of NSCG-M under standard Lipschitz continuity and smoothness conditions to find an ϵ-stationary point and that of NSCG-S under additional, yet less stringent than existing, mean-pth moment Lipschitzness and mean-pth moment smoothness. The sample complexity orders derived in this paper are competitive with the state-of-the-art results for first-order methods in single-level non-convex stochastic optimization under heavy-tailed noise. Finally, we report numerical experiments results showcasing the effectiveness of the proposed methods.

Keywords: Heavy-tailed noise, stochastic compositional optimization, normalization, variance reduction, sample complexity

1 Introduction

In this paper, we consider the stochastic compositional optimization (SCO) under heavy-tailed noise:

$$\min_{x\in\mathbb{R}^d} \Psi(x) := F(G(x)) = \mathbb{E}_{\xi\sim\Xi_1}[f(\mathbb{E}_{\phi\sim\Xi_2}[g(x;\phi)];\xi)], \tag{P}$$

where the outer function $F : \mathbb{R}^q \to \mathbb{R}$ with $F(y) = \mathbb{E}_{\xi\sim\Xi_1}[f(y;\xi)]$ and the inner function $G : \mathbb{R}^d \to \mathbb{R}^q$ with $G(x) = \mathbb{E}_{\phi\sim\Xi_2}[g(x;\phi)]$ are continuously differentiable for almost every random variable $\xi \in \Xi_1$ and $\phi \in \Xi_2$, respectively, and possibly nonconvex. For problem (P), at any inquiry point x the exact objective value $G(x)$, Jacobian $\nabla G(x)$, and gradient $\nabla F(x)$ are not available, while only stochastic estimates $g(x;\phi)$, $\nabla g(x;\phi)$, and $\nabla f(y;\xi)$ can be obtained. Crucially, the underlying distributions are heavy-tailed, meaning the available stochastic estimates possess only bounded pth (central) moment with $p \in (1, 2]$ [36]. Formally, for any $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^q$,

$$\begin{aligned} &\mathbb{E}[\|g(x;\phi) - G(x)\|^p] \leqslant V_g^p, \qquad \mathbb{E}[\|\nabla g(x;\phi) - \nabla G(x)\|^p] \leqslant V_J^p, \\ &\mathbb{E}[\|\nabla f(y;\xi) - \nabla F(y)\|^p] \leqslant V_f^p, \end{aligned} \tag{1}$$

*Southern University of Science and Technology, Shenzhen, China, and Pengcheng Laboratory, Shenzhen, China (hanchh2024@163.com).

†Sun Yat-sen University, Guangzhou, China (wangx936@mail.sysu.edu.cn).

‡Pengcheng Laboratory, Shenzhen, China (xupx@pcl.ac.cn).

§Southern University of Science and Technology, Shenzhen, China (zhangj9@sustech.edu.cn).

where V_g, V_J and V_f are positive scalars.

In many application fields, such as reinforcement learning [14, 2], portfolio optimization [27, 19], modeling the random variables with heavy-tailed distributions (e.g., α-stable or Pareto-type laws) provides a substantially improved statistical fit. Below we present two examples that motivate the general form of problem (P).

Example 1.1 (Policy Evaluation for Markov Decision Processes). Consider an infinite-horizon discounted Markov Decision Process (MDP), denoted as a tuple $M = (\mathcal{S}, \mathcal{A}, R, P, \bar{\gamma})$ consisting of the state space S, the action space $\mathcal{A}$, a controlled transition kernel P, a random reward function R with expectation $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and a discount factor $\bar{\gamma} \in (0, 1)$. Let $\pi : \mathcal{S} \to \Xi(S)$ be a stationary randomized Markov policy, and define the value function $V^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \bar{\gamma}^t r(s_t, a_t) | s_0 = s]$. Estimating V^π amounts to solving the Bellman equation $V^\pi = r^\pi + \bar{\gamma} P^\pi V^\pi$, where r^π and V^π can be viewed as vectors. As shown in [32], this task can be formulated as a SCO problem:

$$\min_{x \in \mathbb{R}^d} \ell(\mathbb{E}[A]x - \mathbb{E}[b]),$$

where $\ell : \mathbb{R}^d \to \mathbb{R}$ is smooth, $\mathbb{E}[A] = (I - \bar{\gamma} P^\pi)$, and $\mathbb{E}[b] = r^\pi$. This can be formulated as problem (P) with the outer function $f(x) = \ell(x)$ and the inner function $G(x) = \mathbb{E}[A]x - \mathbb{E}[b]$. A multitude of real-world online decision-making systems exhibit heavy-tailed rewards, which can be mathematically modeled as the reward functions r^π with finite pth moment [40, 14, 2].

Example 1.2 (Conditional Value-at-Risk). Conditional Value-at-Risk (CVaR) is a widely used risk metric in portfolio optimization and insurance [27, 19, 26]. For a confidence level $\alpha \in (0, 1)$, CVaR_α is defined as the expected loss beyond the α-quantile of the loss distribution. It admits the variational form [27]:

$$\mathrm{CVaR}_\alpha(Y_\phi) = \inf_{u \in \mathbb{R}} \left\{ u + \frac{1}{1-\alpha} \mathbb{E}_\phi[(Y_\phi - u)_+] \right\},$$

where $(x)_+ = \max(0, x)$ and u is an auxiliary variable which is bounded below. To address the non-smoothness of $(x)_+$ at $x = 0$, we approximate it by a Huber-type function $\ell\varepsilon(x)$ with parameter $\varepsilon > 0$, defined piecewise: $\ell_\varepsilon(x) = 0$ for $x \leqslant 0$, $\ell_\varepsilon(x) = x^2/(2\varepsilon)$ for $0 < x \leqslant \varepsilon$, and $\ell_\varepsilon(x) = x - \varepsilon/2$ otherwise. As stated in [26], Y_ϕ may have heavy-tailed distributions with $\mathbb{E}[\|Y_\phi\|^p] < +\infty$. Minimizing CVaR_α is in the form of the SCO problem (P) with the outer function $f(x) = \inf_u\{u + x/(1-\alpha)\}$ and the inner function $G(x) = \mathbb{E}_\phi[\ell_\varepsilon(x - u)]$. In this scenario, we can derive $\mathbb{E}[\|\ell_\varepsilon(Y_\phi - u)\|^p] < +\infty$ through simple calculations, thereby violating the standard assumption of bounded variance.

The presence of heavy-tailed noise directly challenges the standard bounded-variance assumption (i.e., $p = 2$ in (1)), which serves as the cornerstone for the theoretical guarantees of numerous stochastic optimization algorithms. Consequently, conventional SCO algorithms lose their theoretical guarantees and may become ineffective in such settings. Although the variance may not be infinite in practical scenarios, it can be extremely large, rendering existing SCO algorithms impractical for real-world applications. As highlighted by an example in [36, Remark 1], the presence of heavy-tailed noise can adversely affect the convergence of classical stochastic gradient descent algorithms.

1.1 Related work

In recent years, the study of SCO problems with bounded stochastic variances (i.e., problem (P) with $p = 2$) has advanced considerably, driven by growing computational capabilities and large-scale data applications. Wang et al. [31] introduced the stochastic compositional gradient descent (SCGD) algorithm, which employs an auxiliary variable to track the inner function. For non-convex SCO problem, SCGD obtains a sample complexity $\mathcal{O}(\epsilon^{-8})$ to find an ϵ-stationary point, a point with expected gradient norm below ϵ. Subsequent accelerations via extrapolation techniques improved the sample complexity to $\mathcal{O}(\epsilon^{-7})$.

This accelerated variant was later extended to handle non-smooth penalty terms while maintaining the same sample complexity [32]. Chen et al. [3] proposed a stochastically corrected stochastic compositional gradient (SCSC) method with sample complexity in order $\mathcal{O}(\epsilon^{-4})$, incorporating a linear correction term for the inner function estimate. By lifting the problem to a higher-dimensional space, Ghadimi et al. [10] developed the nested averaged stochastic approximation (NASA) algorithm for the non-convex SCO problems with convex set constraints, also achieving $\mathcal{O}(\epsilon^{-4})$ sample complexity. In [33], the authors proposed data-driven schemes for addressing misspecified SCO problems. Moreover, various variance reduction techniques have also been incorporated into SCO algorithms, leading to improved sample complexity under some additional assumptions (such as, uniform Lipschitzness or average smoothness). Specifically, by the idea of stochastic recursive gradient descent, Hu et al. [13] developed an algorithm named SARAH-Compositional, achieving the sample complexity of $\mathcal{O}(\epsilon^{-3})$. In [37], the authors proposed a composite gradient method with incremental variance reduction (denoted as CIVR) for the SCO problem with a deterministic outer function, which employs the stochastic path-integrated differential estimator (SPIDER) [8]. Chen and Zhou [4] improved the CIVR with the momentum scheme, resulting in the MVRC algorithm for non-smooth regularized SCO. Further extensions include a normalized proximal approximate gradient method with nested variance reduction [38] and projection-free variance-reduced methods [17]. Despite these advances, the theoretical analysis in all aforementioned works fundamentally depends on the bounded-variance assumption, rendering them inapplicable to the heavy-tailed noise setting considered in this work.

Existing algorithms for handling heavy-tailed noise have primarily focused on single-level stochastic optimization (SO):

$$\min_{x\in\mathbb{R}^d} G(x) := \mathbb{E}_{\phi\sim\Xi_2}[g(x;\phi)], \tag{2}$$

where the stochastic estimate $\nabla g(x;\phi)$ is unbiased with respect to $G(x)$ and its error possesses a bounded pth moment. A prevalent strategy to mitigate the impact of heavy-tailed noise is the gradient clipping technique, which thresholds gradient norms to suppress outliers. Zhang et al. [36] first applied this technique to non-convex SO problem (2), proposing a clipped stochastic gradient descent method (SGD-C) that attains a sample complexity of $\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$ in expectation, together with a matching sample complexity lower bound. Subsequent works have aimed to improve these guarantees. Cutkosky et al. [5] combined gradient clipping with normalization and momentum, achieving the sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-\frac{3p-2}{p-1}})$ in high-probability sense, where $\tilde{\mathcal{O}}(\cdot)$ further hides the poly-logarithmic factors. Liu et al. [22] further introduced the variance reduction strategy to break the previous lower bound, and then attain an improved sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-\frac{2p-1}{p-1}})$ based on the uniform Lipschitzness of ∇g. Relying solely on gradient clipping technique, Nguyen et al. [24] established a high-probability convergence result for the SGD-C that match the lower bound of the convergence rate. In recent years, a variety of other stochastic optimization algorithms with gradient clipping have also been developed to mitigate the effects of heavy-tailed noise [20, 29, 25, 18, 11, 35].

More recent study indicates that, while gradient clipping-based algorithms can provide convergence guarantees, there is a discrepancy between the theoretical insights and practical applications of this technique. Specifically, as stated in [15, 12], large clipping thresholds are required to ensure theoretically convergence, whereas small thresholds are typically employed in real-world applications. This has motivated the development of clipping-free algorithms. Sun et al. [30] proposed a normalized SGD with momentum (NSGD-M) achieving the sample complexity of $\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$. By incorporating a variance reduction technique, they also developed the NSGD-VR methods, improving sample complexity to $\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$ with the uniform Lipschitzness condition on ∇g. Under generalized heavy-tailed noise, Liu and Zhou [23] proposed a batched variant of NSGD-M. The notable work [15] indicates that normalized SGD (NSGD) algorithm with batched sampling can also effectively handle the heavy-tailed noise. Convergence results

were established in both expectation and high-probability sense, matching existing lower bounds of sample complexity. Moreover, the convergence results of the NSGD-M without batched sampling were analyzed in [15]. He et al. [12] designed NSGD-M variants using three momentum strategies, achieving optimal complexity without the requirement of explicit knowledge of problem-specific quantities. More recently, the vanilla SGD algorithm, devoid of any additional strategies such as gradient clipping or normalization, has been demonstrated to effectively handle heavy-tailed noise in [16]. However, under standard smoothness assumptions, its sample complexity with $\mathcal{O}(\epsilon^{-\frac{2p}{p-1}})$ is suboptimal for non-convex SO problem (2). Despite these extensive advances for SO under heavy-tailed noise, the intricate compositional structure of the SCO problem (P) prevents the direct and efficient application of existing algorithms, calling for novel algorithmic designs tailored to this setting.

1.2 Challenges

The main challenges in solving the SCO problem (P) with heavy-tailed noise lie in its compositional structure and the failure of the standard bounded-variance assumption. In particular, the algorithms developed in this paper needs to resolve the following issues.

- Existing algorithms that effectively handle heavy-tailed noise in SO (2) crucially rely on the unbiasedness of the stochastic gradient estimator. In the SCO problem (P), however, the randomness in both layers of functions makes it difficult to construct an unbiased estimator for the gradient of the compositional objective. This structural limitation prevents the direct application of SO methods to the SCO context.
- Conventional SCO algorithms, whose convergence analysis depends heavily on the bounded-variance assumption, are no longer theoretically justified in the presence of heavy-tailed noise. Moreover, even seemingly mild assumptions, such as the uniform Lipschitzness of the inner function g, must be carefully re-examined to ensure they remain compatible with heavy-tailed distributional assumptions.

1.3 Contributions

In this work we study the SCO problem (P) under heavy-tailed noise. We propose an algorithm framework for Normalized Stochastic Compositional Gradient methods (NSCG), which relies on estimates for the inner function G, its Jacobian ∇G, and the outer gradient ∇F, and updates variables via gradient normalization. We introduce two specific variance-reduced methods, NSCG-M and NSCG-S, and provide a sample complexity analysis for finding an ϵ-stationary point (see (3)). Under standard Lipschitzness and smoothness conditions, NSCG-M achieves a sample complexity of $\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$ with the prior knowledge of p, and $\mathcal{O}(\epsilon^{-\frac{2p}{p-1}})$ when p is unknown. By further introducing the mean-pth moment Lipschitzness and mean-pth moment smoothness (Assumptions 4), NSCG-S attains the sample complexity of $\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$ if p is known. Without prior knowledge of p, the corresponding sample complexity order becomes $\mathcal{O}(\epsilon^{-\frac{3p}{2p-2}})$. Moreover, when $p = 2$, they recover the optimal sample complexity $\mathcal{O}(\epsilon^{-3})$ obtained by first order algorithms for SCO problems. A detailed comparison with existing algorithms is summarized in Table 1.

1.4 Notations and organization

For any integer $T \geqslant 1$, $[T]$ denotes the set $\{1, 2, ..., T\}$. The Euclidean norm of a vector $x \in \mathbb{R}^d$ is $\|x\| = \sqrt{x^\top x}$, and the corresponding operator norm For a matrix $A \in \mathbb{R}^{q\times d}$ is $\|A\| = \max\{\|Ax\| \mid \|x\| = 1\}$. The standard Euclidean inner product is written as $\langle \cdot, \cdot \rangle$.

Table 1: Sample complexity of related algorithms for finding an ϵ-stationary point in non-convex SO and SCO.

Problem	**Algorithm**	**Moment Order**	**Assumptions**	**Sample Complexity**
SO (2)	NSGD-M [23, 15]	$p \in (1, 2]$	Lipschitzness of ∇G	$\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$ $\mathcal{O}(\epsilon^{-\frac{2p}{p-1}})$*
	NSGD-VR [30]	$p \in (1, 2]$	Uniform Lipschitzness of ∇g	$\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$
	NSGD-M [12]	$p \in (1, 2]$	Lipschitzness of ∇G; Mean-pth moment smoothness of g	$\mathcal{O}((\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))^{\frac{2p-1}{p-1}})$ $\mathcal{O}((\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))^{\frac{3p}{2p-2}})$*
SCO (P)	NASA [10]	$p = 2$	Lipschitzness of F, ∇F, G, ∇G	$\mathcal{O}(\epsilon^{-4})$
	SCSC [3]	$p = 2$	Uniform Lipschitzness of ∇f, ∇g	$\mathcal{O}(\epsilon^{-4})$
	MVRC [4]	$p = 2$	Uniform Lipschitzness of f, ∇f, g, ∇g	$\mathcal{O}(\epsilon^{-3})$
	NSCG-M	$p \in (1, 2]$	Lipschitzness of F, ∇F, G, ∇G	$\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$ (Theorem 1) $\mathcal{O}(\epsilon^{-\frac{2p}{p-1}})$* (Theorem 2)
	NSCG-S	$p \in (1, 2]$	Lipschitzness of F, ∇F, G, ∇G; Mean-pth moment Lipschitzness of g; Mean-pth moment smoothness of f, g	$\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$ (Theorem 3) $\mathcal{O}(\epsilon^{-\frac{3p}{2p-2}})$* (Theorem 4)

[1] Uniform Lipschitzness of function g denotes that for any $x, y \in \mathbb{R}^d$, there exits a constant $\ell > 0$ such that $\|g(x;\phi) - g(y;\phi)\| \leqslant \ell\|x - y\|$ for almost every ϕ, and Lipschitzness of function G refers to $\|G(x) - G(y)\| \leqslant \ell\|x - y\|$.
[2] Mean-pth moment Lipschitzness of g refers to Assumption 4 (a). Mean-pth moment smoothness of g and f refers to Assumption 4 (b) and (c), respectively.
[3] The asterisk * indicates the sample complexity without prior knowledge of the tail index p.

The remainder of the paper is structured as follows. Section 2 presents the foundational assumptions. Section 3 introduces a general framework for NSCG methods. Section 4 details the two specific variance-reduced methods: NSCG-M and NSCG-S, and analyzes their sample complexity for solving the SCO problem (P) under heavy-tailed noise. Numerical experiments are reported in Section 5, and conclusions are drawn in Section 6.

2 Preliminaries

This section presents the foundational assumptions for our analysis and discusses the specific challenges of solving the SCO problem (P) under heavy-tailed noise. We begin by stating the assumptions used throughout the remainder of this paper.

Assumption 1 *The objective function $\Psi(x)$ is bounded below, i.e., $\Psi^* = \inf_{x\in\mathbb{R}^d} \Psi(x) > -\infty$.*

Assumption 2 *The function F is ℓ_F-Lipschitz continuous and its gradient function ∇F is L_F-Lipschitz continuous, and the function G is ℓ_G-Lipschitz continuous and its Jacobian ∇G is L_G-Lipschitz continuous.*

The above two assumptions are standard in SCO literature (see e.g., [4, 37, 10]). We proceed by presenting the crucial assumptions of unbiasedness and review the assumptions regarding heavy-tailed noise. For integer $t > 0$, let $\{\xi_k\}_{k=1}^t \sim \Xi_1$, $\{\phi_k\}_{k=1}^t \sim \Xi_2$, and $\{\hat{\phi}_k\}_{k=1}^t \sim \Xi_2$ be mutually independent random variables sampled during the iterative process.

Assumption 3 *For any given $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^q$, the stochastic oracle returns stochastic estimates $g(x;\phi_t)$, $\nabla f(y;\xi_t)$ and $\nabla g(x;\hat{\phi}_t)$ that satisfy*

$$\begin{aligned}
&\mathbb{E}[g(x;\phi_t)] = G(x), && \mathbb{E}[\|g(x;\phi_t) - G(x)\|^p] \leqslant V_g^p,\\
&\mathbb{E}[\nabla g(x;\hat{\phi}_t)] = \nabla G(x), && \mathbb{E}[\|\nabla g(x;\hat{\phi}_t) - \nabla G(x)\|^p] \leqslant V_J^p,\\
&\mathbb{E}[\nabla f(y;\xi_t)] = \nabla F(y), && \mathbb{E}[\|\nabla f(y;\xi_t) - \nabla F(y)\|^p] \leqslant V_f^p.
\end{aligned}$$

Assumption 3 includes the standard bounded-variance condition (when $p = 2$), but in general is weaker. Most existing algorithms designed for handling heavy-tailed noise cannot be directly applied to problem (P) due to its compositional structure. A key limitation lies in their reliance on unbiased gradient estimates of the overall objective. By the chain rule, the gradient of the objective is given by $\nabla\Psi(x) = \nabla G(x)^\top \nabla f(G(x))$. However, under Assumption 3, the stochastic oracles only provide unbiased estimates for each individual component, but not for the full composition. And constructing an unbiased estimate of $\nabla\Psi$ is computationally expensive and even infeasible in practice [31, 32, 3]. This fact prevents the direct application of algorithms designed for the SO problem (2) to the SCO problem (P).

Furthermore, as illustrated by the examples in Introduction, problem (P) involves heavy-tailed noise in the estimate of the inner function G itself. Such random noise may propagate to the Jacobian estimate, rendering its stochastic noise distribution heavy-tailed [2]. Consequently, it meets only the bounded-pth moment condition as stated in Assumption 3, not the standard bounded-variance condition. Moreover, we consider a general setting that the stochastic estimate noise of gradient ∇F also follows the heavy-tailed distribution. Due to the compositional form of the gradient $\nabla G(x)^\top \nabla F(G(x))$, the noise from each source may further accumulate and propagate through the chain of estimation. This phenomenon leads to amplified bias and variance in the full gradient estimate. These factors collectively significantly distort the gradient of $\Psi(x)$, making existing SCO algorithms that rely on finite-variance assumptions ill-suited for the heavy-tailed setting considered here.

Algorithm 1

Require: Initial point $x_1 \in \mathbb{R}^d$, step size $\alpha > 0$, time horizon T.

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

2: Call the stochastic oracles to construct the estimates g_t, ∇g_t and ∇f_t.

3: Compute the estimate $\widehat{\nabla}_t^{f,g} = \nabla g_t^\top \nabla f_t$.

4: Update variable x_{t+1} through

$$x_{t+1} = x_t - \alpha \frac{\widehat{\nabla}_t^{f,g}}{\|\widehat{\nabla}_t^{f,g}\|}. \tag{4}$$

5: **end for**

This work aims to develop tailored algorithms to solve the SCO problem (P) under heavy-tailed noise, and to establish their sample complexity to find an *ϵ-stationary point*, i.e., a point $x \in \mathbb{R}^d$ satisfying

$$\mathbb{E}[\|\nabla \Psi(x)\|] \leqslant \epsilon, \tag{3}$$

where the expectation is taken with respect to all random variables generated during the algorithm process.

3 Algorithm framework of normalized stochastic compositional gradient methods

In this section, we propose the algorithm framework of NSCG methods for solving the SCO problem (P) under heavy-tailed noise, outlined in Algorithm 1. At tth iteration, stochastic oracles provide the estimates g_t, ∇g_t, and ∇f_t to approximate the true values G, ∇G, and ∇F, respectively. Their composition $\widehat{\nabla}_t^{f,g} := \nabla g_t^\top \nabla f_t$ serves as an estimate of the full gradient $\nabla G(x_t)^\top \nabla F(G(x_t))$. We do not require the stochastic estimate $\widehat{\nabla}_t^{f,g}$ be unbiased; instead, we focus on controlling the bias of this estimate. As stated in Assumption 3, heavy-tailed noise violates the standard assumption of bounded variance, leading to significantly increased bias in stochastic estimates. To mitigate this effect, we normalize the estimate $\widehat{\nabla}_t^{f,g}$ before updating x_t along its negative direction. This normalization strategy effectively transfers the estimation bias to be precisely controlled by the step size α [5]. As a result, desired convergence properties can be obtained through a carefully tuned step size schedule. For convenience, let

$$\Delta_t := \Psi(x_t) - \Psi^*, \quad \widehat{\nabla}_t^{f,g} := \nabla g_t^\top \nabla f_t, \quad \widehat{\nabla}_t^{g} := \nabla g_t^\top \nabla F(g_t), \quad \nabla_t^{g} := \nabla G(x_t)^\top \nabla F(g_t).$$

The following lemma provides an estimate on the averaged gradient norm at iterates.

Lemma 1 *Suppose Assumptions 1 and 2 hold. Then, the iterates generated by Algorithm 1 satisfies*

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla \Psi(x_t)\|] \leqslant \frac{2}{T}\sum_{t=1}^{T} \mathbb{E}[\|\widehat{\nabla}_t^{f,g} - \widehat{\nabla}_t^{g}\|] + \frac{2\ell_F}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla g_t - \nabla G(x_t)\|] \\ + \frac{2\ell_G L_F}{T}\sum_{t=1}^{T} \mathbb{E}[\|g_t - G(x_t)\|] + \frac{\Delta_1}{\alpha T} + \frac{\alpha L}{2}, \end{aligned} \tag{5}$$

where $L := \ell_G^2 L_F + L_G \ell_F$.

Proof. From Assumption 2, $\nabla\Psi$ is L-Lipschitz continuous (see, e.g. [10]). It follows that

$$\begin{aligned}
&\Psi(x_{t+1}) - \Psi(x_t) \\
&\leqslant \langle \nabla\Psi(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&= -\alpha\langle \nabla\Psi(x_t), \frac{\widehat{\nabla}_t^{f,g}}{\|\widehat{\nabla}_t^{f,g}\|}\rangle + \frac{\alpha^2 L}{2} \\
&= -\alpha\|\widehat{\nabla}_t^{f,g}\| + \alpha\langle \widehat{\nabla}_t^{f,g} - \nabla\Psi(x_t), \frac{\widehat{\nabla}_t^{f,g}}{\|\widehat{\nabla}_t^{f,g}\|}\rangle + \frac{\alpha^2 L}{2} \\
&\overset{(*)}{\leqslant} -\alpha\|\widehat{\nabla}_t^{f,g}\| + \alpha\|\widehat{\nabla}_t^{f,g} - \nabla\Psi(x_t)\| + \frac{\alpha^2 L}{2} \\
&\leqslant -\alpha\|\nabla\Psi(x_t)\| + 2\alpha\left(\|\widehat{\nabla}_t^{f,g} - \widehat{\nabla}_t^{g}\| + \|\widehat{\nabla}_t^{g} - \nabla_t^{g}\| + \|\nabla_t^{g} - \nabla\Psi(x_t)\|\right) + \frac{\alpha^2 L}{2},
\end{aligned}$$

where $(*)$ is by

$$\langle \widehat{\nabla}_t^{f,g} - \nabla\Psi(x_t), \widehat{\nabla}_t^{f,g}/\|\widehat{\nabla}_t^{f,g}\|\rangle \leqslant \|\widehat{\nabla}_t^{f,g} - \nabla\Psi(x_t)\|\|\widehat{\nabla}_t^{f,g}/\|\widehat{\nabla}_t^{f,g}\|\| = \|\widehat{\nabla}_t^{f,g} - \nabla\Psi(x_t)\|,$$

and the last inequality is due to

$$\|\nabla\Psi(x_t)\| \leqslant \|\widehat{\nabla}_t^{f,g}\| + \|\widehat{\nabla}_t^{f,g} - \widehat{\nabla}_t^{g}\| + \|\widehat{\nabla}_t^{g} - \nabla_t^{g}\| + \|\nabla_t^{g} - \nabla\Psi(x_t)\|.$$

Taking the expectation on both sides gives

$$\begin{aligned}
\mathbb{E}[\Psi(x_{t+1})] - \mathbb{E}[\Psi(x_t)] \leqslant\ & 2\alpha\mathbb{E}[\|\widehat{\nabla}_t^{f,g} - \widehat{\nabla}_t^{g}\|] + 2\alpha\mathbb{E}[\|\widehat{\nabla}_t^{g} - \nabla_t^{g}\|] \\
& + 2\alpha\mathbb{E}[\|\nabla_t^{g} - \nabla\Psi(x_t)\|] - \alpha\mathbb{E}[\|\nabla\Psi(x_t)\|] + \frac{\alpha^2 L}{2}.
\end{aligned} \tag{6}$$

By Assumption 2, we obatin

$$\mathbb{E}[\|\widehat{\nabla}_t^{g} - \nabla_t^{g}\|] \leqslant \mathbb{E}[\|\nabla g_t - \nabla G(x_t)\|\|\nabla F(g_t)\|] \leqslant \ell_F\mathbb{E}[\|\nabla g_t - \nabla G(x_t)\|].$$

An analogous inequality also holds for $\mathbb{E}[\|\nabla_t^{g} - \nabla\Psi(x_t)\|]$:

$$\mathbb{E}[\|\nabla_t^{g} - \nabla\Psi(x_t)\|] \leqslant \mathbb{E}[\|\nabla G(x_t)\|\|\nabla F(g_t) - \nabla F(G(x_t))\|] \leqslant \ell_G L_F\mathbb{E}[\|g_t - G(x_t)\|].$$

Substituting the above two inequalities into (6) and adding up the recursion from $t = 1$ to T yield the desired result. □

According to Lemma 1, a convergence guarantee follows if we can control the cumulative errors in (5), i.e., $\sum_{t=1}^T \mathbb{E}[\|\widehat{\nabla}_t^{f,g} - \widehat{\nabla}_t^{g}\|]$, $\sum_{t=1}^T \mathbb{E}[\|\nabla g_t - \nabla G(x_t)\|]$, and $\sum_{t=1}^T \mathbb{E}[\|g_t - G(x_t)\|]$. This motivates our algorithm designs and analysis in the next section.

4 Variance-reduced NSCG methods

The key step in Algorithm 1 is the construction of stochastic estimates g_t, ∇g_t and ∇f_t. This section presents two variance-reduced methods based on the mini-batch estimator and SPIDER, respectively. We will establish the sample complexity of each method for finding an ϵ-stationary point of the SCO problem (P). To proceed, we first introduce three technical lemmas.

Lemma 2 *([15, Lemma 10]) Let $p \in (1,2]$, and $M_1, \ldots, M_n \in \mathbb{R}^d$ be a martingale difference sequence satisfying $\mathbb{E}[\|M_j\|^p] < +\infty$ for all $j = 1, \ldots, n$. Then,*

$$\mathbb{E}\left[\left\|\sum_{j=1}^{n} M_j\right\|^p\right] \leqslant 2\sum_{j=1}^{n} \mathbb{E}\left[\|M_j\|^p\right].$$

Lemma 3 *([7, Example 4.1.7]) Let X and Y be independent random variables valued in (E_1, Σ_1) and (E_2, Σ_2). For any measurable $h : E_1 \times E_2 \to \mathbb{R}^d$ with $\mathbb{E}[\|h(X,Y)\|] < +\infty$, define $g(x) = \mathbb{E}[h(x,Y)]$. Then it holds almost surely that $\mathbb{E}[h(X,Y) \mid X] = g(X)$.*

Lemma 4 *([28, Theorem 6.3]) For any $a, b \in \mathbb{R}^d$ and $b \neq 0$, it holds that*

$$\|a+b\|^p \leqslant 2^{2-p}\|a\|^p + \|b\|^p + p\frac{\langle a, b\rangle}{\|b\|^{2-p}}.$$

4.1 The NSCG-M method

NSCG-M refers to Algorithm 1 that uses mini-batch estimator to generate stochastic estimates. More specifically, at tth iteration of Algorithm 1, we generate three index sets consisting i.i.d. samples:

$$\mathcal{B}_{t,1} = \{\phi_t^{(1)}, \phi_t^{(2)}, \ldots, \phi_t^{(B_{t,1})}\}, \qquad \mathcal{B}_{t,2} = \{\hat{\phi}_t^{(1)}, \hat{\phi}_t^{(2)}, \ldots, \hat{\phi}_t^{(B_{t,2})}\}, \qquad \mathcal{B}_{t,3} = \{\xi_t^{(1)}, \xi_t^{(2)}, \ldots, \xi_t^{(B_{t,3})}\},$$

where $B_{t,1}, B_{t,2}$ and $B_{t,3}$ are positive integers. Then, we compute the estimates

$$g_t = \frac{1}{B_{t,1}}\sum_{i=1}^{B_{t,1}} g(x_t; \phi_t^{(i)}), \quad \nabla g_t = \frac{1}{B_{t,2}}\sum_{i=1}^{B_{t,2}} \nabla g(x_t; \hat{\phi}_t^{(i)}), \quad \nabla f_t = \frac{1}{B_{t,3}}\sum_{i=1}^{B_{t,3}} \nabla f(g_t; \xi_t^{(i)}). \tag{7}$$

Although NSCG-M can be regarded as an extension of the algorithm in [15] to the SCO problem (P) under heavy-tailed noise, it is not trivial to analyze its theoretical behavior due to the challenges we mentioned earlier. We then provide upper bounds for the three error estimates in NSCG-M.

Lemma 5 *Suppose that Assumptions 2 and 3 hold. Let $\{(x_t, g_t, \nabla g_t, \nabla f_t)\}$ be the sequence generated by NSCG-M. Then, for any $t \in [T]$,*

$$\mathbb{E}[\|\widehat{\nabla}_t^{f,g} - \widehat{\nabla}_t^{g}\|] \leqslant 4(V_J + \ell_G)V_f B_{t,3}^{-\frac{p-1}{p}}, \tag{8}$$

$$\mathbb{E}[\|\nabla g_t - \nabla G(x_t)\|] \leqslant 2V_J B_{t,2}^{-\frac{p-1}{p}}, \tag{9}$$

$$\mathbb{E}[\|g_t - G(x_t)\|] \leqslant 2V_g B_{t,1}^{-\frac{p-1}{p}}. \tag{10}$$

Proof. By the independence of $\mathcal{B}_{t,1}$, $\mathcal{B}_{t,2}$, and $\mathcal{B}_{t,3}$, we obtain

$$\begin{aligned}
\mathbb{E}[\|\widehat{\nabla}_t^{f,g} - \widehat{\nabla}_t^{g}\| | x_t, g_t] &\leqslant \mathbb{E}[\|\nabla g_t\| | x_t]\mathbb{E}[\|\nabla f_t - \nabla F(g_t)\| | g_t] \\
&\leqslant \left(\frac{1}{B_{t,2}}\sum_{i=1}^{B_{t,2}} \mathbb{E}\left[\left\|\nabla g\left(x_t; \hat{\phi}_t^{(i)}\right)\right\| | x_t\right]\right) \\
&\quad \cdot \frac{1}{B_{t,3}}\mathbb{E}\left[\left\|\sum_{i=1}^{B_{t,3}}\left(\nabla f\left(g_t; \xi_t^{(i)}\right) - \nabla F(g_t)\right)\right| g_t\right\|\right].
\end{aligned} \tag{11}$$

For the first term on the right-hand side of the aforementioned inequality, Jensen's inequality implies that

$$\begin{aligned}
&\frac{1}{B_{t,2}}\sum_{i=1}^{B_{t,2}}\mathbb{E}\left[\left\|\nabla g\left(x_t;\hat{\phi}_t^{(i)}\right)\right\||x_t\right] \leqslant \frac{1}{B_{t,2}}\sum_{i=1}^{B_{t,2}}\mathbb{E}\left[\left\|\nabla g\left(x_t;\hat{\phi}_t^{(i)}\right)\right\|^p|x_t\right]^{\frac{1}{p}} \\
&\leqslant \frac{2}{B_{t,2}}\sum_{i=1}^{B_{t,2}}\mathbb{E}\left[\left\|\nabla g\left(x_t;\hat{\phi}_t^{(i)}\right)-\nabla G(x_t)\right\|^p+\|\nabla G(x_t)\|^p|x_t\right]^{\frac{1}{p}} \\
&\leqslant 2(V_J+\ell_G),
\end{aligned} \tag{12}$$

where the last inequality is by Assumptions 2 and 3 and the fact that $(a+b)^{\frac{1}{p}} \leqslant a^{\frac{1}{p}}+b^{\frac{1}{p}}$ with $a,b\geqslant 0$.

Next, we use Lemma 2 to bound $\mathbb{E}[\|\sum_{i=1}^{B_{t,3}}(\nabla f(g_t;\xi_t^{(i)})-\nabla F(g_t))\||g_t]$, which follows the proof of [15, Proposition 1]. Let $M_t^i := \nabla f(g_t;\xi_t^{(i)})-\nabla F(g_t)$, $i=1,\ldots,B_{t,3}$. Based on the fact that $M_t^1,\cdots,M_t^i$ are independent random variables and g_t is $\sigma(M_t^1,\cdots,M_t^{i-1})$ measurable, we have

$$\mathbb{E}[M_t^i|M_t^1,\cdots,M_t^{i-1}] = \mathbb{E}[\nabla f(g_t;\xi_t^{(i)})-\nabla F(g_t)|g_t] = 0.$$

Additionally, by Assumption 3, we obtain

$$\mathbb{E}[\|M_t^i\|^p] = \mathbb{E}[\mathbb{E}[\|\nabla f(g_t;\xi_t^{(i)})-\nabla F(g_t)\|^p|g_t]] \leqslant V_f^p < +\infty.$$

Thus, applying Lemma 2 yields

$$\mathbb{E}\left[\left\|\sum_{i=1}^{B_{t,3}}\left(\nabla f\left(g_t;\xi_t^{(i)}\right)-\nabla F(g_t)\right)\right\|^p\right] \leqslant 2\sum_{i=1}^{B_{t,3}}\mathbb{E}\left[\left\|\nabla f\left(g_t;\xi_t^{(i)}\right)-\nabla F(x_t)\right\|^p\right] \leqslant 2B_{t,3}V_f^p. \tag{13}$$

Let $Z=(\xi_t^{(1)},\ldots,\xi_t^{(B_{t,3})})$ and $s(g_t,Z)=\|\sum_{i=1}^{B_{t,3}}(\nabla f(g_t;\xi_t^{(i)})-\nabla F(g_t))\|^p$. According to the independence of g_t and Z, we apply Lemma 3 obtaining

$$\mathbb{E}\left[s(g_t,Z)|g_t\right]^{\frac{1}{p}} = \mathbb{E}\left[\left\|\sum_{i=1}^{B_{t,3}}\left(\nabla f\left(g_t;\xi_t^{(i)}\right)-\nabla F(g_t)\right)\right\|^p\right]^{\frac{1}{p}} \overset{(13)}{\leqslant} 2B_{t,3}^{\frac{1}{p}}V_f.$$

Furthermore, applying Jensen's inequality gives

$$\mathbb{E}\left[\left\|\sum_{i=1}^{B_{t,3}}\left(\nabla f\left(g_t;\xi_t^{(i)}\right)-\nabla F(g_t)\right)\right\||g_t\right] \leqslant \mathbb{E}[s(g_t,Z)|g_t]^{\frac{1}{p}} \leqslant 2B_{t,3}^{\frac{1}{p}}V_f. \tag{14}$$

Finally, taking the expectation over (11) and substituting (12) and (14) leads to (8). Through completely similar analyses, we can also obtain (9) and (10). □

We now establish the sample complexity results of NSCG-M to find an ϵ-stationary point.

Theorem 1 (complexity with known p) *Suppose Assumptions 1, 2, and 3 hold, and for any given $T\in\mathbb{N}_+$, let $\alpha=(\frac{\Delta_1}{LT})^{\frac{1}{2}}$. The estimates in (7) use the batch sizes $B_{t,1}=\lceil(b_1T)^{\frac{p}{2p-2}}\rceil$ with $b_1=\frac{16\ell_G^2L_F^2V_g^2}{\Delta_1 L}$, $B_{t,2}=\lceil(b_2T)^{\frac{p}{2p-2}}\rceil$ with $b_2=\frac{16\ell_F^2V_J^2}{\Delta_1 L}$, and $B_{t,3}=\lceil(b_3T)^{\frac{p}{2p-2}}\rceil$ with $b_3=\frac{64(V_J+\ell_G)^2V_f^2}{\Delta_1 L}$. Then, the iterates generated by NSCG-M satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\Psi(x_t)\|] = \mathcal{O}\left(T^{-\frac{1}{2}}\right),$$

and the sample complexity to find an ϵ-stationary point is in order $\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$.

Proof. By substituting the results in Lemma 5 into Lemma 1, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\Psi(x_t)\|] \leqslant \frac{4}{T}\sum_{t=1}^{T}\left(2(V_J+\ell_G)V_f B_{t,3}^{-\frac{p-1}{p}} + \ell_F V_J B_{t,2}^{-\frac{p-1}{p}} + \ell_G L_F V_g B_{t,1}^{-\frac{p-1}{p}}\right) + \frac{\Delta_1}{\alpha T} + \frac{\alpha L}{2}. \tag{15}$$

According to the choice of parameters α, $B_{t,1}$ $B_{t,2}$ and $B_{t,3}$, we further have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\Psi(x_t)\|] &\leqslant \frac{8(V_J+\ell_G)V_f}{(b_3T)^{\frac{1}{2}}} + \frac{4\ell_F V_J}{(b_2T)^{\frac{1}{2}}} + \frac{4\ell_G L_F V_g}{(b_1T)^{\frac{1}{2}}} + \frac{\sqrt{\Delta_1 L}}{T^{\frac{1}{2}}} + \frac{\sqrt{\Delta_1 L}}{2T^{\frac{1}{2}}} \\ &\leqslant \frac{5\sqrt{\Delta_1 L}}{T^{\frac{1}{2}}} = \mathcal{O}\left(T^{-\frac{1}{2}}\right).\end{aligned}$$

To find an ϵ-stationary point, we set the number of iterations $T = \mathcal{O}(\epsilon^{-2})$. Consequently, the total sample complexity (denoted as N) is calculated as

$$\begin{aligned}N = \sum_{t=1}^{T}(B_{t,1}+B_{t,2}+B_{t,3}) &= \mathcal{O}\left(\left((b_1)^{\frac{p}{2p-2}} + (b_2)^{\frac{p}{2p-2}} + (b_3)^{\frac{p}{2p-2}}\right)\left(\frac{\Delta_1 L}{\epsilon^2}\right)^{1+\frac{p}{2p-2}}\right) \\ &= \mathcal{O}\left(\epsilon^{-\frac{3p-2}{p-1}}\right).\end{aligned}$$

This completes the proof. □

NSCG-M achieves the same order of sample complexity as the algorithms for non-convex SO problems (2) [36, 15, 30, 23, 12]. Moreover, when $p = 2$ (i.e., the finite variance case), the sample complexity of NSCG-M reduces to $\mathcal{O}(\epsilon^{-4})$, matching the optimal results in the existing work of non-convex SCO problems [10, 3].

Theorem 1 establishes the sample complexity of NSCG-M when the tail index p is known. However, in many practical computations it is difficult to determine the value of p. Under such circumstances, NSCG-M owns the following properties.

Theorem 2 (complexity with unknown p) *Suppose Assumptions 1, 2, and 3 hold, and for any given $T \in \mathbb{N}_+$, let $\alpha = (\frac{\Delta_1}{LT})^{\frac{1}{2}}$. The estimates in (7) use the batch sizes $B_{t,1} = \lceil b_1 T\rceil$ with $b_1 = \ell_G^2 L_F^2 V_g^2$, $B_{t,2} = \lceil b_2 T\rceil$ with $b_2 = \ell_F^2 V_J^2$, and $B_{t,3} = \lceil b_3 T\rceil$ with $b_3 = (V_J+\ell_G)^2 V_f^2$. Then, the iterates generated by NSCG-M satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\Psi(x_t)\|] = \mathcal{O}\left(\frac{((V_J+\ell_G)V_f)^{\frac{2-p}{p}} + (\ell_F V_J)^{\frac{2-p}{p}} + (\ell_G L_F V_g)^{\frac{2-p}{p}}}{T^{\frac{p-1}{p}}} + \frac{\sqrt{\Delta_1 L}}{T^{\frac{1}{2}}}\right),$$

which is in order $\mathcal{O}(T^{-\frac{p-1}{p}})$. Moreover, the sample complexity of NSCG-M to find an ϵ-stationary point of problem (P) is in order $\mathcal{O}(\epsilon^{-\frac{2p}{p-1}})$.

Proof. Inequality (15) in the proof of Theorem 1 remains valid. With the chosen parameters α, $B_{t,1}$, $B_{t,2}$, and $B_{t,3}$, it follows directly that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\Psi(x_t)\|] &\leqslant \frac{8(V_J+\ell_G)V_f}{(b_3T)^{\frac{p-1}{p}}} + \frac{4\ell_F V_J}{(b_2T)^{\frac{p-1}{p}}} + \frac{4\ell_G L_F V_g}{(b_1T)^{\frac{p-1}{p}}} + \frac{\sqrt{\Delta_1 L}}{T^{\frac{1}{2}}} + \frac{\sqrt{\Delta_1 L}}{2T^{\frac{1}{2}}} \\ &\leqslant \frac{8((V_J+\ell_G)V_f)^{\frac{2-p}{p}} + 4(\ell_F V_J)^{\frac{2-p}{p}} + 4(\ell_G L_F V_g)^{\frac{2-p}{p}}}{T^{\frac{p-1}{p}}} + \frac{2\sqrt{\Delta_1 L}}{T^{\frac{1}{2}}} \\ &= \mathcal{O}\left(\frac{((V_J+\ell_G)V_f)^{\frac{2-p}{p}} + (\ell_F V_J)^{\frac{2-p}{p}} + (\ell_G L_F V_g)^{\frac{2-p}{p}}}{T^{\frac{p-1}{p}}} + \frac{\sqrt{\Delta_1 L}}{T^{\frac{1}{2}}}\right) \\ &= \mathcal{O}\left(T^{-\frac{p-1}{p}}\right).\end{aligned}$$

Following the same derivation logic as in Theorem 1, the sample complexity to find an ϵ-stationary point is

$$N = \mathcal{O}\left((b_1 + b_2 + b_3)\left(\frac{\Delta_1 L}{\epsilon^{\frac{p}{p-1}}}\right)^{1+1}\right) = \mathcal{O}\left(\epsilon^{-\frac{2p}{p-1}}\right).$$

This completes the proof. □

The proof of Theorem 2 follows a reasoning similar to that of Theorem 1. Note that Theorem 2 presents a sample complexity of $\mathcal{O}(\epsilon^{-\frac{2p}{p-1}})$, matching the result for the SO methods without the prior knowledge of p [12, 23]. Moreover, for $p = 2$, this result can still be reduced to the sample complexity of $\mathcal{O}(\epsilon^{-4})$.

4.2 The NSCG-S method

NSCG-S refers to Algorithm 1 using the variance reduction technique SPIDER [8]. Formally, NSCG-S is a double-loop method. In each outer iteration $t \in [T]$, we use large sample sizes $\mathcal{B}_{t,1}$, $\mathcal{B}_{t,2}$ and $\mathcal{B}_{t,3}$ to construct the estimates of $G(x_{t,0})$, $\nabla G(x_{t,0})$ and $\nabla F(g_{t,0})$, respectively. For inner iteration $j \in [\tau_t - 1]$ with integer $\tau_t \geqslant 1$, the SPIDER estimator is employed in combination with relatively smaller sample sets $\mathcal{S}_{t,j,1}$, $\mathcal{S}_{t,j,2}$ and $\mathcal{S}_{t,j,3}$ to construct the estimates of $G(x_{t,j})$, $\nabla G(x_{t,j})$ and $\nabla F(g_{t,j})$, respectively. More specifically, at tth outer iteration, stochastic estimates are constructed through

$$\begin{aligned} g_{t,0} &= \frac{1}{B_{t,1}} \sum_{i=1}^{B_{t,1}} g(x_{t,0}; \phi_t^{(i)}), \\ \nabla g_{t,0} &= \frac{1}{B_{t,2}} \sum_{i=1}^{B_{t,2}} \nabla g(x_{t,0}; \hat{\phi}_t^{(i)}), \\ \nabla f_{t,0} &= \frac{1}{B_{t,3}} \sum_{i=1}^{B_{t,3}} \nabla f(g_{t,0}; \xi_t^{(i)}), \end{aligned} \tag{16}$$

where $B_{t,k} = |\mathcal{B}_{t,k}|$, $k = 1, 2.3$. Within the inner iterations, they are built in the following manner, that is for $j \in [\tau_t - 1]$,

$$g_{t,j} = g_{t,j-1} + \frac{1}{S_{t,1}} \sum_{i=1}^{S_{t,1}} \left(g\big(x_{t,j}; \phi_{t,j}^{(i)}\big) - g\big(x_{t,j-1}; \phi_{t,j}^{(i)}\big) \right), \tag{17}$$

$$\nabla g_{t,j} = \Pi_R \left[\nabla g_{t,j-1} + \frac{1}{S_{t,2}} \sum_{i=1}^{S_{t,1}} \left(\nabla g\big(x_{t,j}; \hat{\phi}_{t,j}^{(i)}\big) - \nabla g\big(x_{t,j-1}; \hat{\phi}_{t,j}^{(i)}\big) \right) \right], \tag{18}$$

$$\nabla f_{t,j} = \nabla f_{t,j-1} + \frac{1}{S_{t,3}} \sum_{i=1}^{S_{t,1}} \left(\nabla f\big(g_{t,j}; \xi_{t,j}^{(i)}\big) - \nabla f\big(g_{t,j-1}; \xi_{t,j}^{(i)}\big) \right), \tag{19}$$

where $S_{t,k} = |\mathcal{S}_{t,j,k}|$ for $k = 1, 2, 3$, and the projection operator Π_R for $R > 0$ is defined as $\Pi_R(x) = \arg\min_{\|z\| \leqslant R} \|z - x\|^2$. The projection operator plays a crucial role in bounding the stochastic Jacobian estimate $\nabla g_{t,j}$. Finally, the inner iteration ends after setting $x_{t+1,0} = x_{t,\tau_t}$.

To establish the sample complexity of SCO algorithms based on variance reduction techniques, existing analysis (such as those in [13, 4, 38, 21]) typically requires $g(x; \phi)$ satisfy uniform Lipschitzness, i.e., for any $x, y \in \mathbb{R}^d$, $\|g(x; \phi) - g(y; \phi)\| \leqslant \ell_g \|x - y\|$ for almost every ϕ, with $\ell_g > 0$. However, this assumption is not applicable to the problem (P) under heavy-tailed noise. Specifically, the uniform Lipschitzness and

almost-sure differentiability of $g(\cdot;\phi)$ implies $\|\nabla g(x;\phi)\| \leqslant \ell_g$ for any $x \in \mathbb{R}^d$. Then, for some $x \in \mathbb{R}^d$ satisfying $\|\nabla G(x)\| < +\infty$, there exists $a > p$ such that

$$\mathbb{E}[\|\nabla g(x;\phi) - \nabla G(x)\|^a] \leqslant \mathbb{E}[\|\nabla g(x;\phi)\|^a + \|\nabla G(x)\|^a] < +\infty,$$

which contradicts the Assumption 3 that only guarantees a finite pth moment. Therefore, a weaker assumption is needed for (P) under heavy-tailed noise.

Assumption 4 *The following conditions hold concerning problem* (P).

(a) (Mean-pth moment Lipschitzness of g) *There exists $\ell_g > 0$ such that*

$$\mathbb{E}[\|g(x;\phi) - g(y;\phi)\|^p] \leqslant \ell_g^p \|x - y\|^p, \quad \forall x, y \in \mathbb{R}^d.$$

(b) (Mean-pth moment smoothness of g) *There exists $L_g > 0$ such that*

$$\mathbb{E}[\|\nabla g(x;\phi) - \nabla g(y;\phi)\|^p] \leqslant L_g^p \|x - y\|^p, \quad \forall x, y \in \mathbb{R}^d.$$

(c) (Mean-pth moment smoothness of f) *There exists $L_f > 0$ such that*

$$\mathbb{E}[\|\nabla f(x;\xi) - \nabla f(y;\xi)\|^p] \leqslant L_f^p \|x - y\|^p, \quad \forall x, y \in \mathbb{R}^q.$$

Remark 1 *Note that Assumption 4 (a) combining the almost-sure differentiability of $g(\cdot;\phi)$ implies $\mathbb{E}[\|\nabla g(x;\phi)\|^p] \leqslant \ell_g^p$ for any $x \in \mathbb{R}^d$. It is noteworthy that, Assumption 4 (a) can be implied by the uniform Lipschitzness of $g(x;\phi)$. Moreover, it also recovers the condition of bounded second moment (i.e., $\mathbb{E}[\|\nabla g(x;\phi)\|^2] \leqslant \ell_g^2$), which is standard in the SCO literature [3, 10, 34]. When $p = 2$, Assumption 4 (a) reduces to the mean-squared Lipschitzness (such as in [17, 39]). Similarly, Assumption 4 (b) and (c) are relaxed conditions tailored for heavy-tailed noise settings. Similar conditions in the SO literature relate to the so-called "weakly average smoothness" [12, 9]. For $p = 2$, they recover the standard mean-squared smoothness as seen in [8, 6, 17, 39]. Therefore, when $1 < p < 2$, our Assumption 4 is strictly milder and better aligned with the behavior of heavy-tailed noise.*

Next, we redefine the relevant notation for NSCG-S. For any $t \in [T]$ and $j \in [\tau_t - 1]$, let

$$\begin{aligned} \Delta_{t,j} &:= \Psi(x_{t,j}) - \Psi^*, & \widehat{\nabla}_{t,j}^{f,g} &:= \nabla g_{t,j}^\top \nabla f_{t,j}, \\ \widehat{\nabla}_{t,j}^{g} &:= \nabla g_{t,j}^\top \nabla F(g_{t,j}), & \nabla_{t,j}^{g} &:= \nabla G(x_{t,j})^\top \nabla F(g_{t,j}), \end{aligned}$$

and filtration $\mathcal{F}_{t,j}$ be the σ-algebra generated by all random variables used in the first t outer iterations and the first j inner iterations. Let $\{(x_{t,j}, g_{t,j}, \nabla g_{t,j}, \nabla f_{t,j})\}$ be the sequence generated by NSCG-S. We state the averaged gradient norm estimate tailored for NSCG-S below, omitting its proof as it follows closely that of Lemma 5.

Lemma 6 *Suppose Assumptions 1 and 2 hold, the iterates generated by NSCG-S satisfies*

$$\begin{aligned} &\frac{1}{\tau_t T} \sum_{t=1}^{T} \sum_{j=0}^{\tau_t - 1} \mathbb{E}[\|\nabla \Psi(x_{t,j})\|] \\ &\leqslant \frac{2}{\tau_t T} \sum_{t=1}^{T} \sum_{j=0}^{\tau_t - 1} \mathbb{E}[\|\widehat{\nabla}_{t,j}^{f,g} - \widehat{\nabla}_{t,j}^{g}\|] + \frac{2\ell_F}{\tau_t T} \sum_{t=1}^{T} \sum_{j=0}^{\tau_t - 1} \mathbb{E}[\|\nabla g_{t,j} - \nabla G(x_{t,j})\|] \\ &\quad + \frac{2\ell_G L_F}{\tau_t T} \sum_{t=1}^{T} \sum_{j=0}^{\tau_t - 1} \mathbb{E}[\|g_{t,j} - G(x_{t,j})\|] + \frac{\Delta_{1,0}}{\alpha \tau_t T} + \frac{\alpha L}{2}, \end{aligned}$$

where $L := \ell_G^2 L_F + L_G \ell_F$.

The following lemma provides the upper bounds for the three error terms.

Lemma 7 *Suppose Assumptions 2, 3, and 4 hold, and $R \geqslant \ell_G$. Then for any $t \in [T]$ and $j \in [\tau_t - 1]$,*

$$\begin{aligned}
&\mathbb{E}\left[\|g_{t,j} - G(x_{t,j})\|\right] \leqslant 8\tau_t^{\frac{1}{p}}\alpha(\ell_g + \ell_G)S_{t,1}^{-\frac{p-1}{p}} + 2V_g B_{t,1}^{-\frac{p-1}{p}},\\
&\mathbb{E}[\|\nabla g_{t,j} - \nabla G(x_{t,j})\|] \leqslant 8\tau_t^{\frac{1}{p}}\alpha(L_g + L_G)S_{t,2}^{-\frac{p-1}{p}} + 2V_J B_{t,2}^{-\frac{p-1}{p}},\\
&\mathbb{E}[\|\widehat{\nabla}_{t,j}^{f,g} - \widehat{\nabla}_{t,j}^{g}\|] \leqslant R(8\tau_t^{\frac{1}{p}}\alpha(L_f + L_F)S_{t,3}^{-\frac{p-1}{p}} + 4(V_J + \ell_G)V_f B_{t,3}^{-\frac{p-1}{p}}).
\end{aligned}$$

Proof. Our proof focuses on deriving the upper bound for $\mathbb{E}\left[\|g_{t,j} - G(x_{t,j})\|\right]$. Bounds for the other two terms are obtained similarly with only minor modifications.

For $\mathbb{E}\left[\|g_{t,j} - G(x_{t,j})\|\right]$, we apply Lemma 4 obtaining

$$\begin{aligned}
&\mathbb{E}\left[\|g_{t,j} - G(x_{t,j})\|^p|\mathcal{F}_{t,j-1}\right]\\
&\leqslant 2^{2-p}\mathbb{E}\left[\|g_{t,j} - \mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right]\|^p|\mathcal{F}_{t,j-1}\right] + \mathbb{E}\left[\|\mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right] - G(x_{t,j})\|^p|\mathcal{F}_{t,j-1}\right]\\
&\quad + p\mathbb{E}\left[\frac{\langle g_{t,j} - \mathbb{E}[g_{t,j}|\mathcal{F}_{t,j-1}], \mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right] - G(x_{t,j})\rangle}{\|\mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right] - G(x_{t,j})\|^{2-p}}|\mathcal{F}_{t,j-1}\right]\\
&\leqslant 2^{2-p}\mathbb{E}\left[\|g_{t,j} - \mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right]\|^p|\mathcal{F}_{t,j-1}\right] + \mathbb{E}\left[\|\mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right] - G(x_{t,j})\|^p|\mathcal{F}_{t,j-1}\right]\\
&\quad + p\frac{\langle \mathbb{E}\left[g_{t,j} - \mathbb{E}[g_{t,j}|\mathcal{F}_{t,j-1}]|\mathcal{F}_{t,j-1}\right], \mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right] - G(x_{t,j})\rangle}{\|\mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right] - G(x_{t,j})\|^{2-p}}\\
&\leqslant 2^{2-p}\mathbb{E}\left[\|g_{t,j} - \mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right]\|^p|\mathcal{F}_{t,j-1}\right] + \mathbb{E}\left[\|\mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right] - G(x_{t,j})\|^p|\mathcal{F}_{t,j-1}\right],
\end{aligned}$$

where the second inequality follows from the $\mathcal{F}_{t,j-1}$-measurability of $G(x_{t,j})$ and $\mathbb{E}[g_{t,j}|\mathcal{F}_{t,j-1}]$, and the last inequality is due to $\mathbb{E}[g_{t,j} - \mathbb{E}[g_{t,j}|\mathcal{F}_{t,j-1}]|\mathcal{F}_{t,j-1}] = 0$. Taking total expectation on both sides of the above inequality yields

$$\begin{aligned}
\mathbb{E}\left[\|g_{t,j} - G(x_{t,j})\|^p\right] &\leqslant 2^{2-p}\mathbb{E}\left[\|g_{t,j} - \mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right]\|^p\right] + \mathbb{E}\left[\|\mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right] - G(x_{t,j})\|^p\right]\\
&\leqslant 2\underbrace{\mathbb{E}\left[\|g_{t,j} - \mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right]\|^p\right]}_{T_1} + \underbrace{\mathbb{E}\left[\|\mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right] - G(x_{t,j})\|^p\right]}_{T_2}. \qquad (20)
\end{aligned}$$

Below, we derive an upper bound for T_1 and the equivalent form of T_2.

Upper bound for T_1: By the definition of (17), it follows that

$$\begin{aligned}
g_{t,j} - \mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right] &= g_{t,j-1} + \frac{1}{S_{t,1}}\sum_{i=1}^{S_{t,1}}\left(g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right)\right)\\
&\quad - \mathbb{E}\left[g_{t,j-1} + \frac{1}{S_{t,1}}\sum_{i=1}^{S_{t,1}}\left(g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right)\right)|\mathcal{F}_{t,j-1}\right] \qquad (21)\\
&= \frac{1}{S_{t,1}}\sum_{i=1}^{S_{t,1}}\left(g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right) - G(x_{t,j}) + G(x_{t,j-1})\right).
\end{aligned}$$

We also apply Lemma 2 to bound the right side of above equality. We denote $M_{t,j} := g(x_{t,j};\phi_{t,j}^{(i)}) - g(x_{t,j-1};\phi_{t,j}^{(i)}) - G(x_{t,j}) + G(x_{t,j-1})$, and check that for any $j \in [\tau_t - 1]$,

$$\begin{aligned}
&\mathbb{E}\left[M_{t,j}|\mathcal{F}_{t,j-1}\right]\\
&= \mathbb{E}\left[g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right) - G(x_{t,j}) + G(x_{t,j-1})|\mathcal{F}_{t,j-1}\right]\\
&= \mathbb{E}\left[g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - G(x_{t,j})|x_{t,j}\right] + \mathbb{E}\left[G(x_{t,j-1}) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right)|x_{t,j-1}\right] = 0,
\end{aligned}$$

where the second inequality holds because $x_{t,j}$ and $x_{t,j-1}$ are $\mathcal{F}_{s-1}$-measurable and $\phi_{t,j}^{(i)}$ is independent of $\mathcal{F}_{t,j-1}$, and the last equality is due to the unbiasedness of function $g(\cdot;\phi)$. By following a reasoning analogous to Lemma 5 and invoking the bounded pth moment assumption with the tower property of expectation, we obtain

$$\begin{aligned}\mathbb{E}[\|M_{t,j}\|^p] &= \mathbb{E}\left[\left\|g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right) - G(x_t) + G(x_{t,j-1})\right\|^p\right] \\ &\leqslant 2\left(\mathbb{E}\left[\left\|g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - G(x_{t,j})\right\|^p\right] + \mathbb{E}\left[\left\|g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right) - G(x_{t,j-1})\right\|^p\right]\right) \\ &\leqslant 4V_g^p < +\infty.\end{aligned}$$

Hence, Lemma 2 gives

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{i=1}^{S_{t,1}}\left(g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right) - G(x_{t,j}) + G(x_{t,j-1})\right)\right\|^p\right] \\ &\leqslant 2\sum_{i=1}^{S_{t,1}}\mathbb{E}\left[\left\|g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right) - G(x_{t,j}) + G(x_{t,j-1})\right\|^p\right]\end{aligned}$$

Combining the fact that $\|a+b\|^p \leqslant 2\|a\|^p + 2\|b\|^p$, we further obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{i=1}^{S_{t,1}}\left(g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right) - G(x_{t,j}) + G(x_{t,j-1})\right)\right\|^p\right] \\ &\leqslant 4\sum_{i=1}^{S_{t,1}}\left(\mathbb{E}\left[\left\|g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right)\right\|^p\right] + \|G(x_{t,j}) - G(x_{t,j-1})\|^p\right) \\ &\leqslant 4S_{t,1}(\ell_g^p + \ell_G^p)\,\|x_{t,j} - x_{t,j-1}\|^p \\ &\leqslant 4\alpha^p S_{t,1}(\ell_g^p + \ell_G^p),\end{aligned} \tag{22}$$

where we apply Assumption 4 (a) and Assumption 2 in the second inequality. From (21), it follows that

$$\begin{aligned}T_1 &= \mathbb{E}\left[\|g_{t,j} - \mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right]\|^p\right] \\ &\leqslant \mathbb{E}\left[\left\|\frac{1}{S_{t,1}}\sum_{i=1}^{S_{t,1}}\left(g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right) - G(x_{t,j}) + G(x_{t,j-1})\right)\right\|^p\right] \\ &= \frac{1}{S_{t,1}^p}\mathbb{E}\left[\left\|\sum_{i=1}^{S_{t,1}}\left(g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right) - G(x_{t,j}) + G(x_{t,j-1})\right)\right\|^p\right] \\ &\overset{(22)}{\leqslant} 4\alpha^p(\ell_g^p + \ell_G^p)S_{t,1}^{1-p}.\end{aligned} \tag{23}$$

Equivalent form of T_2: Based on the fact that $g_{t,j-1}$ is a constant conditioning on $\mathcal{F}_{t,j-1}$ and the assumption of the unbiasedness for $g(x_{t,j};\phi_{t,j}^{(i)})$ and $g(x_{t,j-1};\phi_{t,j}^{(i)})$, we obtain

$$\begin{aligned}T_2 &= \mathbb{E}\left[\|\mathbb{E}\left[g_{t,j}|\mathcal{F}_{t,j-1}\right] - G(x_{t,j})\|^p\right] \\ &= \mathbb{E}\left[\left\|\mathbb{E}\left[g_{t,j-1} + \frac{1}{S_{t,1}}\sum_{i=1}^{S_{t,1}}\left(g\left(x_{t,j};\phi_{t,j}^{(i)}\right) - g\left(x_{t,j-1};\phi_{t,j}^{(i)}\right)\right)|\mathcal{F}_{t,j-1}\right] - G(x_{t,j})\right\|^p\right] \\ &= \mathbb{E}\left[\|g_{t,j-1} - G(x_{t,j-1})\|^p\right].\end{aligned}$$

Therefore, substituting the bounds obtained in (23) and (24) into (20) gives

$$\begin{aligned}\mathbb{E}\left[\|g_{t,j}-G(x_{t,j})\|^p\right] &\leqslant 8\alpha^p(\ell_g^p+\ell_G^p)S_{t,1}^{1-p}+\mathbb{E}\left[\|g_{t,j-1}-G(x_{t,j-1})\|^p\right]\\ &\leqslant 8\tau_t\alpha^p(\ell_g^p+\ell_G^p)S_{t,1}^{1-p}+\mathbb{E}\left[\|g_{t,0}-G(x_{t,0})\|^p\right].\end{aligned}$$

Applying Jensen's inequality and the fact that $(a+b)^{1/p}\leqslant a^{1/p}+b^{1/p}$ with $a,b\geqslant 0$ yields

$$\begin{aligned}\mathbb{E}\left[\|g_{t,j}-G(x_{t,j})\|\right] &\leqslant 8\tau_t^{\frac{1}{p}}\alpha(\ell_g+\ell_G)S_{t,1}^{-\frac{p-1}{p}}+\mathbb{E}\left[\|g_{t,0}-G(x_{t,0})\|\right]\\ &\leqslant 8\tau_t^{\frac{1}{p}}\alpha(\ell_g+\ell_G)S_{t,1}^{-\frac{p-1}{p}}+2V_gB_{t,1}^{-\frac{p-1}{p}},\end{aligned}$$

where $\mathbb{E}\left[\|g_{t,0}-G(x_{t,0})\|\right]\leqslant 2V_gB_{t,1}^{-\frac{p-1}{p}}$ is a direct result of Lemma 5. This completes the proof of the upper bound for $\mathbb{E}[\|g_{t,j}-G(x_{t,j})\|]$.

For $\mathbb{E}[\|\nabla g_{t,j}-\nabla G(x_{t,j})\|]$, let

$$v_{t,j}=\nabla g_{t,j-1}+\frac{1}{S_{t,2}}\sum_{i=1}^{S_{t,1}}(\nabla g(x_{t,j};\hat{\phi}_{t,j}^{(i)})-\nabla g(x_{t,j-1};\hat{\phi}_{t,j}^{(i)})),$$

which implies $\nabla g_{t,j}=\Pi_R[v_{t,j}]$. With $R\geqslant\ell_G$, we then apply the non-expansive property of the projection operator to get

$$\mathbb{E}[\|\nabla g_{t,j}-\nabla G(x_{t,j})\|]\leqslant\mathbb{E}[\|\Pi_R(v_{t,j})-\Pi_R(\nabla G(x_{t,j}))\|]\leqslant\mathbb{E}[\|v_{t,j}-\nabla G(x_{t,j})\|].$$

The desired result can be obtained by an entirely analogous proof process.

For $\mathbb{E}[\|\widehat{\nabla}_{t,j}^{f,g}-\widehat{\nabla}_{t,j}^{g}\|]$, it holds that

$$\mathbb{E}[\|\widehat{\nabla}_{t,j}^{f,g}-\widehat{\nabla}_{t,j}^{g}\|]\leqslant\mathbb{E}[\|\nabla g_{t,j}\|\|\nabla f_{t,j}-\nabla F(g_{t,j})\|]\leqslant R\mathbb{E}[\|\nabla f_{t,j}-\nabla F(g_{t,j})\|],$$

where the last inequality is because the projection operator maps any matrix into the norm ball of radius R. Note that $g_{t,j}$ depends on the newly sampled random variables $\{\phi_{t,j}^{(1)},\ldots,\phi_{t,j}^{(S_{t,1})}\}$, hence we need to enlarge the filtration $\mathcal{F}_t$ accordingly. Let $\widetilde{\mathcal{F}}_t:=\mathcal{F}_t\vee\sigma(\phi_{t,j}^{(1)},\ldots,\phi_{t,j}^{(S_{t,1})})$ be the smallest σ-algebra containing both $\mathcal{F}_t$ and $\sigma(\phi_{t,j}^{(1)},\ldots,\phi_{t,j}^{(S_{t,1})})$. A similar proof process then yields the desired result. □

Now, we are ready to present the main result concerns with NSCG-S method.

Theorem 3 (complexity with known p) *Suppose Assumptions 1, 2, 3, and 4 hold. For any given $T\in\mathbb{N}_+$, let $\tau_t=\tau=T$ and $\alpha=\frac{2}{T(L+2)}$. The estimates in (16) use the batch sizes $B_{t,1}=\lceil(b_1T)^{\frac{p}{p-1}}\rceil$ with $b_1=12\ell_GL_FV_g$, $B_{b,2}=\lceil(b_2T)^{\frac{p}{p-1}}\rceil$ with $b_2=12\ell_FV_J$, $B_{t,3}=\lceil(b_3T)^{\frac{p}{p-1}}\rceil$ with $b_3=24R(V_J+\ell_G)V_f$, and the estimates in (18), (18), and (19) use the batch sizes $S_{t,1}=\lceil(s_1\tau^{\frac{1}{p}})^{\frac{p}{p-1}}\rceil$ with $s_1=48\ell_GL_F(\ell_g+\ell_G)$, $S_{t,2}=\lceil(s_2\tau^{\frac{1}{p}})^{\frac{p}{p-1}}\rceil$ with $s_2=48\ell_F(L_g+L_G)$, and $S_{t,3}=\lceil(s_3\tau^{\frac{1}{p}})^{\frac{p}{p-1}}\rceil$ with $s_3=48R(L_f+L_F)$, respectively. Then, the iterates generated by NSCG-S satisfy*

$$\frac{1}{\tau_tT}\sum_{t=1}^{T}\sum_{j=0}^{\tau_t-1}\mathbb{E}[\|\nabla\Psi(x_{t,j})\|]=\mathcal{O}\left(T^{-1}\right),$$

and the sample complexity to find an ϵ-stationary point of problem (P) is in order $\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$.

Proof. Substituting the results in Lemma 7 into the Lemma 6 gives

$$\begin{aligned}
&\frac{1}{\tau_t T}\sum_{t=1}^{T}\sum_{j=0}^{\tau_t-1}\mathbb{E}[\|\nabla\Psi(x_{t,j})\|] \\
&\leqslant \frac{2R}{\tau_t T}\sum_{t=1}^{T}\sum_{j=0}^{\tau_t-1}\left(8\tau_t^{\frac{1}{p}}\alpha(L_f+L_F)S_{t,3}^{-\frac{p-1}{p}}+4(V_J+\ell_G)V_f B_{t,3}^{-\frac{p-1}{p}}\right) \\
&\quad+\frac{2\ell_F}{\tau_t T}\sum_{t=1}^{T}\sum_{j=0}^{\tau_t-1}\left(8\tau_t^{\frac{1}{p}}\alpha(L_g+L_G)S_{t,2}^{-\frac{p-1}{p}}+2V_J B_{t,2}^{-\frac{p-1}{p}}\right) \\
&\quad+\frac{2\ell_G L_F}{\tau_t T}\sum_{t=1}^{T}\sum_{j=0}^{\tau_t-1}\left(8\tau_t^{\frac{1}{p}}\alpha(\ell_g+\ell_G)S_{t,1}^{-\frac{p-1}{p}}+2V_g B_{t,1}^{-\frac{p-1}{p}}\right)+\frac{\Delta_{1,0}}{\alpha\tau_t T}+\frac{\alpha L}{2}.
\end{aligned} \tag{24}$$

Noting that τ_t, $B_{t,1}$, $B_{t,2}$, $B_{t,3}$, $S_{t,1}$, $S_{t,2}$, and $S_{t,3}$ are independent of t, we then obtain

$$\begin{aligned}
\frac{1}{\tau_t T}\sum_{t=1}^{T}\sum_{j=0}^{\tau_t-1}\mathbb{E}[\|\nabla\Psi(x_{t,j})\|]\leqslant&\frac{8R(V_J+\ell_G)V_f}{b_3 T}+\frac{4\ell_F V_J}{b_2 T}+\frac{4\ell_G L_F V_g}{b_1 T}+\frac{16\alpha R(L_f+L_F)}{s_3} \\
&+\frac{16\alpha\ell_F(L_g+L_G)}{s_2}+\frac{16\alpha\ell_G L_F(\ell_g+\ell_G)}{s_1}+\frac{\Delta_{1,0}}{\alpha\tau T}+\frac{\alpha L}{2}.
\end{aligned}$$

By the choice of parameters $b_1=12\ell_G L_F V_g$, $b_2=12\ell_F V_J$, $b_3=24R(V_J+\ell_G)V_f$, $s_1=48\ell_G L_F(\ell_g+\ell_G)$, $s_2=48\ell_F(L_g+L_G)$, $s_3=48R(L_f+L_F)$, and $\tau=T$, we further have

$$\frac{1}{\tau_t T}\sum_{t=1}^{T}\sum_{j=0}^{\tau_t-1}\mathbb{E}[\|\nabla\Psi(x_{t,j})\|]\leqslant\frac{\Delta_{1,0}}{\alpha T^2}+\frac{1}{T}+\frac{\alpha(L+2)}{2}$$

According to $\alpha=\frac{2}{T(L+2)}$, it holds that

$$\frac{1}{\tau_t T}\sum_{t=1}^{T}\sum_{j=0}^{\tau_t-1}\mathbb{E}[\|\nabla\Psi(x_{t,j})\|]\leqslant\frac{\Delta_{1,0}(L+2)}{2T}+\frac{2}{T}=\mathcal{O}\left(T^{-1}\right).$$

Note that NSCG-S is a double-loop method whose inner and outer iterations require different sample sizes. Hence, to find an ϵ-stationary point, the total number of samples required by NSCG-S with setting $T=\mathcal{O}(\epsilon^{-1})$ is

$$\begin{aligned}
N&=\sum_{t=1}^{T}(B_{t,1}+B_{t,2}+B_{t,3}+\tau_t(S_{t,1}+S_{t,2}+S_{t,3})) \\
&=\mathcal{O}\left(\left((b_1)^{\frac{p}{p-1}}+(b_2)^{\frac{p}{p-1}}+(b_3)^{\frac{p}{p-1}}\right)\epsilon^{-1-\frac{p}{p-1}}+\left((s_1)^{\frac{p}{p-1}}+(s_2)^{\frac{p}{p-1}}+(s_3)^{\frac{p}{p-1}}\right)\epsilon^{-1-1-\frac{1}{p-1}}\right) \\
&=\mathcal{O}\left(\epsilon^{-\frac{2p-1}{p-1}}\right).
\end{aligned}$$

This completes the proof. □

It can be observed that NSCG-S achieves a lower sample complexity order than NSCG-M. Moreover, NSCG-S attains the same order of the sample complexity as the algorithms with variance reduction strategy for the non-convex SO problem (2), such as NSGD-VR [30] and NSGD-M [12]. In the finite-variance case (i.e., $p=2$), NSCG-S recovers the optimal sample complexity $\mathcal{O}(\epsilon^{-3})$ well-established for variance-reduced SCO [37, 4].

The following theorem also provides the convergence results of NSCG-S without the prior knowledge of p.

Theorem 4 (complexity with unknown p) *Suppose Assumptions 1, 2, 3, and 4 hold. For any given $T \in \mathbb{N}_+$, let $\tau_t = \tau = T$ and $\alpha = T^{-1}$. The estimates in (16) use the batch sizes $B_{t,1} = \lceil (b_1 T)^2 \rceil$ with $b_1 = 4\ell_G L_F V_g$, $B_{b,2} = \lceil (b_2 T)^2 \rceil$ with $b_2 = 4\ell_F V_J$, $B_{t,3} = \lceil (b_3 T)^2 \rceil$ with $b_3 = 8R(V_J + \ell_G)V_f$, and the estimates in (18), (18), and (19) use the batch sizes $S_{t,1} = \lceil s_1 \tau \rceil$ with $s_1 = (16\ell_G L_F(\ell_g + \ell_G))^2$, $S_{t,2} = \lceil s_2 \tau \rceil$ with $s_2 = (16\ell_F(L_g + L_G))^2$, and $S_{t,3} = \lceil s_3 \tau \rceil$ with $s_3 = (16R(L_f + L_F))^2$, respectively. Then, the iterates generated by NSCG-S satisfy*

$$\frac{1}{\tau_t T} \sum_{t=1}^{T} \sum_{j=0}^{\tau_t - 1} \mathbb{E}[\|\nabla \Psi(x_{t,j})\|] = \mathcal{O}\left(T^{-\frac{2p-2}{p}} + T^{-1}\right) = \mathcal{O}\left(T^{-\frac{2p-2}{p}}\right),$$

and the sample complexity to find an ϵ-stationary point of problem (P) is in order $\mathcal{O}(\epsilon^{-\frac{3p}{2p-2}})$.

Proof. Since all parameters τ_t, $B_{t,1}$, $B_{t,2}$, $B_{t,3}$, $S_{t,1}$, $S_{t,2}$, and $S_{t,3}$ are also independent of t in this theorem, substituting them into (24) yields analogously

$$\begin{aligned}\frac{1}{\tau_t T} \sum_{t=1}^{T} \sum_{j=0}^{\tau_t - 1} \mathbb{E}[\|\nabla \Psi(x_{t,j})\|] \leqslant & \frac{8R(V_J + \ell_G)V_f}{(b_3 T)^{\frac{2p-2}{p}}} + \frac{4\ell_F V_J}{(b_2 T)^{\frac{2p-2}{p}}} + \frac{4\ell_G L_F V_g}{(b_1 T)^{\frac{2p-2}{p}}} + \frac{16\tau^{\frac{1}{p}} \alpha R(L_f + L_F)}{(s_3 \tau)^{\frac{p-1}{p}}} \\ & + \frac{16\tau^{\frac{1}{p}} \alpha \ell_F (L_g + L_G)}{(s_2 \tau)^{\frac{p-1}{p}}} + \frac{16\tau^{\frac{1}{p}} \alpha \ell_G L_F (\ell_g + \ell_G)}{(s_1 \tau)^{\frac{p-1}{p}}} + \frac{\Delta_{1,0}}{\alpha \tau T} + \frac{\alpha L}{2}.\end{aligned}$$

With the selected values of b_1, b_2, b_3, s_1, s_2, and s_3, it follows that

$$\begin{aligned}\frac{1}{\tau_t T} \sum_{t=1}^{T} \sum_{j=0}^{\tau_t - 1} \mathbb{E}[\|\nabla \Psi(x_{t,j})\|] \leqslant & \frac{b_1^{\frac{2-p}{p}} + b_2^{\frac{2-p}{p}} + b_3^{\frac{2-p}{p}}}{T^{\frac{2p-2}{p}}} + \alpha \left(s_1^{\frac{2-p}{p}} + s_2^{\frac{2-p}{p}} + s_3^{\frac{2-p}{p}}\right) \tau^{\frac{2-p}{p}} \\ & + \frac{\Delta_{1,0}}{\alpha \tau T} + \frac{\alpha L}{2}.\end{aligned}$$

Setting $\tau = T$ and $\alpha = T^{-1}$ gives

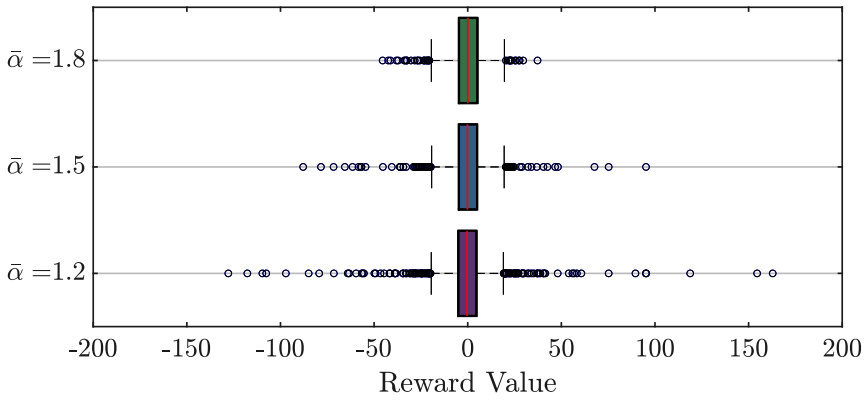
$$\begin{aligned}\frac{1}{\tau_t T} \sum_{t=1}^{T} \sum_{j=0}^{\tau_t - 1} \mathbb{E}[\|\nabla \Psi(x_{t,j})\|] \leqslant & \frac{b_1^{\frac{2-p}{p}} + b_2^{\frac{2-p}{p}} + b_3^{\frac{2-p}{p}} + s_1^{\frac{2-p}{p}} + s_2^{\frac{2-p}{p}} + s_3^{\frac{2-p}{p}}}{T^{\frac{2p-2}{p}}} + \frac{\Delta_{1,0} + L/2}{T} \\ = & \mathcal{O}\left(T^{-\frac{2p-2}{p}} + T^{-1}\right) = \mathcal{O}(T^{-\frac{2p-2}{p}}).\end{aligned}$$

By setting $T = \mathcal{O}(\epsilon^{-\frac{p}{2p-2}})$, the sample complexity to find an ϵ-stationary point is

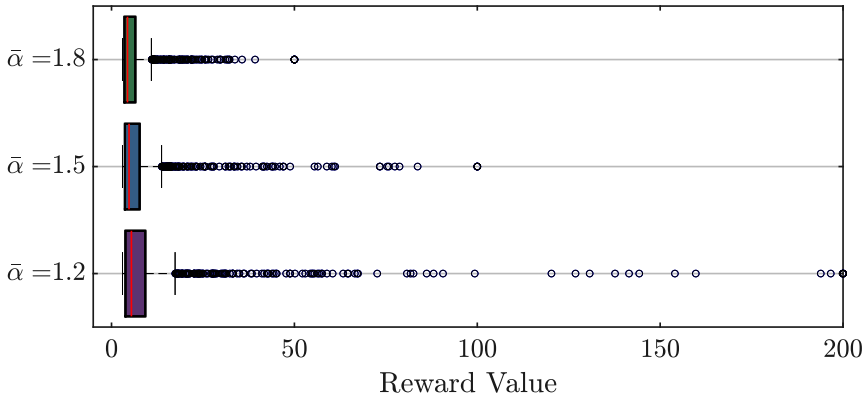
$$\begin{aligned}N &= \mathcal{O}\left(\left((b_1)^2 + (b_2)^2 + (b_3)^2\right)\left(\epsilon^{-\frac{p}{2p-2}}\right)^{1+2} + (s_1 + s_2 + s_3)\left(\epsilon^{-\frac{p}{2p-2}}\right)^{1+1+1}\right) \\ &= \mathcal{O}\left(\epsilon^{-\frac{3p}{2p-2}}\right).\end{aligned}$$

This completes the proof. □

In the absence of prior knowledge of p, NSCG-S can still achieve a sample complexity superior to that established in Theorem 2 and improves the optimal result for the SO methods up to a logarithmic factor [12], i.e., $\mathcal{O}((\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))^{\frac{3p}{2p-2}})$. Moreover, NSCG-S continues to recover the optimal sample complexity $\mathcal{O}(\epsilon^{-3})$ when $p = 2$.



(a) SαS distribution



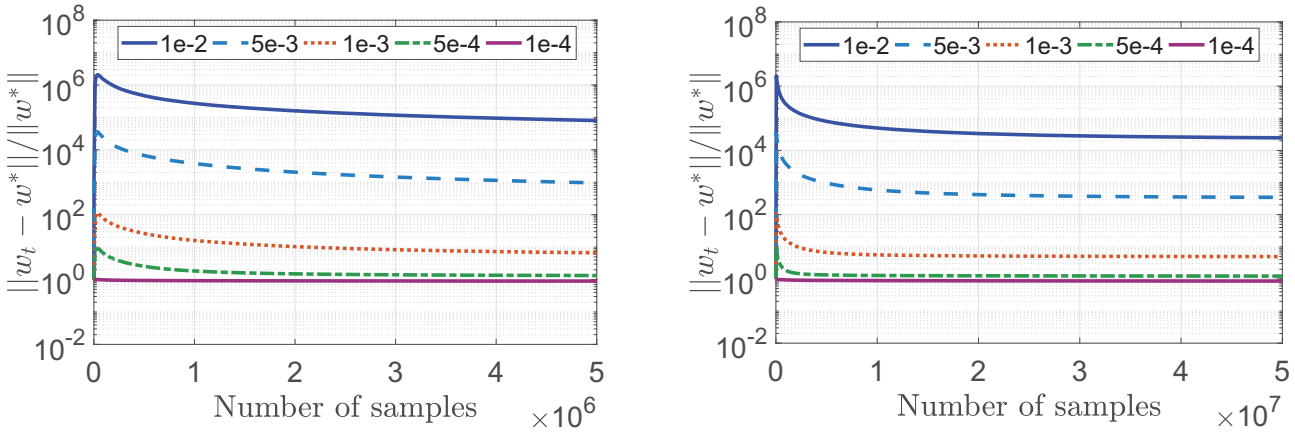
(b) Pareto distribution

Figure 1: The SαS and Pareto distributions with various tail indices.

5 Numerical experiments

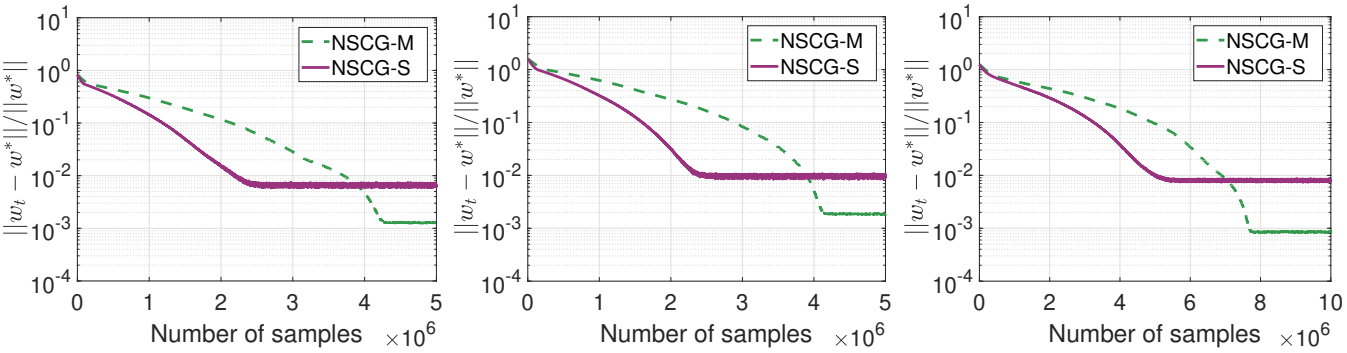
In this section, we apply our NSCG methods to a policy evaluation task following the experimental setup in [32, Section 4]. Recalling Example 1.1 in Introduction, we consider a MDP denoted as a tuple $M = (\mathcal{S}, \mathcal{A}, R, P, \bar{\gamma})$. The objective is to evaluate the value function $v^\pi \in \mathbb{R}^d$ associated with a policy π, where each component $v^\pi(s)$ represents the expected cumulative return starting from state s. Let r_{s_1,s_2} denote the reward when transitioning from state s_1 to s_2. The value function satisfies the Bellman equation $v^\pi(s) = \mathbb{E}_\pi[r_{s_1,s_2} + \bar{\gamma} v^\pi(s_2)|s_1]$ for all $s_1, s_2 \in \{1, \ldots, S\}$. The solution of this equation, denoted v^*, satisfies $v^* = v^\pi$. In this experiment, we adopt a tabular representation to construct the linear mapping $\varphi_s \in \mathbb{R}^S$ for the feature of $v^\pi(s)$. The tabular representation encodes each state as a one-hot vector, where a value of one appears exclusively at the dimension corresponding to the current state index [1]. Formally, the value function $v^\pi(s)$ is approximated by $v^\pi(s) \approx \varphi_s^\top w^*$ for some $w^* \in \mathbb{R}^S$. As illustrated in [32, Section 4], this problem can be formulated as a Bellman residual minimization problem, i.e.,

$$\min_{w \in \mathbb{R}^S} \sum_{s=1}^{S} (\varphi_s^\top w - q_s^\pi(w))^2,$$



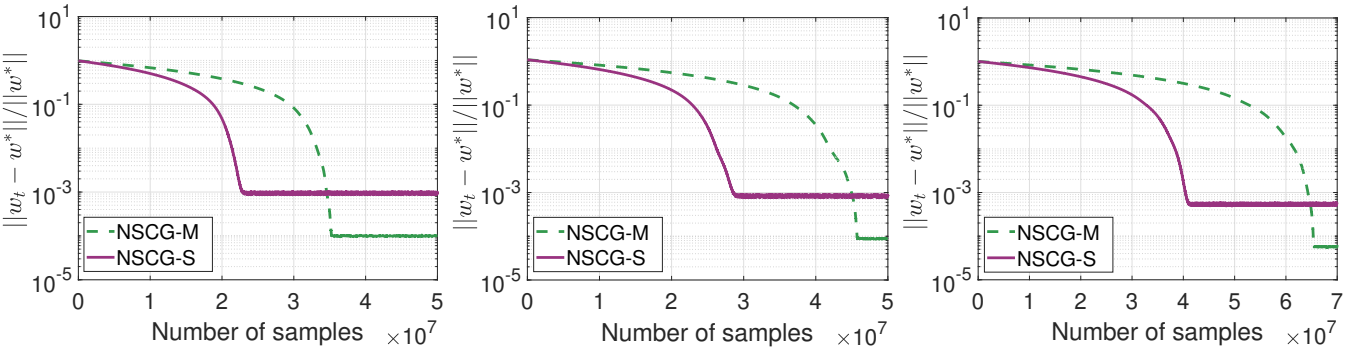
(a) SαS distribution with $\bar{\alpha} = 1.8$ (b) Pareto distribution with $\bar{\alpha} = 1.8$

Figure 2: Convergence performance of the ASC-PG method with different step sizes.



(a) $\bar{\alpha} = 1.8$ (b) $\bar{\alpha} = 1.5$ (c) $\bar{\alpha} = 1.2$

Figure 3: Convergence performance of the NSCG methods under SαS distribution.



(a) $\bar{\alpha} = 1.8$ (b) $\bar{\alpha} = 1.5$ (c) $\bar{\alpha} = 1.2$

Figure 4: Convergence performance of the NSCG methods under Pareto distribution.

where $q_s^\pi(w) = \mathbb{E}_\pi[r_{s,s'} + \bar{\gamma}\varphi_{s'}^\top w] = \sum_{s'} P_{s,s'}^\pi(\{r_{s,s'} + \bar{\gamma}\varphi_{s'}^\top w\})$. This problem can be formulated into problem (P) with

$$G(w) = (\varphi_1^\top w, q_1^\pi(w), \ldots, \varphi_S^\top w, q_S^\pi(w)) \in \mathbb{R}^{2S}, \qquad f(G(w)) = \sum_{s=1}^{S} (\varphi_s^\top w - q_s^\pi(w))^2 \in \mathbb{R}.$$

In our simulation, each MDP instance consists of $S = 100$ states and three actions per state. For any given state-action pair, the agent transitions to one of four possible subsequent states. The transition probabilities are sampled uniformly from the interval $[0, 1]$ and then normalized. For the reward of each transition, unlike in [32] where it is uniformly generated from $[0, 1]$, we consider two heavy-tailed rewards drawn from either a Symmetric α-Stable (SαS) or Pareto distribution. The characteristic function of the SαS distribution is defined as $\varphi(t) = \exp(-c|t|^{\bar{\alpha}})$, where $\bar{\alpha} \in (1, 2]$ denotes the tail index and c is a scale parameter. The Pareto distribution exhibits the probability density function $f(x) = \bar{\alpha} x_m^{\bar{\alpha}} / x^{\bar{\alpha}+1}$ with $x \geqslant x_m$, where x_m denotes the threshold parameter and we reuse the notation $\bar{\alpha}$ to denote the tail index of the Pareto distribution for convenience. When the tail index $\bar{\alpha}$ fails within $(1, 2)$, both distributions exhibit heavy-tailed behavior, possessing only finite pth moment for $p < \bar{\alpha}$. With parameters set as $c = 5$, $x_m = 3$ and $\bar{\alpha} \in \{1.8, 1.5, 1.2\}$, these two distributions are visualized in Fig. 1. For visual clarity, the reward values are truncated within $[-50, 50]$, $[-100, 100]$, and $[-200, 200]$ for $\bar{\alpha} = 1.8$, 1.5, and 1.2, respectively, preserving at least 98% of the data. It is evident that the SαS and Pareto distributions exhibit a significant number of outliers beyond the upper and lower bounds. As $\bar{\alpha}$ decreases, both distributions display more frequent extreme values, indicating more severe heavy-tailed behavior.

We first verify that the accelerated stochastic compositional proximal gradient (ASC-PG) method from [32] fails under heavy-tailed environments, as demonstrated in Fig. 2. The step size is set as $\alpha = \min\{\eta, t^{-1}\}$, where $\eta \in \{10^{-2}, 5\times10^{-3}, 10^{-3}, 5\times10^{-4}, 10^{-4}\}$ and t denotes the iteration step. This modification is necessary because the original schedule (i.e., t^{-1}) directly causes divergence in $\|w_t - w^*\|/\|w^*\|$. The results demonstrate that ASC-PG method fails to converge under these step sizes when $\bar{\alpha} = 1.8$, not to mention the lower $\bar{\alpha}$. Moreover, larger step sizes further degrade convergence performances, while smaller step sizes do not lead to successful convergence.

We proceed to evaluate the two NSCG methods with their step sizes tuned over the set $\{5\times10^{-2}, 10^{-2}, 5\times10^{-3}, \ldots, 10^{-5}\}$. For NSCG-M, the number of batch sizes is set to 50. To ensure a fair comparison, we align the sample cost of iteration t across algorithms. For NSCG-M, the batch size is fixed at 50. For NSCG-S, we set the inner iteration count to 4 with and inner/outer batch sizes to 5 and 20, respectively, resulting in the same per-iteration cost of 50 samples. The convergence performance of the NSCG methods under SαS distribution are shown in Fig. 3, which plots the average of 20 independent runs. It is evident that NSCG-S requires significantly fewer samples to converge than NSCG-M. The log-scale plots demonstrate that NSCG-M achieves higher convergence accuracy, which benefits from its larger batch sizes per iteration. In the more severe heavy-tailed case ($\bar{\alpha} = 1.2$), convergence demands a larger number of samples. Similar convergence behavior is observed under the Pareto distribution with the average of 20 independent runs, as shown in Fig. 4. Due to the larger initial bias, achieving convergence in these cases requires a greater number of samples. These observations are consistent with our theoretical insights.

6 Conclusions

In this paper, we have presented a generic framework of normalized stochastic compositional gradient methods for non-convex SCO under heavy-tailed noise. Two concrete methods are derived within this framework: NSCG-M and NSCG-S, each employing a distinct stochastic estimator to reliably approximate the inner function, its Jacobian, and the outer gradient in the presence of heavy-tailed noise. We established the sample complexity of both methods for finding an ϵ-stationary point under certain conditions, with

and without knowledge of p. These results align with the best-known complexity of first-order methods for single-level stochastic optimization algorithms under heavy-tailed noise and, for $p = 2$, recover the optimal complexity for the stochastic composite optimization algorithms. Numerical experiments on a policy-evaluation task with heavy-tailed rewards confirm the practical efficacy of the proposed methods.

Acknowledgments

The authors would like to thank Dr. Luo Luo for his insightful comments and suggestions on the early version of this manuscript.

References

[1] Adam Adam and Martha White. Investigating practical linear temporal difference learning. In *International Conference on Autonomous Agents & Multiagent Systems*, pages 494–502, 2016.

[2] Amrit Singh Bedi, Anjaly Parayil, Junyu Zhang, Mengdi Wang, and Alec Koppel. On the sample complexity and metastability of heavy-tailed policy search in continuous control. *Journal of Machine Learning Research*, 25(39):1–58, 2024.

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

[4] Ziyi Chen and Yi Zhou. Momentum with variance reduction for nonconvex composition optimization. *arXiv preprint arXiv:2005.07755*, 2020.

[5] Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems*, volume 34, pages 4883–4895, 2021.

[6] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[7] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[8] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[9] Adrien Fradin, Abdurakhmon Sadiev, Laurent Condat, and Peter Richtárik. Tight lower bounds and optimal algorithms for stochastic nonconvex optimization with heavy-tailed noise. *arXiv preprint arXiv:2512.18713*, 2025.

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

[11] Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. High-probability complexity bounds for non-smooth stochastic convex optimization with heavy-tailed noise. *Journal of Optimization Theory and Applications*, 203(3):2679–2738, 2024.

[12] Chuan He, Zhaosong Lu, Defeng Sun, and Zhanwang Deng. Complexity of normalized stochastic first-order methods with momentum under heavy-tailed noise. *arXiv preprint arXiv:2506.11214*, 2025.

[13] Wenqing Hu, Chris Junchi Li, Xiangru Lian, Ji Liu, and Huizhuo Yuan. Efficient smooth non-convex stochastic compositional optimization via stochastic recursive gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[14] Jiayi Huang, Han Zhong, Liwei Wang, and Lin Yang. Tackling heavy-tailed rewards in reinforcement learning with function approximation: Minimax optimal and instance-dependent regret bounds. In *Advances in Neural Information Processing Systems*, volume 36, pages 56576–56588, 2023.

[15] Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *International Conference on Artificial Intelligence and Statistics*, volume 258, 2025.

[16] Guanghui Lan Ilyas Fatkhullin, Florian Hübler. Can SGD handle heavy-tailed noise? *arXiv preprint arXiv:2508.04860*, 2025.

[17] Wei Jiang, Sifan Yang, Wenhao Yang, Yibo Wang, Yuanyu Wan, and Lijun Zhang. Projection-free variance reduction methods for stochastic constrained multi-level compositional optimization. *arXiv preprint arXiv:2406.03787*, 2024.

[18] Nikita Kornilov, Ohad Shamir, Aleksandr Lobanov, Darina Dvinskikh, Alexander Gasnikov, Innokentiy Shibaev, Eduard Gorbunov, and Samuel Horváth. Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance. In *Advances in Neural Information Processing Systems*, volume 36, pages 64083–64102, 2023.

[19] Pavlo Krokhmal, Jonas Palmquist, and Stanislav Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk*, 4:43–68, 2002.

[20] Shaojie Li and Yong Liu. High probability analysis for non-convex stochastic optimization with clipping. *arXiv preprint arXiv:2307.13680*, 2023.

[21] Jin Liu, Xiaokang Pan, Junwen Duan, Hong-Dong Li, Youqi Li, and Zhe Qu. Faster stochastic variance reduction methods for compositional minimax optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13927–13935, 2024.

[22] Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *Annual Conference on Learning Theory*, pages 2266–2290, 2023.

[23] Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. *arXiv preprint arXiv:2412.19529*, 2024.

[24] Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems*, volume 36, pages 24191–24222, 2023.

[25] Daniela Angela Parletta, Andrea Paudice, Massimiliano Pontil, and Saverio Salzo. High probability bounds for stochastic subgradient schemes with heavy tailed noise. *SIAM Journal on Mathematics of Data Science*, 6(4):953–977, 2024.

[26] LA Prashanth, Krishna Jagannathan, and Ravi Kumar Kolla. Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions. In *International Conference on Machine Learning*, pages 5577–5586, 2020.

[27] R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

[28] Anton Rodomanov and Yurii Nesterov. Smoothness parameter of power of euclidean norm. *Journal of Optimization Theory and Applications*, 185(2):303–326, 2020.

[29] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, pages 29563–29648, 2023.

[30] Tao Sun, Xinwang Liu, and Kun Yuan. Gradient normalization provably benefits nonconvex SGD under heavy-tailed noise. *arXiv preprint arXiv:2410.16561*, 2024.

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

[32] Mengdi Wang, Ji Liu, and Ethan X Fang. Accelerating stochastic composition optimization. *Journal of Machine Learning Research*, 18(105):1–23, 2017.

[33] Shuoguang Yang, Ethan X Fang, and Uday V Shanbhag. Data-driven compositional optimization in misspecified regimes. *Operations Research*, 73(3):1395–1411, 2025.

[34] Shuoguang Yang, Wei You, Zhe Zhang, and Ethan X Fang. Stochastic compositional optimization with compositional constraints. *INFORMS Journal on Optimization*, 2025.

[35] Yuchen Yang, Kaihong Lu, and Long Wang. Online distributed optimization with clipped stochastic gradients: High probability bound of regrets. *Automatica*, 182:112525, 2025.

[36] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, volume 33, pages 15383–15393, 2020.

[37] Junyu Zhang and Lin Xiao. A stochastic composite gradient method with incremental variance reduction. In *Advances in Neural Information Processing Systems*, volume 32, pages 9075–9085, 2019.

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

[39] Junyu Zhang and Lin Xiao. Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization. *Mathematical Programming*, 195(1):649–691, 2022.

[40] Vincent Zhuang and Yanan Sui. No-regret reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pages 3385–3393, 2021.