Normalized stochastic proximal approximation methods for nonsmooth composite optimization under heavy-tailed noise

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

May 21, 2026

Abstract

In this paper, we study nonsmooth composite optimization problems under heavy-tailed noise, with the objective being a summation of a nested function and a nonsmooth convex regularizer. We propose stochastic proximal approximation methods incorporating a normalization technique to handle the potential challenges caused by the nonsmooth regularizer and heavy-tailed noise. For the case where the outer function of the nested structure is smooth, our proposed algorithms achieve sample complexities that match the best known results for single-layer nonconvex stochastic optimization under heavy-tailed noise. For the case where the outer function is convex but nonsmooth, to the best of our knowledge, the corresponding normalized stochastic proximal gradient methods are new, with sample complexity bounds provided. For each of the above proposed algorithms, we consider two variants with constant parameter sequences and decaying ones, respectively. The effectiveness of the proposed methods is validated through numerical experiments on sparse phase retrieval problem and policy evaluation for Markov decision processes.

Keywords: Heavy-tailed Noise, Stochastic Composite Optimization, Nonsmooth Optimization, Normalization, Variance Reduction

1 Introduction

In this work, we consider the following nonsmooth composite optimization (NCO):

$$\min_{x\in\mathbb{R}^n} \Psi(x) := f(G(x)) + r(x) = f(\mathbb{E}_{\xi\sim\Xi}[g(x;\xi)]) + r(x), \tag{P}$$

where $G:\mathbb{R}^n\to\mathbb{R}^m$ with $G(x)=\mathbb{E}_{\xi\sim\Xi}[g(x;\xi)]$ is a smooth vector mapping for *a.e.* ξ; $r:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ is a closed, convex and nonsmooth function. For $f:\mathbb{R}^m\to\mathbb{R}$, we study two cases.

- Case I: f is smooth and possibly nonconvex;
- Case II: f is nonsmooth and convex.

We focus on the setting in which samples ξ are available, and stochastic oracle returns $g(x;\xi)$ and $\nabla g(x;\xi)$ as unbiased stochastic estimates of $G(x)$ and its Jacobian $\nabla G(x)$, respectively.

The NCO problem (P) serves as a general mathematical framework for a wide array of applications. For instance, in sparse phase retrieval, one can only measure the intensity (i.e., the squared magnitude)

*Southern University of Science and Technology, 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).

of linear projections, rather than the linear measurements. The mainstream approach minimizes a loss function that measures the discrepancy between predicted intensities and the observations, regularized by a nonsmooth penalty [4, 5, 47]. This leads to a composite optimization problem that takes the exact form of NCO problem (P). Further examples include policy evaluation in reinforcement learning [12, 46], robust blind deconvolution [7], and risk-averse mean-variance optimization [38, 39], all of which naturally align with the NCO framework.

Classical methods for solving the NCO problem have been extensively studied under the assumption that stochastic estimate errors have bounded variances, i.e.

$$\mathbb{E}[\|g(x;\xi) - G(x)\|^2] \leqslant V_g^2, \qquad \mathbb{E}[\|\nabla g(x;\xi) - \nabla G(x)\|^2] \leqslant V_J^2, \quad \forall x \in \mathbb{R}^n \tag{1}$$

for some positive constants V_g and V_J. In [50], a stochastic composite gradient method was proposed for NCO in Case I, achieving a sample complexity of $\mathcal{O}(\epsilon^{-3})$ for finding an ϵ-stationary point[1]. An enhanced version incorporating a momentum update strategy was developed in [9], attaining the same order of sample complexity. Subsequently, a framework of normalized proximal approximate gradient methods was designed in [51] to solve the multilevel composite stochastic optimization problem with smooth nested mapping. The variant employing mini-batch estimator yields a sample complexity of $\mathcal{O}(\epsilon^{-4})$, whereas the version equipped with the SPIDER estimator (a variance reduction technique [15]) exhibits an improved complexity $\mathcal{O}(\epsilon^{-3})$. For the NCO problem in Case II, significant progress has been made through stochastic *prox-linear* algorithms [52]. Using a mini-batch estimator, this algorithm yields sample complexities of $\mathcal{O}(\epsilon^{-6})$ and $\mathcal{O}(\epsilon^{-4})$ for the inner function and its Jacobian, respectively. These bounds were further improved to $\mathcal{O}(\epsilon^{-5})$ and $\mathcal{O}(\epsilon^{-3})$ by incorporating the SPIDER estimator. Meanwhile, two stochastic Gauss-Newton algorithms were developed in [43], establishing matching sample complexity bounds under the same settings. In more general settings, Wang et al. investigated the doubly stochastic NCO problem and established a sample complexity of $\mathcal{O}(\epsilon^{-4.5})$ [46] . This framework was further extended to include inequality constraints in [26]. Additionally, algorithms for solving doubly/multiply stochastic NCO problems with $r(x) \equiv 0$ have also been extensively studied [45, 48, 44, 8, 1, 19, 29, 30, 25, 53].

Notably, the theoretical guarantees of the aforementioned algorithms rely heavily on the assumption (1). However, recent empirical evidence suggests that due to factors such as data distribution and environmental noise, the estimation noise frequently exhibits heavy-tailed characteristics. For example, in high-energy coherent X-ray imaging, impulse disturbances induced by X-ray radiation on charge-coupled devices lead to heavy-tailed noise in the measurements [18]. Moreover, the rapid advancement of large language models, characterized by massive data and deep architectures, has fundamentally altered the statistical properties of stochastic approximations, causing gradient noise to exhibit heavy-tailed behavior [40, 41, 49]. Mathematically, the stochastic estimates under heavy-tailed noise are assumed to be unbiased and to have bounded p-th central moments with $1 < p \leqslant 2$ [49, 21]. Specifically, these conditions are summarized in the following assumption.

Assumption 1 *For any $x \in \mathbb{R}^n$, there exist constants $V_g, V_J > 0$, and $p \in (1, 2]$ such that*

$$\begin{aligned} &\mathbb{E}[g(x;\xi)] = G(x), && \mathbb{E}[\|g(x;\xi) - G(x)\|^p] \leqslant V_g^p, \\ &\mathbb{E}[\nabla g(x;\xi)] = \nabla G(x), && \mathbb{E}[\|\nabla g(x;\xi) - \nabla G(x)\|^p] \leqslant V_J^p. \end{aligned}$$

This assumption relaxes the cornerstone condition of bounded variance for stochastic noise in NCO. It represents a weaker condition, reducing to the bounded variance assumption when $p = 2$ (under which the noise can still be non-sub-Gaussian and thus heavy-tailed). The extreme outliers generated in this heavy-tailed regime can severely disrupt the convergence behavior of the aforementioned NCO algorithms and may even lead to divergence.

[1]There exists a point $\bar{x} \in \mathbb{R}^n$ satisfying $\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \epsilon$, where $\epsilon > 0$ and $\mathcal{G}(\bar{x})$ is the proximal-gradient mapping defined in (7).

1.1 Related Work

The study of stochastic optimization under heavy-tailed noise has gained interest in recent years, yet it remains largely concentrated on single-layer stochastic optimization problems, i.e., $\min_x G(x) := \mathbb{E}[g(x;\xi)]$. In nonconvex and smooth settings, gradient estimates are frequently contaminated by extreme outliers arising from heavy-tailed noise. To address this issue, gradient clipping (a technique that truncates large gradient estimates at a constant threshold) was introduced to improve the stochastic gradient descent (SGD) algorithm, achieving a sample complexity of $\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$ for finding an ϵ-stationary point[2] [49, 35, 20, 28]. Alternatively, gradient normalization, defined as $\nabla g(x;\xi)/\|\nabla g(x;\xi)\|$, provides an adaptive mechanism for mitigating extreme gradients. A variant of SGD incorporating both gradient clipping and normalization along with momentum was proposed in [10], yielding the same order of complexity. Subsequently, the STORM estimator (a variance reduction technique [11]) was introduced to achieve an improved sample complexity of $\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$ under the uniform Lipschitz continuity of the stochastic gradient [32]. Sun et al. examined the specific roles of clipping and normalization in SGD's convergence, developing effective algorithms that operate without gradient clipping [42]. More recently, normalized SGD (NSGD) methods have gained popularity due to their advantage of obviating the need to tune a clipping threshold, and related work continues to emerge. Notably, the convergence performance of the NSGD method can be analyzed without prior knowledge of the tail index p, achieving a sample complexity of $\mathcal{O}(\epsilon^{-\frac{2p}{p-1}})$ [23, 24, 34]. With the incorporation of variance reduction techniques, this complexity was improved to $\mathcal{O}(\epsilon^{-\frac{3p}{2p-2}})$ under the weakly average smoothness condition [23].

For the convex problem $\min_x G(x) := \mathbb{E}[g(x;\xi)]$, where only stochastic subgradients are available, the clipping strategy was incorporated into subgradient estimation, leading to the clipped stochastic subgradient method under the setting $p = 2$ with non-sub-Gaussian noise [36]. This method achieves a sample complexity of $\mathcal{O}(\epsilon^{-2})$ to find $\bar{x}$ with an expected suboptimality $\mathbb{E}[G(\bar{x}) - G(x^*)] \leqslant \epsilon$, where x^* denotes the optimal solution. The general case $1 < p \leqslant 2$ was subsequently studied in [33], yielding sample complexities of $\mathcal{O}(\epsilon^{-\frac{p}{p-1}})$ for nonsmooth convex problems and $\mathcal{O}(\epsilon^{-\frac{p}{2p-2}})$ for strongly convex problems. When stochastic subgradients are unavailable, zeroth-order clipped stochastic similar triangles methods based on two-point and one-point estimation were developed in [27] and [2], respectively. Moreover, a proximal clipped stochastic gradient method was designed for convex optimization problem, i.e., $\min_x G(x) + r(x)$, under heavy-tailed noise [31]. He and Lu considered more general hybrid conditions (including Lipschitz smoothness, Hölder smoothness and Lipschitz continuity of G) and demonstrated that the vanilla stochastic proximal subgradient method already suffices [22]. However, these results rely on the convexity of the objective function, which generally does not hold for our NCO problem (P).

Despite the maturity of research on single-layer stochastic optimization under heavy-tailed noise, the NCO problem (P) presents fundamental challenges in this regime. Specifically, the stochastic estimates of both the inner function and its Jacobian may be contaminated by heavy-tailed noise. First, the inner noise may be distorted by the outer nonlinear function. Furthermore, by the chain rule for composite functions, this nonlinearly distorted noise is multiplicatively coupled with the Jacobian estimation noise in the full gradient. This compounding effect results in a stochastic full gradient that suffers from more severe bias. Moreover, several favorable assumptions commonly used in single-layer optimization fail to hold in the composite setting. These factors collectively imply that NCO under heavy-tailed noise is not a simple extension of its single-layer counterpart, but rather a challenge of multifaceted complexity. Han et al. developed two normalized stochastic composite gradient methods for the doubly stochastic NCO with $r(x) \equiv 0$ [21]. For smooth nonconvex stochastic objectives, the algorithm based on a mini-batch sampling strategy achieves a sample complexity of $\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$, while the variant employing the SPIDER

[2]There exists a point $\bar{x} \in \mathbb{R}^n$ satisfying $\mathbb{E}[\|\nabla G(\bar{x})\|] \leqslant \epsilon$ for any $\epsilon > 0$.

estimator attains $\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$. However, both algorithms strictly rely on large batch sizes for their theoretical guarantees, severely limiting their applicability in scenarios where obtaining a large number of effective samples is expensive. Additionally, there remains a lack of suitable analytical frameworks for solving the nonsmooth NCO problem (P) in the presence of heavy-tailed noise.

1.2 Contributions

This paper develops Normalized Stochastic Proximal Approximation (NSPA) methods for NCO problems (P) in the presence of heavy-tailed noise. The key idea lies in leveraging proximal approximation to tackle the composite structure and normalized update steps to control the potentially severe bias induced by heavy-tailed noise. We propose variants of NSPA methods tailored to problem (P) in Case I and Case II, respectively, establishing their sample complexity bounds (summarized in Table 1) under both constant and decaying parameter settings. Notably, when $p = 2$ (i.e., bounded variance setting), our bounds recover the optimal results established by Zhang and Xiao [50, 52].

- For Case I where f is smooth and possibly nonconvex, we develop normalized stochastic proximal gradient methods: NSPA-PM and NSPA-ST, which use Polyak momentum and the STORM estimator, respectively. Crucially, both methods can guarantee robust convergence under heavy-tailed noise even with small batch sizes, bringing significant benefits to scenarios characterized by limited sample availability. Furthermore, we establish sample complexities for NSPA-PM under standard assumptions, and for NSPA-ST under additional mean p-th moment Lipschitzness and smoothness conditions (Assumption 4). Moreover, we introduce a sliding mechanism that flexibly adjusts the interplay among the batch sizes, momentum parameters, and normalization threshold sequence, thereby enhancing its applicability to various scenarios.

- For Case II where f is nonsmooth and convex, we propose normalized stochastic proximal linearization methods: NSPA-B and NSPA-SP, which employ batch sampling alone and the SPIDER estimator, respectively. Both methods enjoy provable convergence with established sample complexity bounds. Notably, our analysis indicate that the known phenomenon of the nonsmooth outer function leading to different sample size requirements for estimating the inner function and its Jacobian continues to hold under heavy-tailed noise. To the best of our knowledge, these two methods provide the first sample complexity guarantees for solving NCO problem (P) involving a nonsmooth nested structure under heavy-tailed noise.

1.3 Notations

Throughout this paper, let $\|\cdot\|$ denote the Euclidean norm for vectors and the spectral norm for matrices. We use $\langle\cdot,\cdot\rangle$ to denote the standard inner product. Let $[T] = \{1, 2, \ldots, T\}$ for any integer $T \geqslant 1$. We use $\lceil\cdot\rceil$ to denote the ceiling operation. The symbol $\mathbb{E}[\cdot]$ denotes the expectation over the underlying probability space. We write $a_t = \mathcal{O}(b_t)$ when there exists a constant $0 < C < +\infty$ such that $a_t \leqslant Cb_t$. $\tilde{\mathcal{O}}(\cdot)$ additionally hides poly-logarithmic factors.

1.4 Outline

The rest of this paper is organized as follows. Section 2 discusses the challenges of solving the NCO problem under heavy-tailed noise, introduces necessary preliminaries, and presents the generic algorithmic framework. Section 3 develops the NSPA-PM and NSPA-ST methods and details their sample complexity analysis for solving (P) in Case I. Section 4 proposes the NSPA-B and NSPA-SP methods and provides their sample complexity analysis for solving (P) in Case II. Numerical experiments are reported in Section

Table 1: Sample complexities of NSPA methods for finding an ϵ-stationary point of NCO.

Problem	Algorithm	Parameter Type	Sample Complexity ($1 < p < 2$)	Sample Complexity ($p = 2$)	Reference
NCO (P) in Case I	NSPA-PM	Constant	$\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$	$\mathcal{O}(\epsilon^{-4})$	Thm 1
	NSPA-PM	Decaying	$\widetilde{\mathcal{O}}(\epsilon^{-\frac{2p}{p-1}})$	$\mathcal{O}(\epsilon^{-4})$	Thm 2
	NSPA-ST	Constant	$\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$	$\mathcal{O}(\epsilon^{-3})$	Thm 3
	NSPA-ST	Decaying	$\widetilde{\mathcal{O}}(\epsilon^{-\frac{3p}{2p-2}})$	$\mathcal{O}(\epsilon^{-3})$	Thm 4
NCO (P) in Case II	NSPA-B	Constant	$\mathcal{O}(\epsilon^{-\frac{4p-2}{p-1}})$; $\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$	$\mathcal{O}(\epsilon^{-6})$; $\mathcal{O}(\epsilon^{-4})$	Thm 5
	NSPA-B	Decaying	$\widetilde{\mathcal{O}}(\epsilon^{-\frac{3p}{p-1}})$; $\widetilde{\mathcal{O}}(\epsilon^{-\frac{2p}{p-1}})$	$\mathcal{O}(\epsilon^{-6})$; $\mathcal{O}(\epsilon^{-4})$	Thm 6
	NSPA-SP	Constant	$\mathcal{O}(\epsilon^{-\frac{3p-1}{p-1}})$; $\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$	$\mathcal{O}(\epsilon^{-5})$; $\mathcal{O}(\epsilon^{-3})$	Thm 7
	NSPA-SP	Decaying	$\widetilde{\mathcal{O}}(\epsilon^{-\frac{p(3p-1)}{2(p-1)^2}})$; $\widetilde{\mathcal{O}}(\epsilon^{-\frac{3p}{2p-2}})$	$\mathcal{O}(\epsilon^{-5})$; $\mathcal{O}(\epsilon^{-3})$	Thm 8

[1] For NCO in Case II, the two sample complexities correspond to estimating the stochastic inner function and its Jacobian, respectively.

5. Section 6 concludes this paper with further discussion. For brevity, the proofs of all theorems and lemmas are deferred to the Appendix.

2 Normalized stochastic proximal approximation methods

This section lays the foundation for our proposed methods. Following a detailed discussion of the specific challenges induced by heavy-tailed noise and the introduction of necessary preliminaries, we formally present the NSPA framework for solving NSO problem.

2.1 Challenges of heavy-tailed noise

In the NCO problem, heavy-tailed noise does not simply affect the overall model as inner-layer noise; rather, it propagates through the nested structure in a nonlinear manner. Specifically, the gradient of a stochastic composite optimization problem exhibits a multiplicative coupling between the inner Jacobian and the outer gradient, i.e., $\nabla G^\top \cdot \nabla f(G)$. This structure causes the heavy-tailed noise from the inner layer to be propagated through the nonlinear mapping of the outer gradient and multiplied by the inner Jacobian, thereby significantly altering the statistical properties of the full gradient estimate [21]. We examine the behavior of the empirical second moment of the gradient estimates for both the composite function $\Psi(x) = (\xi x)^2$ and the inner function $G(x) = \xi x$ under light-tailed (Gaussian) and heavy-tailed (α-stable) noise in Figure 1. For heavy-tailed distributions with a power-law tail $\Pr(|\xi| > t) \sim t^{-\alpha}$ and $\alpha < 2$, the empirical second moment (i.e., $\sum_{i=1}^N |\xi_i^2|/N$) diverges at a rate of $N^{2/\alpha-1}$ (in probability or almost surely) as the sample size N increases [16]. As illustrated in Figure 1, the second moment of the inner gradient noise diverges at a rate of approximately $N^{1/3}$, while that of the full gradient estimate diverges at a much faster rate of approximately $N^{5/3}$. This contrasts sharply with the behavior observed under Gaussian noise, revealing that the composite structure fundamentally amplifies heavy-tailed noise. Consequently, traditional analyses relying on bounded variance no longer hold, necessitating new theoretical tools and robust algorithms for such settings.

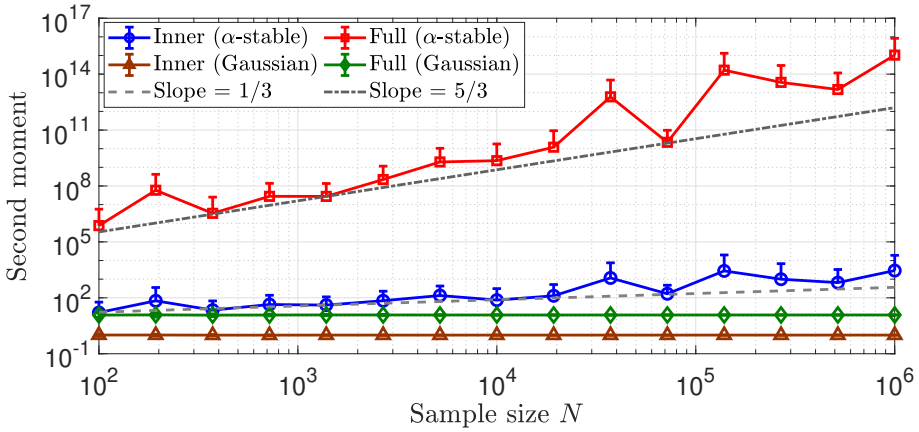


Figure 1: Comparison of the empirical second moment for the full gradient of $\Psi(x) = (\xi x)^2$ and the gradient of inner function $G(x) = \xi x$ under Gaussian noise (light-tailed) and α-stable (with $\alpha = 1.5$) noise (heavy-tailed).

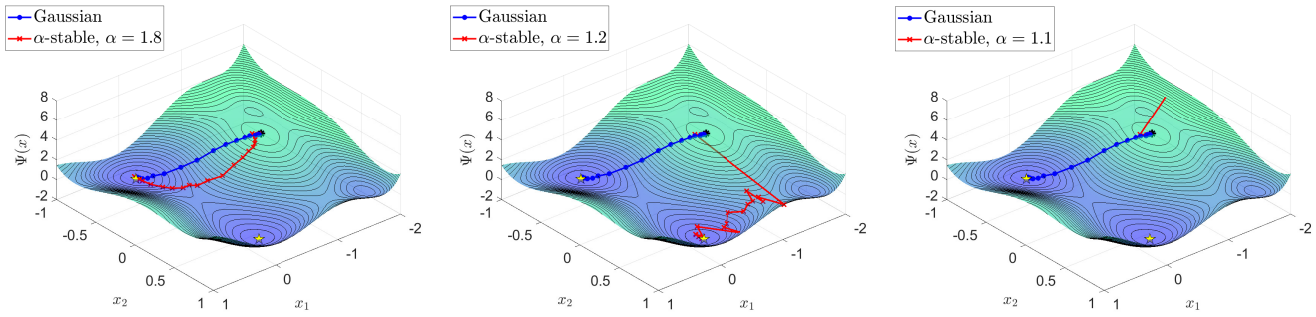


Figure 2: Convergence trajectories of the proximal gradient method under Gaussian noise (light-tailed) and α-stable noise (heavy-tailed) with different tail indices.

Furthermore, we demonstrate the impact of heavy-tailed noise on traditional methods through a simple numerical example. Figure 2 depicts the convergence trajectories of the proximal gradient method [see, e.g., 3] applied to the composite function $\Psi(x) = (4-2.1x_1^2+x_1^4/3)x_1^2+x_1x_2+(-4+4x_2^2)x_2^2+0.1(|x_1|+|x_2|)$ under both Gaussian noise and α-stable noise with different tail indices ($\alpha \in \{1.8, 1.2, 1.1\}$). Under Gaussian noise, the algorithm exhibits smooth and steady descent. In contrast, for heavy-tailed noise, as the noise tail becomes heavier (i.e., smaller α), the convergence trajectories of the algorithm exhibit increasing instability. Specifically, at $\alpha = 1.8$, the trajectory displays only mild fluctuations. At $\alpha = 1.2$, however, the convergence trajectory is suddenly disrupted by extreme outliers in the early iterations, eventually converging to a different local minimum. As the tail index becomes heavier ($\alpha = 1.1$), the trajectory even diverges and fails to stabilize. This highlights the significant effect of heavy-tailed noise on the convergence of optimization algorithms.

2.2 Preliminaries

We now present the assumptions underlying our analysis of the NCO problem.

Assumption 2 *The objective Ψ is bounded below, i.e., $\Psi^* = \inf_{x\in\mathbb{R}^n} \Psi(x) > -\infty$.*

Assumption 3

(a) The function $f : \mathbb{R}^m \to \mathbb{R}$ is ℓ_f-Lipschitz continuous.

(b) The function $G : \mathbb{R}^n \to \mathbb{R}^m$ is ℓ_G-Lipschitz continuous and L_G-smooth.

(c) The function $r : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed and convex.

These assumptions are commonly used in the NCO literature [50, 9, 51, 52], and Assumption 3 (b) directly implies

$$G(x) - G(y) - \nabla G(y)^\top (x - y) \leqslant \frac{L_G}{2}\|x - y\|^2, \qquad \forall x, y \in \mathbb{R}^n. \tag{2}$$

Let $r : \mathbb{R}^n \to \mathbb{R}$ be proper, closed and convex and $\mu > 0$. Then the *proximal operator* $\mathrm{prox}_r^\mu : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$\mathrm{prox}_r^\mu(x) = \arg\min_{z \in \mathbb{R}^n} \left\{ r(z) + \frac{1}{2\mu}\|z - x\|^2 \right\}, \qquad \forall x \in \mathrm{dom}\ r.$$

In this paper, we assume that r is relatively simple, meaning its proximal operator can be efficiently computed. Moreover, since r is closed and convex, the proximal operator is naturally non-expansive [37, section 31], i.e., for any $x, y \in \mathrm{dom}\ r$, $\|\mathrm{prox}_r^\mu(x) - \mathrm{prox}_r^\mu(y)\| \leqslant \|x - y\|$.

Furthermore, we introduce *mean p-th moment Lipschitzness* and *mean p-th moment smoothness* conditions for the stochastic function $g(\cdot; \xi)$ [21], which are crucial for the complexity analysis of variance-reduced algorithms under heavy-tailed noise.

Assumption 4

(a) (mean p-th moment Lipschitzness) There exists a constant $\ell_g > 0$ such that

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

(b) (mean p-th moment smoothness) There exists a constant $L_g > 0$ such that

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

It is notable that these assumptions are compatible with the bounded p-th moment assumption on the noise (i.e., Assumption 1). Moreover, they are strictly weaker than their standard mean-squared or uniform counterparts. The necessity of Assumption 4 arises from the failure of standard variance reduction analyses in heavy-tailed regimes. Since stochastic estimates with heavy-tailed noise possess unbounded variance, traditional mean-squared conditions are mathematically ill-posed. Assumption 4 resolves this limitation by ensuring that the recursive differences in variance-reduced estimators remain bounded in the p-th moment sense. A related concept in the single-layer stochastic optimization literature is the so-called weakly average smoothness [23, 17].

2.3 The generic algorithm framework

We now present a generic framework of normalized stochastic proximal approximation methods for solving the NCO problem (P) under heavy-tailed noise.

For Case I where f is smooth, the classical proximal gradient update is given by

$$\text{(Case I)} \quad \hat{x} = \mathrm{prox}_r^\mu(x - \mu \nabla \Phi(x)) \text{ with } \Phi(x) := f(G(x)),$$

where $\nabla \Phi(x) = \nabla G(x)^\top \nabla f(G(x))$. For Case II where f is nonsmooth, the exact gradient of Φ is not available, and thus a proximal linearization technique [13, 14, 52] is typically employed to update the iterates, i.e.,

$$\text{(Case II)} \quad \hat{x} = \arg\min_{z \in \mathbb{R}^n} \left\{ r(z) + f\big(G(x) + \nabla G(x)^\top (z - x)\big) + \frac{1}{2\mu}\|z - x\|^2 \right\}. \tag{3}$$

Algorithm 1 NSPA method

Require: Initial point $x_1 \in \mathbb{R}^n$, proximal parameter $\mu > 0$, sequence $\rho_t > 0$, and the number of iterations T.

for $t = 1, \ldots, T$ **do**
  Query stochastic oracles to construct the stochastic estimates $\tilde{g}_t$ and $\nabla\tilde{g}_t$ at x_t.
  if Case I **then**
    Compute $\nabla_t^{\tilde{g}} := \nabla\tilde{g}_t^\top \nabla f(\tilde{g}_t)$ and obtain $\tilde{x}_{t+1}$ via (4).
  else if Case II **then**
    Obtain $\tilde{x}_{t+1}$ via (5).
  end if
  Update x_{t+1} through (6).
end for
return $\bar{x}$ chosen at random from $\{x_t\}_{t=1,\ldots,T}$ with x_t selected with probability $\frac{\rho_t}{\sum_{k=1}^T \rho_k}$.

However, since exact evaluations of $G(x)$ and its Jacobian $\nabla G(x)$ are often costly or even prohibitive in practice, we instead rely on their stochastic estimates. At each iteration t, stochastic oracles are queried to construct stochastic estimates $\tilde{g}_t$ for $G(x_t)$ and $\nabla\tilde{g}_t$ for $\nabla G(x_t)$. For Case I, we compute the approximate gradient $\nabla_t^{\tilde{g}} := \nabla\tilde{g}_t^\top \nabla f(\tilde{g}_t)$ and then solve the proximal subproblem to produce a tentative iterate

$$\text{(Case I)} \quad \tilde{x}_{t+1} = \operatorname{prox}_r^\mu(x_t - \mu\nabla_t^{\tilde{g}}). \tag{4}$$

For Case II, we solve the stochastic proximal linearization subproblem, i.e.,

$$\text{(Case II)} \quad \tilde{x}_{t+1} = \arg\min_{z\in\mathbb{R}^n} \left\{ r(z) + f\big(\tilde{g}_t + \nabla\tilde{g}_t^\top(z - x_t)\big) + \frac{1}{2\mu}\|z - x_t\|^2 \right\}. \tag{5}$$

Finally, the algorithm performs a normalized proximal approximation step

$$x_{t+1} = x_t + \kappa_t(\tilde{x}_{t+1} - x_t) \quad \text{with} \quad \kappa_t = \min\left\{ \frac{\mu\rho_t}{\|\tilde{x}_{t+1} - x_t\|}, 1 \right\}, \tag{6}$$

where $\rho_t > 0$ is a predefined normalization threshold sequence, and thus step size $\kappa_t \in (0, 1]$ for any $t \in [T]$. Furthermore, this normalization step guarantees that consecutive iterates satisfy

$$\|x_{t+1} - x_t\| = \kappa_t\|\tilde{x}_{t+1} - x_t\| \leqslant \mu\rho_t.$$

Such a bounding mechanism adaptively mitigates the extreme impact of the heavy-tailed noise, thereby enabling a rigorous convergence analysis through refined step-size control. While the algorithmic framework for Case I largely follows Algorithm 1 in [51], our work incorporates two key distinctions: (i) the use of stochastic estimators to $\tilde{g}_t$ and $\nabla\tilde{g}_t$, and (ii) the sample complexity analysis tailored for the heavy-tailed noise scenario.

The goal of this paper is to establish the sample complexity for proposed methods finding an ϵ-stationary point x that satisfies

$$\mathbb{E}[\|\mathcal{G}(x)\|] \leqslant \epsilon, \quad \text{with} \quad \mathcal{G}(x) := \frac{x - \hat{x}}{\mu}, \tag{7}$$

given tolerance $\epsilon > 0$.

3 The smooth case

In this section, two stochastic estimators are integrated into the NSPA framework for the smooth case, i.e. Case I, yielding the methods NSPA-PM (Algorithm 2 with Option I) and NSPA-ST (Algorithm 2 with Option II). The smoothness of the outer function f is formally assumed below.

Assumption 5 *The function f is differentiable and its gradient is L_f-Lipschitz continuous.*

Combining this assumption with Assumptions 3 (a) and (b), it follows that the function $\Phi(x) = f(G(x))$ is L-smooth with $L := \ell_G^2 L_f + \ell_f L_G$ [see, e.g., 50].

Algorithm 2 NSPA-PM and NSPA-ST methods

Require: Initial point $x_1 \in \mathbb{R}^n$, proximal parameter $\mu > 0$, sequence $\rho_t > 0$, momentum parameters $\beta_t, \gamma_t \in (0, 1]$, sample sizes $B_{t,1}, B_{t,2}$, and iteration number T.

Initialize $x_0 = x_1$, $\beta_1 = \gamma_1 = 1$, $\tilde{g}_0 = 0$, $\nabla\tilde{g}_0 = 0$.

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

Sample the sets $\mathcal{B}_{t,1} = \{\xi_t^{(i)}\}_{i=1}^{B_{t,1}}$ and $\mathcal{B}_{t,2} = \{\hat{\xi}_t^{(i)}\}_{i=1}^{B_{t,2}}$ from distribution Ξ.

Query stochastic oracles to construct the estimates.

Option I (NSPA-PM):

$$\tilde{g}_t = (1-\beta_t)\tilde{g}_{t-1} + \frac{\beta_t}{B_{t,1}} \sum_{i=1}^{B_{t,1}} g(x_t; \xi_t^{(i)}), \quad \nabla\tilde{g}_t = (1-\gamma_t)\nabla\tilde{g}_{t-1} + \frac{\gamma_t}{B_{t,2}} \sum_{i=1}^{B_{t,2}} \nabla g(x_t; \hat{\xi}_t^{(i)}). \tag{8}$$

Option II (NSPA-ST):

$$\tilde{g}_t = (1-\beta_t)\tilde{g}_{t-1} + \frac{\beta_t}{B_{t,1}} \sum_{i=1}^{B_{t,1}} g(x_t; \xi_t^{(i)}) + \frac{1-\beta_t}{B_{t,1}} \sum_{i=1}^{B_{t,1}} \left(g(x_t; \xi_t^{(i)}) - g(x_{t-1}; \xi_t^{(i)}) \right), \tag{9}$$

$$\nabla\tilde{g}_t = (1-\gamma_t)\nabla\tilde{g}_{t-1} + \frac{\gamma_t}{B_{t,2}} \sum_{i=1}^{B_{t,2}} \nabla g(x_t; \hat{\xi}_t^{(i)}) + \frac{1-\gamma_t}{B_{t,2}} \sum_{i=1}^{B_{t,2}} \left(\nabla g(x_t; \hat{\xi}_t^{(i)}) - \nabla g(x_{t-1}; \hat{\xi}_t^{(i)}) \right). \tag{10}$$

Compute the composite gradient $\nabla_t^{\tilde{g}} := \nabla\tilde{g}_t^\top \nabla f(\tilde{g}_t)$ and obtain $\tilde{x}_{t+1}$ via (4).

Update x_{t+1} via (6).

end for

return $\bar{x}$ chosen at random from $\{x_t\}_{t=1,\ldots,T}$ with x_t selected with probability $\frac{\rho_t}{\sum_{k=1}^T \rho_k}$.

Depending on the choice of the estimator, Algorithm 2 branches into two distinct update schemes.

- Option I (NSPA-PM) adopts batch sampling and Polyak momentum to track both the inner function value and its Jacobian, as formulated in (8). This approach stabilizes the update direction by aggregating historical stochastic estimates weighted by the parameters β and γ.
- Option II (NSPA-ST) employs batch sampling and the STORM estimator, which incorporates a recursive correction mechanism ((9) and (10)). By explicitly adding the correction terms (i.e., $g(x_t; \xi_t^{(i)}) - g(x_{t-1}; \xi_t^{(i)})$ and $\nabla g(x_t; \hat{\xi}_t^{(i)}) - \nabla g(x_{t-1}; \hat{\xi}_t^{(i)})$), NSPA-ST effectively mitigates the estimation error induced by heavy-tailed noise. This mechanism enables the estimators to maintain a high-quality approximation of G and ∇G without requiring large batches.

After the estimators $\tilde{g}_t$ and $\nabla\tilde{g}_t$ are constructed, the algorithm evaluates the composite gradient approximation $\nabla_t^{\tilde{g}}$. Finally, it executes a proximal step followed by a normalized update step.

Let $\{x_t, \tilde{g}_t, \nabla\tilde{g}_t\}_{t=1,\dots,T}$ be the sequence generated by Algorithm 2. We proceed to bound $\mathbb{E}[\|\mathcal{G}(\bar{x})\|]$, which serves as a cornerstone of our theoretical analysis in this section.

Lemma 1 *Suppose Assumptions 3 and 5 hold, and $\mu = \frac{1}{2L}$. Then for integers $T \geqslant 1$, it holds that*

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{8L(\Psi(x_1)-\Psi^*)}{\sum_{k=1}^T \rho_k} + \frac{5\ell_f \sum_{t=1}^T \rho_t \mathbb{E}[\|\nabla\tilde{g}_t - \nabla G(x_t)\|]}{\sum_{k=1}^T \rho_k} + \frac{5\ell_G L_f \sum_{t=1}^T \rho_t \mathbb{E}[\|\tilde{g}_t - G(x_t)\|]}{\sum_{k=1}^T \rho_k} + \frac{\sum_{t=1}^T \rho_t^2}{4\sum_{k=1}^T \rho_k}.$$

By Lemma 1, the setting of ρ_t and the expectation errors $\mathbb{E}[\|\nabla\tilde{g}_t - \nabla G(x_t)\|]$ and $\mathbb{E}[\|\tilde{g}_t - G(x_t)\|]$ determine the stationarity of the proposed methods. Consequently, establishing the bound for the expected stationarity measure $\mathbb{E}[\|\mathcal{G}(\bar{x})\|]$ primarily hinges on controlling these two estimation errors. This serves as the foundation of analysis in the remainder of this section.

3.1 NSPA-PM method

This subsection is devoted to analyzing the sample complexity of NSPA-PM for finding an ϵ-stationary point under constant and decaying parameter sequences. For simplicity, let

$$\bar{\beta}_{i:j} := \prod_{k=i}^{j}(1-\beta_k), \qquad \bar{\gamma}_{i:j} := \prod_{k=i}^{j}(1-\gamma_k).$$

with the convention that $\bar{\beta}_{i:j} = \bar{\gamma}_{i:j} = 1$ when $i > j$. Upper bounds on the estimation errors $\mathbb{E}[\|\tilde{g}_t - G(x_t)\|]$ and $\mathbb{E}[\|\nabla\tilde{g}_t - \nabla G(x_t)\|]$ (for NSPA-PM) are established below.

Lemma 2 *Suppose Assumptions 1, 3, and 5 hold. Then for any $t \in [T]$, the output of NSPA-PM satisfies*

$$\mathbb{E}[\|\tilde{g}_t - G(x_t)\|] \leqslant 2V_g \left(\sum_{r=1}^{t} \bar{\beta}_{(r+1):t}^p \beta_r^p B_{r,1}^{1-p}\right)^{\frac{1}{p}} + \mu\ell_G \sum_{r=2}^{t} \bar{\beta}_{r:t}\rho_{r-1},$$

$$\mathbb{E}[\|\nabla\tilde{g}_t - \nabla G(x_t)\|] \leqslant 2V_J \left(\sum_{r=1}^{t} \bar{\gamma}_{(r+1):t}^p \gamma_r^p B_{r,2}^{1-p}\right)^{\frac{1}{p}} + \mu L_G \sum_{r=2}^{t} \bar{\gamma}_{r:t}\rho_{r-1}.$$

The following theorem provides the sample complexity of NSPA-PM with constant parameter sequences.

Theorem 1 *Suppose Assumptions 1, 2, 3, and 5 hold, and let $\mu = \frac{1}{2L}$ and $\rho_t = \rho = \epsilon^c$ with $1 \leqslant c \leqslant \frac{2p-1}{p-1}$. If the batch sizes in (8) are chosen as*

$$B_{t,1} = b_1 = \left\lceil \frac{20\ell_f L_G (80\ell_G L_f V_g)^{\frac{p}{p-1}} \epsilon^{c-\frac{2p-1}{p-1}}}{L} \right\rceil, \quad B_{t,2} = b_2 = \left\lceil \frac{20 L_f \ell_G^2 (80\ell_f V_J)^{\frac{p}{p-1}} \epsilon^{c-\frac{2p-1}{p-1}}}{L} \right\rceil,$$

the momentum parameters are defined as

$$\beta_t = \beta = \min\left\{1, \frac{20\ell_f L_G \epsilon^{c-1}}{L}\right\}, \qquad \gamma_t = \gamma = \min\left\{1, \frac{20 L_f \ell_G^2 \epsilon^{c-1}}{L}\right\},$$

and the number of iterations satisfies $T \geqslant 32L(\Psi(x_1)-\Psi^*)\epsilon^{-c-1}$, then the output $\bar{x}$ of NSPA-PM satisfies $\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \epsilon$. Consequently, the sample complexity to find an ϵ-stationary point is of order $\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$.

Remark 1 *It is worth highlighting that the exponent c serves as a flexible sliding mechanism to adjust the interplay among the sequence ρ_t, batch sizes $B_{t,1}, B_{t,2}$, and momentum parameters β_t, γ_t. This flexibility enables the NSPA-PM to adapt to diverse requirements, thereby enhancing robust performance across different problem regimes. Specifically, the boundary values of c lead to two distinct regimes:*

- *$c = 1$. The iteration complexity in Theorem 1 becomes $T = \mathcal{O}(\epsilon^{-2})$. Here, NSPA-PM sets $\beta_t = \gamma_t = 1$ and relies on larger batch sizes $B_{t,1}, B_{t,2} = \lceil \epsilon^{-\frac{p}{p-1}} \rceil$ to control the estimation errors. In this case, the stochastic estimates for $\tilde{g}_t$ and $\nabla\tilde{g}_t$ in NSPA-PM (i.e., (8) in Algorithm 2) become*

$$\tilde{g}_t = \frac{1}{B_{t,1}} \sum_{i=1}^{B_{t,1}} g(x_t; \xi_t^{(i)}), \qquad \nabla\tilde{g}_t = \frac{1}{B_{t,2}} \sum_{i=1}^{B_{t,2}} \nabla g(x_t; \hat{\xi}_t^{(i)}).$$

  NSPA-PM reduces to its batch-only variant, for which Theorem 1 yields the same sample complexity of $\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$, thereby covering the results in [21].

- *$c = \frac{2p-1}{p-1}$. The number of iterations in Theorem 1 is $T = \mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$, with minimal batch sizes $B_{t,1}, B_{t,2} = \mathcal{O}(1)$. The bias reduction is primarily achieved through an exponential moving average of historical gradient information, with the corresponding momentum parameters satisfying $\beta_t, \gamma_t = \mathcal{O}(\epsilon^{\frac{p}{p-1}})$. When batch sizes $B_{t,1}$ and $B_{t,2}$ are set to 1, the stochastic estimates of $\tilde{g}_t$ and $\nabla\tilde{g}_t$ in (8) take the following form*

$$\tilde{g}_t = (1-\beta_t)\tilde{g}_{t-1} + \beta_t g(x_t; \xi_t^{(i)}), \quad \nabla\tilde{g}_t = (1-\gamma_t)\nabla\tilde{g}_{t-1} + \gamma_t \nabla g(x_t; \hat{\xi}_t^{(i)}),$$

  which reduces NSPA-PM to a pure Polyak momentum variant of NSPA. Theorem 1 also guarantees the sample complexity $\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$ for finding an ϵ-stationary point.

Note that in Theorem 1 the step size ρ_t is heavily coupled with the target precision ϵ, which results in the step size κ_t being very small. Although this yields the optimal sample complexity under standard Lipschitz continuity and smoothness assumptions, it may cause slow initial convergence. To overcome this limitation, we introduce a decaying sequence $\rho_t = t^{-\frac{3}{4}}$ that is relatively large in early iterations to accelerate convergence and then gradually decays to guarantee desired convergence. The corresponding convergence result for the decaying parameter sequences is provided in the following theorem.

Theorem 2 *Suppose Assumptions 1, 2, 3, and 5 hold, and let $T \geqslant 3$, $\mu = \frac{1}{2L}$, and $\rho_t = t^{-\frac{3}{4}}$. If the batch sizes in (8) are chosen as $B_{t,1} = B_{t,2} = 1$, and the momentum parameters are defined as $\beta_t = \gamma_t = t^{-\frac{1}{2}}$, then the output $\bar{x}$ of NSPA-PM satisfies*

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{\Delta_1 \log(T)}{T^{\frac{p-1}{2p}}},$$

where $\Delta_1 = 8L(\Psi(x_1) - \Psi^*) + 20D(\ell_f V_g + \ell_G L_f V_J) + 5D(\ell_f \ell_G + \ell_G L_f L_G)/L + 3/4$. Consequently, by setting $T = \mathcal{O}(\log(\epsilon^{-1})/\epsilon)^{\frac{2p}{p-1}})$, the sample complexity to find an ϵ-stationary point is of order $\widetilde{\mathcal{O}}(\epsilon^{-\frac{2p}{p-1}})$.

The sample complexity results in Theorems 1 and 2 (up to logarithmic factors) match those achieved in the corresponding single-layer stochastic optimization settings [34, 24, 23]. In the special case of $p = 2$, these bounds reduce to $\mathcal{O}(\epsilon^{-4})$ [51, 52]. Notably, under the settings of Theorem 1 with $c = \frac{2p-1}{p-1}$ and Theorem 2, NSPA-PM matches the sample complexities order of [21] despite using merely two samples (more generally, $\mathcal{O}(1)$) per iteration, thereby avoiding the reliance on large batch sizes.

3.2 NSPA-ST method

In this subsection, we provide the corresponding theoretical results for NSPA-ST method (Algorithm 2 with Option II). The following lemma establishes bounds on the estimation errors $\mathbb{E}[\|\tilde{g}_t - G(x_t)\|]$ and $\mathbb{E}[\|\nabla\tilde{g}_t - \nabla G(x_t)\|]$ for NSPA-ST.

Lemma 3 *Suppose Assumptions 1, 3, 4, and 5 hold. Then for any $t \in [T]$, the output of NSPA-ST satisfies*

$$\mathbb{E}[\|\tilde{g}_t - G(x_t)\|] \leqslant 2V_g \left(\sum_{r=1}^{t} \bar{\beta}_{(r+1):t}^{p} \beta_r^p B_{r,1}^{1-p}\right)^{\frac{1}{p}} + 8\mu(\ell_g + \ell_G)\left(\sum_{r=2}^{t} \bar{\beta}_{r:t}^{p} \rho_{r-1}^{p} B_{r,1}^{1-p}\right)^{\frac{1}{p}},$$

$$\mathbb{E}[\|\nabla\tilde{g}_t - \nabla G(x_t)\|] \leqslant 2V_J \left(\sum_{r=1}^{t} \bar{\gamma}_{(r+1):t}^{p} \gamma_r^p B_{r,2}^{1-p}\right)^{\frac{1}{p}} + 8\mu(L_g + L_G)\left(\sum_{r=2}^{t} \bar{\gamma}_{r:t}^{p} \rho_{r-1}^{p} B_{r,2}^{1-p}\right)^{\frac{1}{p}}.$$

The sample complexity analysis of NSPA-ST is given in the following theorem.

Theorem 3 *Suppose Assumptions 1, 2, 3, 4, and 5 hold, and let $\mu = \frac{1}{2L}$ and $\rho_t = \rho = \epsilon^c$ with $1 \leqslant c \leqslant \frac{p}{p-1}$. If the batch sizes in (9) and (10) are chosen as*

$$B_{t,1} = b_1 = \max\left\{\left\lceil (80\ell_G L_f V_g)^{\frac{p}{p-1}} \cdot \epsilon^{c-\frac{p}{p-1}} \right\rceil, \left\lceil \left(\frac{160\ell_G L_f(\ell_g + \ell_G)}{L}\right)^{\frac{p}{p-1}} \cdot \epsilon^{c-\frac{p}{p-1}} \right\rceil\right\},$$

$$B_{t,2} = b_2 = \max\left\{\left\lceil (80\ell_f V_J)^{\frac{p}{p-1}} \cdot \epsilon^{c-\frac{p}{p-1}} \right\rceil, \left\lceil \left(\frac{160\ell_f(L_g + L_G)}{L}\right)^{\frac{p}{p-1}} \cdot \epsilon^{c-\frac{p}{p-1}} \right\rceil\right\},$$

the momentum parameters are defined as $\beta_t = \beta = \epsilon^c$ and $\gamma_t = \gamma = \epsilon^c$, and the number of iterations satisfies $T \geqslant 32L(\Psi(x_1) - \Psi^*)\epsilon^{-c-1}$, then the output $\bar{x}$ of NSPA-ST satisfies $\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \epsilon$. Consequently, the sample complexity to find an ϵ-stationary point is of order $\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$.

Remark 2 *Similar to Remark 1, the exponent $c \in [1, \frac{p}{p-1}]$ provides a sliding mechanism that interpolates between batch sizes and momentum parameters for NSPA-ST. At one extreme ($c = 1$), the algorithm operates in a hybrid regime requiring $\mathcal{O}(\epsilon^{-\frac{1}{p-1}})$ batch sizes, $T = \mathcal{O}(\epsilon^{-2})$ iterations, and momentum parameters $\beta_t = \gamma_t = \epsilon$. At the other extreme ($c = \frac{p}{p-1}$), it recovers a pure STORM-like update with $\mathcal{O}(1)$ batch sizes, $T = \mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$ iterations, and momentum parameters $\beta_t = \gamma_t = \epsilon^{\frac{p}{p-1}}$. Both regimes consistently yield the optimal sample complexity of $\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$.*

Parallel to the analysis of NSPA-PM in Theorem 2, we now present the convergence analysis for NSPA-ST with decaying parameter sequences.

Theorem 4 *Suppose Assumptions 1, 2, 3, 4, and 5 hold, and let $T \geqslant 3$, $\mu = \frac{1}{2L}$, and $\rho_t = t^{-\frac{2}{3}}$. If the batch sizes in (9) and (10) are chosen as $B_{t,1} = B_{t,2} = 1$, and the momentum parameters are defined as $\beta_t = \gamma_t = t^{-\frac{2}{3}}$, then the output $\bar{x}$ of NSPA-ST satisfies*

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{\Delta_2 \log(T)}{T^{\frac{2p-2}{3p}}},$$

where $\Delta_2 = 8L(\Psi(x_1) - \Psi^*) + 20D\left(\ell_f V_g + \ell_G L_f V_J + 2(\ell_f + \ell_G L_f)(\ell_g + \ell_G)L^{-1}\right) + 1$. Consequently, by setting $T = \mathcal{O}((\log(\epsilon^{-1})/\epsilon)^{\frac{3p}{2p-2}})$, the sample complexity to find an ϵ-stationary point is of order $\widetilde{\mathcal{O}}(\epsilon^{-\frac{3p}{2p-2}})$.

Analogously, the sample complexities established in Theorems 3 and 4 are consistent with the corresponding results for single-layer methods with variance reduction [42, 23]. Moreover, when $p = 2$, NSPA-ST matches the optimal complexity results (i.e., $\mathcal{O}(\epsilon^{-3})$) for NCO problems under light-tailed noise [51]. Under the settings of Theorem 3 with $c = \frac{p}{p-1}$ and Theorem 4, NSPA-ST achieves the same sample complexity order as in [21] with only $\mathcal{O}(1)$ samples per iteration.

4 The nonsmooth case

This section focuses on the case where the outer function f is nonsmooth and convex. By incorporating two stochastic estimators into the NSPA framework (Algorithm 1 under Case II), we develop the NSPA-B and NSPA-SP methods and then establish their sample complexities for finding an ϵ-stationary point.

Both NSPA-B and NSPA-SP employ a batch sampling strategy to construct stochastic estimates of $\tilde{g}_t$ and $\nabla\tilde{g}_t$. The detailed procedure is outlined in Algorithm 3. The update scheme in Algorithm 3 takes two different forms, depending on the choice of τ_t:

- When $\tau_t > 1$, Algorithm 3 operates as NSPA-SP. It adopts a double-loop structure consisting of T epochs and τ_t inner iterations per epoch. Specifically, at the beginning of each inner iteration (i.e., $j = 0$), it constructs the estimates for $G(x_{t,0})$ and $\nabla G(x_{t,0})$ using relatively large sample sizes $B_{t,1}$ and $B_{t,2}$ (as shown in (11)). This step serves to periodically reset the accumulated variance introduced by the recursive estimators, providing high-precision reference points. In subsequent inner iterations (i.e., $1 \leqslant j \leqslant \tau_t - 1$), it updates the estimates for both $G(x_{t,j})$ and $\nabla G(x_{t,j})$ using current and previous stochastic information with relatively smaller batch sizes $S_{t,1}$ and $S_{t,2}$ (see (12) and (13)). This recursive structure reduces the estimation bias by effectively tracking the variations in the function and its Jacobian.

- In the special case of $\tau_t = 1$, the inner loop is bypassed, and thus Algorithm 3 reduces to NSPA-B. In this setting, the estimators are constructed solely based on the mini-batch samples at each outer iteration t, without utilizing historical information for recursive updates.

Subsequently, a proximal linearization step followed by a normalization step is performed to complete one iteration.

Let $\{x_{t,j}, \tilde{g}_{t,j}, \nabla\tilde{g}_{t,j}\}_{j=0,\ldots,\tau_t-1}^{t=1,\ldots,T}$ be the sequence generated by Algorithm 3. The bound for $\mathbb{E}[\|\mathcal{G}(\bar{x})\|]$ in the nonsmooth setting is provided below.

Lemma 4 *Suppose Assumption 3 holds and $\mu = \frac{1}{2\ell_f L_G}$. Then for integers $T \geqslant 1$ and $\tau_t \geqslant 1$, it holds that*

$$\begin{aligned}
&\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \\
&\leqslant \frac{24\ell_f L_G(\Psi(x_{1,0}) - \Psi^*)}{\sum_{k=1}^T \tau_k\rho_k} + \frac{4\ell_f\sqrt{L_G}\sum_{t=1}^T \rho_t \sum_{j=0}^{\tau_t-1} \mathbb{E}[\|\tilde{g}_{t,j} - G(x_{t,j})\|^{\frac{1}{2}}]}{\sum_{k=1}^T \tau_k\rho_k} + \frac{3\sum_{t=1}^T \tau_t\rho_t^2}{4\sum_{k=1}^T \tau_k\rho_k} \\
&\quad + \frac{16\ell_f \sum_{t=1}^T \rho_t \sum_{j=0}^{\tau_t-1} \mathbb{E}[\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\|]}{\sum_{k=1}^T \tau_k\rho_k} + \frac{48\ell_f^2 L_G \sum_{t=1}^T \sum_{j=0}^{\tau_t-1} \mathbb{E}[\|\tilde{g}_{t,j} - G(x_{t,j})\|]}{\sum_{k=1}^T \tau_k\rho_k}.
\end{aligned}$$

Lemma 4 reveals that obtaining an upper bound for $\mathbb{E}[\|\mathcal{G}(\bar{x})\|]$ requires controlling the estimation errors $\mathbb{E}[\|\tilde{g}_{t,j} - G(x_{t,j})\|]$ and $\mathbb{E}[\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\|]$, which are presented in the following two subsections.

Algorithm 3 NSPA-B and NSPA-SP methods

Require: Initial point $x_{1,0} \in \mathbb{R}^n$, proximal parameter $\mu > 0$, sequence $\rho_t > 0$ and $\tau_t \geqslant 1$, total outer iterations T, and sample sizes $B_{t,1}, B_{t,2}, S_{t,1}, S_{t,2}$.

for $t = 1, \ldots, T$ **do**
  for $j = 0, \ldots, \tau_t - 1$ **do**
    if $j == 0$ **then**
      Sample the sets $\mathcal{B}_{t,1} = \{\xi_{t,0}^{(i)}\}_{i=1}^{B_{t,1}}$ and $\mathcal{B}_{t,2} = \{\hat{\xi}_{t,0}^{(i)}\}_{i=1}^{B_{t,2}}$ from distribution Ξ.
      Query stochastic oracles to construct the estimates

$$\tilde{g}_{t,0} = \frac{1}{B_{t,1}} \sum_{i=1}^{B_{t,1}} g(x_{t,0}; \xi_t^{(i)}), \qquad \nabla \tilde{g}_{t,0} = \frac{1}{B_{t,2}} \sum_{i=1}^{B_{t,2}} \nabla g(x_{t,0}; \hat{\xi}_t^{(i)}). \tag{11}$$

    else
      Sample the sets $\mathcal{S}_{t,1} = \{\xi_{t,j}^{(i)}\}_{i=1}^{S_{t,1}}$ and $\mathcal{S}_{t,2} = \{\hat{\xi}_{t,j}^{(i)}\}_{i=1}^{S_{t,2}}$ from distribution Ξ.
      Query stochastic oracles to construct the estimates

$$\tilde{g}_{t,j} = \tilde{g}_{t,j-1} + \frac{1}{S_{t,1}} \sum_{i=1}^{S_{t,1}} \big(g(x_{t,j}; \xi_{t,j}^{(i)}) - g(x_{t,j-1}; \xi_{t,j}^{(i)})\big), \tag{12}$$

$$\nabla \tilde{g}_{t,j} = \nabla \tilde{g}_{t,j-1} + \frac{1}{S_{t,2}} \sum_{i=1}^{S_{t,2}} \big(\nabla g(x_{t,j}; \hat{\xi}_{t,j}^{(i)}) - \nabla g(x_{t,j-1}; \hat{\xi}_{t,j}^{(i)})\big). \tag{13}$$

    end if
    Compute the tentative proximal update and the normalized step

$$\tilde{x}_{t,j+1} = \arg\min_{z \in \mathbb{R}^n} \left\{ r(z) + f\big(\tilde{g}_{t,j} + \nabla \tilde{g}_{t,j}^\top (z - x_{t,j})\big) + \frac{1}{2\mu} \|z - x_{t,j}\|^2 \right\},$$

$$x_{t,j+1} = x_{t,j} + \kappa_{t,j+1}(\tilde{x}_{t,j+1} - x_{t,j}) \quad \text{with} \quad \kappa_{t,j+1} = \min\left\{ \frac{\mu \rho_t}{\|\tilde{x}_{t,j+1} - x_{t,j}\|}, 1 \right\}.$$

  end for
  Set $x_{t+1,0} = x_{t,\tau_t}$.
end for
return $\bar{x}$ chosen at random from $\{x_{t,j}\}_{t=1,\ldots,T}^{j=0,\ldots,\tau_t-1}$ with $x_{t,j}$ selected w. p. $\frac{\rho_t}{\sum_{k=1}^T \tau_k \rho_k}$.

4.1 NSPA-B method

This subsection presents the sample complexity analysis of NSPA-B (Algorithm 3 with $\tau_t \equiv 1$). Based on the estimates for $G(x_{t,0})$ and $\nabla G(x_{t,0})$ in NSPA-B, the following lemma establishes upper bounds on the estimation errors $\mathbb{E}[\|\tilde{g}_{t,j} - G(x_{t,j})\|]$ and $\mathbb{E}[\|\nabla \tilde{g}_{t,j} - \nabla G(x_{t,j})\|]$. Its proof can be found in [21, Lemma 5] and is omitted here.

Lemma 5 *Suppose Assumptions 1 and 3 hold. Then for any $t \in [T]$, it holds that*

$$\mathbb{E}[\|\tilde{g}_{t,0} - G(x_{t,0})\|] \leqslant 2V_g B_{t,1}^{-\frac{p-1}{p}}, \qquad \mathbb{E}[\|\nabla \tilde{g}_{t,0} - \nabla G(x_{t,0})\|] \leqslant 2V_J B_{t,2}^{-\frac{p-1}{p}}.$$

We now establish the sample complexity required for NSPA-B to achieve an ϵ-stationary point.

Theorem 5 *Suppose Assumptions 1, 2, and 3 hold, and let* $\mu = \frac{1}{2\ell_f L_G}$ *and* $\rho_t = \rho = \epsilon$. *If the batch sizes in (11) are chosen as*

$$B_{t,1} = b_1 = \left\lceil \left(64\sqrt{2L_G V_g}\ell_f \epsilon^{-1}\right)^{\frac{2p}{p-1}} \right\rceil, \qquad B_{t,2} = b_2 = \left\lceil \left(512\ell_f V_J \epsilon^{-1}\right)^{\frac{p}{p-1}} \right\rceil,$$

and the number of iterations satisfies $T \geqslant 384\ell_f L_G(\Psi(x_{1,0}) - \Psi^*)\epsilon^{-2}$, *then the output* $\bar{x}$ *of NSPA-B satisfies* $\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \epsilon$. *Consequently, the sample complexities for the estimates* $\tilde{g}_t$ *and* $\nabla\tilde{g}_t$ *to find an* ϵ*-stationary point are of order* $\mathcal{O}(\epsilon^{-\frac{4p-2}{p-1}})$ *and* $\mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$, *respectively.*

The convergence result for NSPA-B with decaying ρ_t is presented below.

Theorem 6 *Suppose Assumptions 1, 2, and 3 hold, and for any given integer* $T \geqslant 3$, *let* $\mu = \frac{1}{2\ell_f L_G}$ *and* $\rho_t = t^{-\frac{1}{2}}$. *If the batch sizes in (11) are chosen as* $B_{t,1} = \lceil bT^2 \rceil$ *and* $B_{t,2} = \lceil bT \rceil$ *with* $b > 0$, *then the output* $\bar{x}$ *satisfies*

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{\Delta_3 \log(T)}{T^{\frac{p-1}{p}}},$$

where $\Delta_3 = 24\ell_f L_G(\Psi(x_{1,0}) - \Psi^*) + 4\ell_f\sqrt{2L_G V_g}b^{-\frac{p-1}{2p}} + 96\ell_f^2 L_G V_g b^{-\frac{p-1}{p}} + 32\ell_f V_J b^{-\frac{p-1}{p}} + 3/2$. *Consequently, by setting* $T = \mathcal{O}((\log(\epsilon^{-1})/\epsilon)^{\frac{p}{p-1}})$, *the sample complexities for the estimates* $\tilde{g}_t$ *and* $\nabla\tilde{g}_t$ *to find an* ϵ*-stationary point are of order* $\widetilde{\mathcal{O}}(\epsilon^{-\frac{3p}{p-1}})$ *and* $\widetilde{\mathcal{O}}(\epsilon^{-\frac{2p}{p-1}})$.

Note that Theorems 5 and 6 establish the first sample complexity results for NCO (P) in Case II. In the special case of bounded variance (i.e., $p = 2$), they both recover the sample complexities established in [52], i.e., $\mathcal{O}(\epsilon^{-6})$ and $\mathcal{O}(\epsilon^{-4})$.

4.2 NSPA-SP method

This subsection focuses on the complexity analysis for NSPA-SP (Algorithm 3 with $\tau_t > 1$). The following lemma characterizes the bounds on the estimation errors $\mathbb{E}[\|\tilde{g}_{t,j} - G(x_{t,j})\|]$ and $\mathbb{E}[\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\|]$. See [21, Lemma 7] for the proof.

Lemma 6 *Suppose Assumptions 1, 3, and 4 hold. Then for any* $t \in [T]$ *and* $j \in [\tau_t - 1]$, *it holds that*

$$\mathbb{E}\left[\|\tilde{g}_{t,j} - G(x_{t,j})\|\right] \leqslant 8\mu\rho_t\tau_t^{\frac{1}{p}}(\ell_g + \ell_G)S_{t,1}^{-\frac{p-1}{p}} + 2V_g B_{t,1}^{-\frac{p-1}{p}},$$
$$\mathbb{E}[\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\|] \leqslant 8\mu\rho_t\tau_t^{\frac{1}{p}}(L_g + L_G)S_{t,2}^{-\frac{p-1}{p}} + 2V_J B_{t,2}^{-\frac{p-1}{p}}.$$

Next, we present the sample complexity of NSPA-SP in the following theorem.

Theorem 7 *Suppose Assumptions 1, 2, 3, and 4 hold, and let* $\mu = \frac{1}{2\ell_f L_G}$ *and* $\rho_t = \rho = \epsilon$. *If the batch sizes in (11), (12), and (13) are chosen as*

$$B_{t,1} = b_1 = \left\lceil \left(112\sqrt{2L_G V_g}\ell_f\epsilon^{-1}\right)^{\frac{2p}{p-1}} \right\rceil, \qquad B_{t,2} = b_2 = \left\lceil \left(896\ell_f V_J\epsilon^{-1}\right)^{\frac{p}{p-1}} \right\rceil,$$
$$S_{t,1} = s_1 = \left\lceil \left(224\sqrt{\ell_f(\ell_g + \ell_G)}\epsilon^{-\frac{p+1}{2p}}\right)^{\frac{2p}{p-1}} \right\rceil, \quad S_{t,2} = s_2 = \left\lceil \left(1792(L_g + L_G)L_G^{-1}\epsilon^{-\frac{1}{p}}\right)^{\frac{p}{p-1}} \right\rceil,$$

and the number of iterations satisfies $T \geqslant 672\ell_f L_G(\Psi(x_{1,0}) - \Psi^*)\epsilon^{-1}$ *and* $\tau_t = \tau = \epsilon^{-1}$, *then the output* $\bar{x}$ *satisfies* $\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \epsilon$. *Consequently, the sample complexities for the estimates* $\tilde{g}_t$ *and* $\nabla\tilde{g}_t$ *to find an* ϵ*-stationary point are of order* $\mathcal{O}(\epsilon^{-\frac{3p-1}{p-1}})$ *and* $\mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$, *respectively.*

Under the setting of decaying ρ_t, the convergence guarantee for NSPA-SP is stated in the theorem below.

Theorem 8 *Suppose Assumptions 1, 2, 3, and 4 hold, and for any given integer $T \geqslant 3$, let $\mu = \frac{1}{2\ell_f L_G}$, $\rho_t = t^{-1}$, and $\tau_t = t$. If the batch sizes in (11), (12), and (13) are chosen as $B_{t,1} = \lceil bT^{\frac{2p}{p-1}} \rceil$ and $B_{t,2} = \lceil bT^2 \rceil$ with $b > 0$, and $S_{t,1} = \lceil sT^{\frac{p+1}{p-1}} \rceil$ and $S_{t,2} = \lceil sT \rceil$ with $s > 0$, then the output $\bar{x}$ satisfies*

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{\Delta_4 \log(T)}{T^{\frac{2p-2}{p}}},$$

where

$$\Delta_4 = 24\ell_f L_G(\Psi(x_{1,0}) - \Psi^*) + \frac{16p\sqrt{\ell_f(\ell_g + \ell_G)}s^{-\frac{p-1}{2p}}}{p+1} + 4\ell_f\sqrt{2L_G V_g}b^{-\frac{p-1}{2p}}$$
$$+ 192\ell_f(\ell_g + \ell_G)s^{-\frac{p-1}{p}} + 96\ell_f^2 L_G V_g b^{-\frac{p-1}{p}} + \frac{3}{2} + \frac{64p\ell_f(L_g + L_G)s^{-\frac{p-1}{p}}}{L_G} + 32\ell_f V_J b^{-\frac{p-1}{p}}.$$

Consequently, by setting $T = \mathcal{O}((\log(\epsilon^{-1})/\epsilon)^{\frac{p}{2p-2}})$, the sample complexities for the estimates $\tilde{g}_t$ and $\nabla\tilde{g}_t$ to find an ϵ-stationary point are of order $\widetilde{\mathcal{O}}(\epsilon^{-\frac{p(3p-1)}{2(p-1)^2}})$ and $\widetilde{\mathcal{O}}(\epsilon^{-\frac{3p}{2p-2}})$.

Under the additional Assumption 4, Theorems 7 and 8 establish the first sample complexity results for variance-reduced algorithms in solving the NCO problem under heavy-tailed noise. Similarly, in the bounded variance case, these results coincide with those established in [52], namely $\mathcal{O}(\epsilon^{-5})$ and $\mathcal{O}(\epsilon^{-3})$.

5 Numerical experiments

5.1 Sparse phase retrieval problem

We consider the standard sparse phase retrieval problem, which can be expressed as recovering an s-sparse signal $x^* \in \mathbb{R}^n$ from the measurements

$$y_i = |\langle a_i, x^* \rangle|^2 + \sigma\varepsilon_i, \quad \text{with} \quad i = 1, \ldots, m,$$

where $\{a_i\}_{i=1}^m$ are sensing vectors; $\{y_i\}_{i=1}^m$ denote the phaseless measurements; $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_m]^\top$ denotes the noise vector, and $\sigma > 0$ determines the intensity of the noise [4, 5, 47]. This model can be naturally cast into the framework considered in this paper as follows:

$$\min_{x \in \mathbb{R}^n} \frac{1}{m}\sum_{j=1}^m f(|\langle a_j, x \rangle|^2 - y_j) + \lambda\|x\|_1,$$

where $\lambda > 0$ is the regularization parameter, and f denotes the loss function. To fit the problem framework considered in this paper, we focus on two prevalent loss functions: pseudo-Huber loss $f_1(\cdot) = \sqrt{1 + (\cdot)^2} - 1$ (smooth) and ℓ_1 loss $f_2(\cdot) = |\cdot|$ (nonsmooth and convex), which are referred to as Model I and Model II. Accordingly, we apply NSPA-PM and NSPA-ST to solve Model I, while NSPA-B and NSPA-SP are employed to solve Model II.

For the sparse phase retrieval problem, we set $m = 256$, $n = 128$, $s = 16$, and $\sigma = 1$ and $\lambda = 0.1$. The sensing vectors $\{a_i\}_{i=1}^m$ and the true signal x^* is independently generated from the standard Gaussian distribution $\mathcal{N}(0, 1)$, with the nonzero entries of x^* sampled uniformly without replacement. The heavy-tailed noise is modeled as a symmetric α-stable (SαS) distribution with tail indices $\alpha \in \{1.8, 1.5, 1.2\}$,

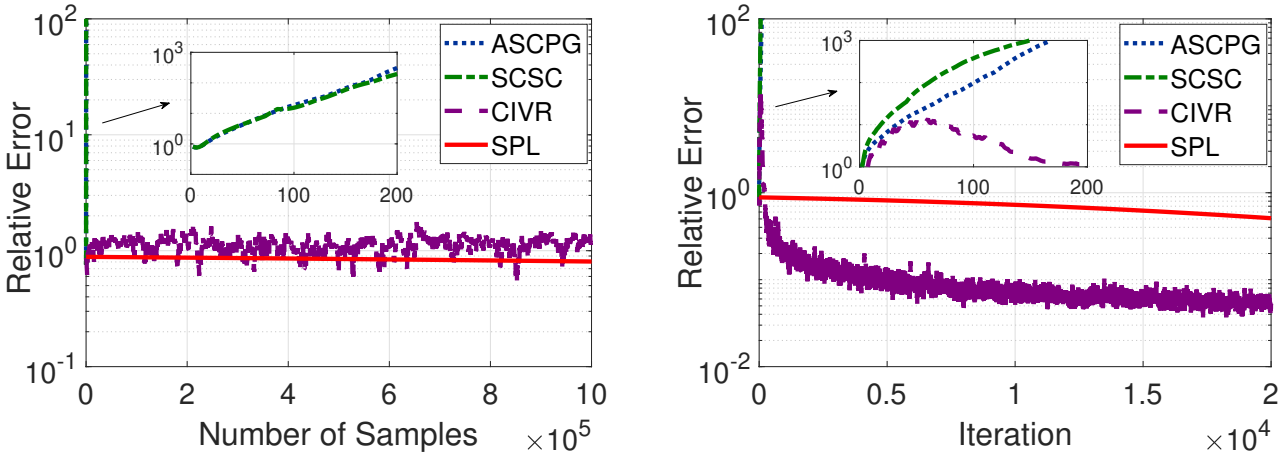
whose characteristic function is $\varphi(t) = \exp(-|t|^\alpha)$. Moreover, we obtain a favorable initial guess via *spectral initialization*, a technique that has been widely adopted in nonconvex phase retrieval problems [4, 5, 6]. To evaluate the recovery performance, the relative error is defined as follows [47]:

$$\textit{Relative Error} := \frac{\min(\|x - x^*\|, \|x + x^*\|)}{\|x^*\|}.$$

While the quadratic inner mapping is not globally Lipschitz continuous, its local Lipschitzness along bounded algorithmic trajectories suffices for our experiments. Guided by our theoretical setting, we set the proximal parameter to $\mu = 0.25$ for NSPA-PM and NSPA-ST, and relax it to $\mu = 0.5$ for NSPA-B and NSPA-SP. Subproblems in the former two are solved via proximal gradient steps with soft-thresholding, whereas the latter two employ subgradient descent with Barzilai-Borwein initialization and a backtracking line search. Following the theoretical discussion, we examine the algorithmic performance with two distinct settings: constant sequences $\{\rho, \beta, \gamma\}$ and decaying sequences $\{\rho_t, \beta_t, \gamma_t\}$, corresponding to the parameter specifications detailed in Table2.

Table 2: Parameter specifications for the proposed methods.

	NSPA-PM	**NSPA-ST**	**NSPA-B**	**NSPA-SP**
Constant	$\rho = 0.01$, $\gamma = \beta = 0.2$, $B_1 = B_2 = 10$	$\rho = 0.01$, $\gamma = \beta = 0.2$, $B_1 = B_2 = 10$	$\rho = 0.01$, $B_1 = 200$, $B_2 = 100$	$\rho = 0.01$, $\tau_t = 5$, $B_1 = 200$, $B_2 = 100$, $S_1 = 100$, $S_2 = 50$
Decaying	$\rho_t = t^{-\frac{3}{4}}$, $\gamma_t = \beta_t = t^{-\frac{1}{2}}$, $B_1 = B_2 = 200$	$\rho_t = t^{-\frac{2}{3}}$, $\gamma_t = \beta_t = t^{-\frac{2}{3}}$, $B_1 = B_2 = 200$	$\rho_t = t^{-\frac{1}{2}}$, $B_1 = 200$, $B_2 = 100$	$\rho_t = t^{-1}$, $\tau_t = 5$, $B_1 = 200$, $B_2 = 100$, $S_1 = 100$, $S_2 = 50$

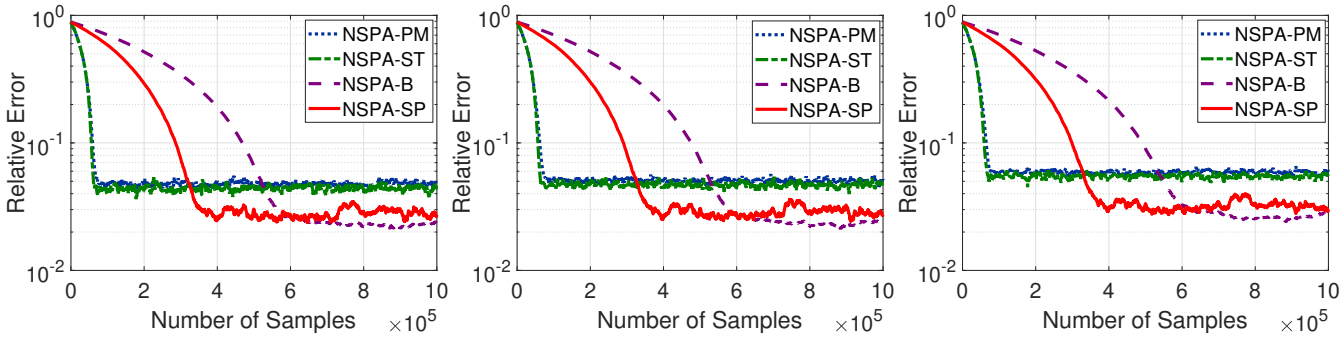


(a) constant parameter set

(b) decaying parameter set

Figure 3: Convergence results of bounded-variance-based algorithms under heavy-tailed noise ($\alpha = 1.5$).

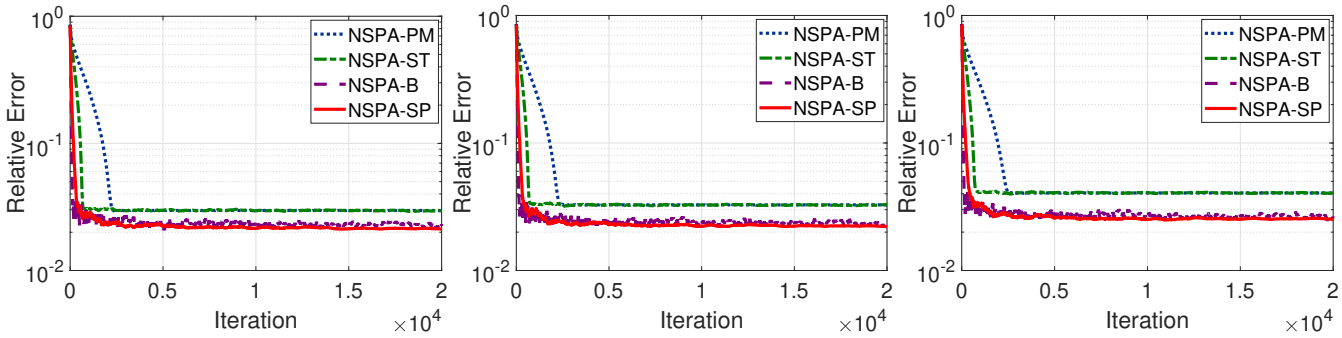
To motivate our algorithms, we first illustrate the convergence limitations of several bounded-variance-based algorithms under heavy-tailed noise. As depicted in Figure 3, we evaluate ASCPG [46], SCSC [8] (adapted to its proximal variant), and CIVR [50] on Model I, alongside SPL [52] on Model II. The general parameters remain identical to those described previously, except that for methods requiring decaying schedules, both the proximal parameter μ and the momentum parameters are set to $t^{-\frac{1}{2}}$. It is evident that the momentum-based ASCPG and SCSC methods diverge rapidly within the the early iterations. Meanwhile, the SPIDER-based CIVR and SPL algorithms suffer from severe trajectory fluctuations or slow convergence rates, demonstrating their inability to effectively handle heavy-tailed noise.



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

Figure 4: Convergence results of the proposed methods with constant parameter set under different noise levels.

Under the constant parameter set, the convergence trajectories of the relative error with respect to the number of samples are illustrated in Figure 4. All four proposed methods converge to stable states with low relative error across different noise regimes, confirming their effectiveness. Specifically, NSPA-PM and NSPA-ST exhibit similar convergence behavior and stabilize with only a small number of samples. Moreover, the relative error increases marginally as the noise becomes heavier-tailed (i.e., as the tail index α decreases). Although NSPA-B and NSPA-SP require more samples to reach stability, they achieve higher convergence accuracy. Additionally, NSPA-SP requires fewer samples than NSPA-B, consistent with theoretical insights.



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

Figure 5: Convergence results of the proposed methods with decaying parameter set under different noise levels.

When using the decaying parameter set, all four algorithms remain convergent under different degrees of heavy-tailed noise, with the evolution of the relative error over iterations depicted in Figure 5. NSPA-B and NSPA-SP exhibit the most superior performance, achieving not only the fastest iteration-wise convergence but also the lowest final relative error. NSPA-ST decreases quickly at first but then levels off at a relatively high error, similar to NSPA-PM, which takes the most iterations to reach stability. Consistent with the results shown in Figure 4, NSPA-B and NSPA-SP exhibit strong robustness to heavy-tailed noise, whereas NSPA-ST and NSPA-PM are more affected. The increased fluctuation arises from the relatively weaker robustness of Model I to heavy-tailed noise; nonetheless, all proposed methods consistently guarantee stable and acceptable convergence.

Next, we present the recovery result for NSPA-SP in Figure 6 (other methods yield similar results and are omitted for brevity. The recovered signal closely coincides with the true signal, confirming that the algorithm yields an accurate approximation of the true signal. To rigorously assess empirical robustness,

Tables 3 and 4 in the Appendix report the final relative errors under various noise scales σ and tail indices α. The results indicate that all algorithms successfully maintain convergence across a diverse range of noise environments. As expected, the final relative error generally increases with a larger σ or a smaller α, reflecting the intrinsic difficulty of optimization in heavy-tailed regimes. Comparing the two parameter choices, the decaying schedule proves more advantageous for achieving lower final relative errors, particularly in low-noise regimes. When comparing the algorithmic variants, NSPA-B and NSPA-SP generally achieve lower relative errors in highly noisy environments, showcasing strong robustness.

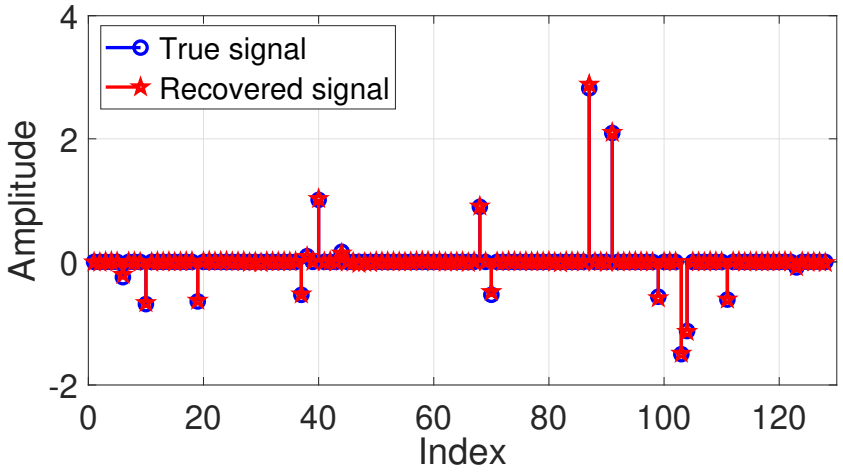


Figure 6: Recover results of NSPA-SP with decaying parameter set and $\alpha = 1.2$.

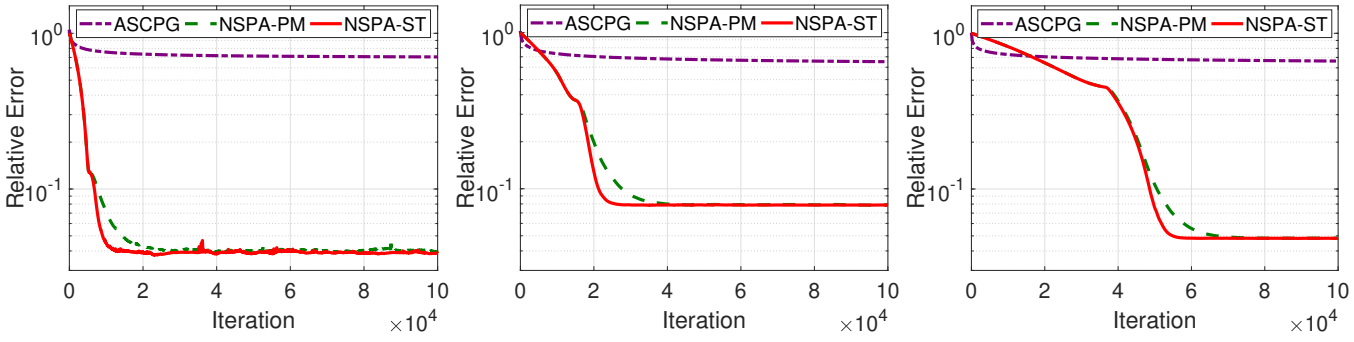
5.2 Policy evaluation for Markov decision processes

In this subsection, we consider a policy evaluation task to validate the effectiveness of NSPA-PM and NSPA-ST with small batch sizes. The experimental set is adapted from Experiment 3 in [46, Section 4]. Specifically, we consider a Markov Decision Process (MDP) tuple $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, characterized by its state space $\mathcal{S}$, action space $\mathcal{A}$, transition probabilities P, reward function R, and discount factor γ. Our goal is to estimate the state-value function $v^\pi \in \mathbb{R}^{|\mathcal{S}|}$ for a target policy π. The function v^π acts as the unique fixed point of the Bellman expectation equation: $v^\pi(s) = \mathbb{E}_\pi[r_{s,s'} + \gamma v^\pi(s')|s]$, with $r_{s,s'}$ representing the immediate transition reward. Under a tabular encoding, we approximate this value function as $v^\pi(s) \approx \varphi_s^\top w^*$, where $\varphi_s \in \mathbb{R}^{|\mathcal{S}|}$ is a one-hot state indicator and $w^* \in \mathbb{R}^{|\mathcal{S}|}$ is the learnable weight. The evaluation task is thus formulated as the regularized Bellman residual minimization:

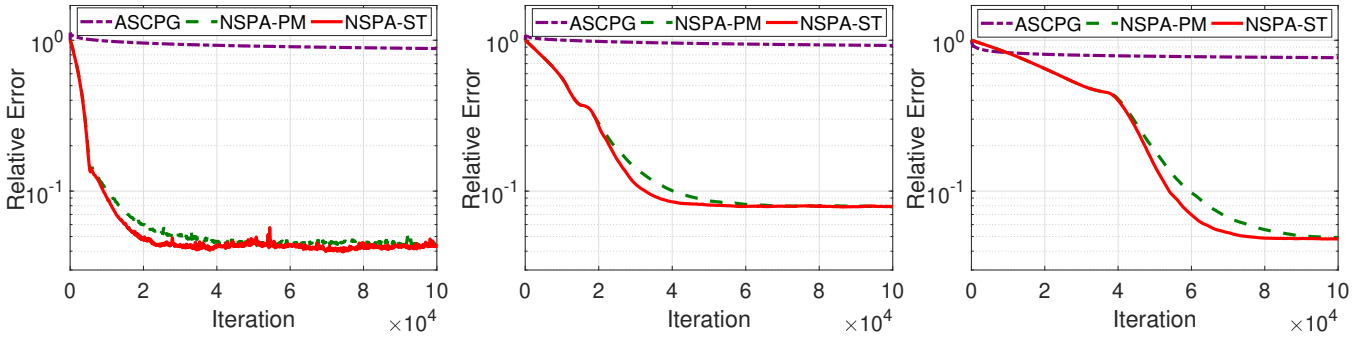
$$\min_{w \in \mathbb{R}^S} \sum_{s=1}^{S} f_1(\varphi_s^\top w - q_s^\pi(w)) + \lambda \|w\|_1,$$

where $q_s^\pi(w) = \mathbb{E}_\pi[r_{s,s'} + \bar{\gamma}\varphi_{s'}^\top w]$ and $f_1(\cdot)$ is the pseudo-Huber loss defined as before. In this simulation, the MDP consists of $|\mathcal{S}| = 500$ states and $|\mathcal{A}| = 5$ actions, utilizing a discount factor of $\gamma = 0.95$. These transition probabilities are generated uniformly at random from $[0, 1]$ and subsequently normalized. We set the regularization parameter $\lambda = 0.005$. The heavy-tailed transition rewards are simulated using a SαS distribution with $\alpha \in \{1.8, 1.5, 1.2\}$.

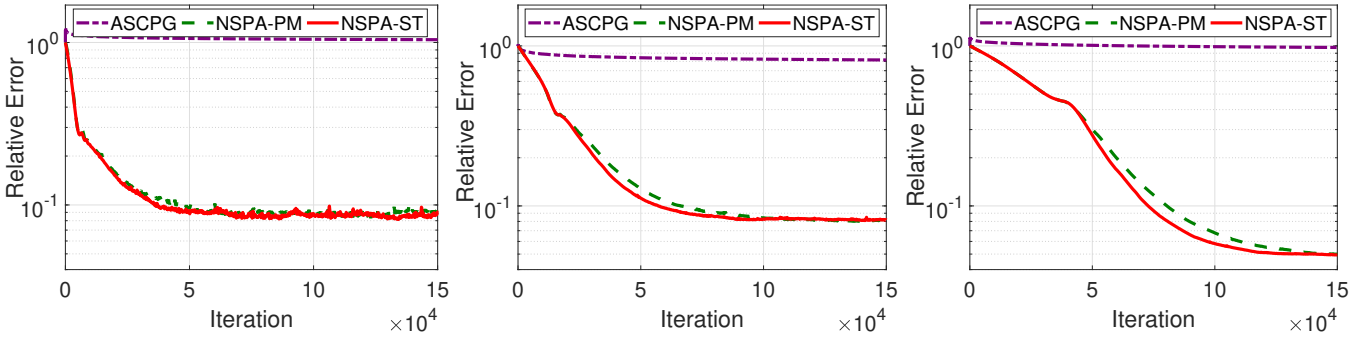
We specifically investigate the convergence behavior under small-sample conditions by drawing merely $B \in \{20, 10, 5, 2\}$ samples per iteration for stochastic estimation. Figure 7 plots the convergence behaviors of ASCPG, NSPA-PM, and NSPA-ST under different noise level. To ensure fairness, the baseline ASCPG adopts the same batch-sampling scheme for its inner function and Jacobian, with other settings identical



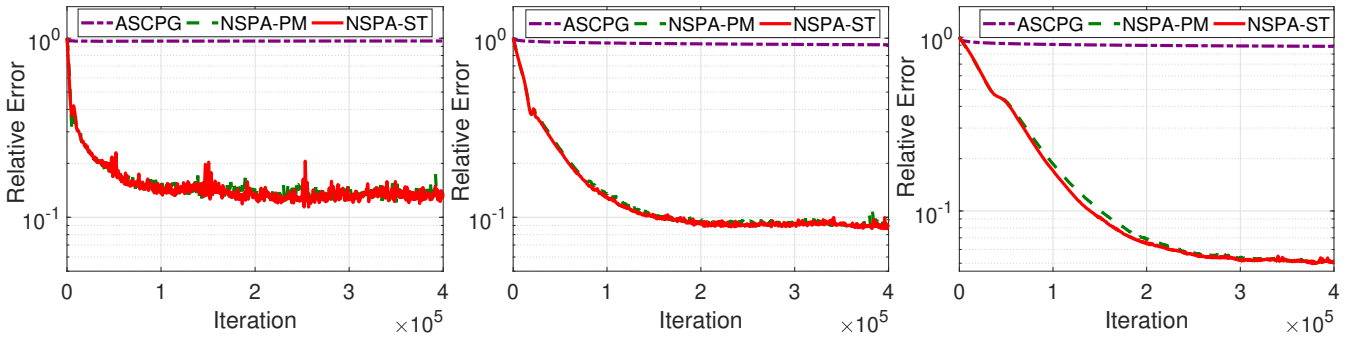
(a) $B = 20$, $\alpha = 1.8$ (b) $B = 20$, $\alpha = 1.5$ (c) $B = 20$, $\alpha = 1.2$



(d) $B = 10$, $\alpha = 1.8$ (e) $B = 10$, $\alpha = 1.5$ (f) $B = 10$, $\alpha = 1.2$



(g) $B = 5$, $\alpha = 1.8$ (h) $B = 5$, $\alpha = 1.5$ (i) $B = 5$, $\alpha = 1.2$



(j) $B = 2$, $\alpha = 1.8$ (k) $B = 2$, $\alpha = 1.5$ (l) $B = 2$, $\alpha = 1.2$

Figure 7: Convergence results of ASCPG, NSPA-PM, and NSPA-ST under different sampling sizes and noise levels.

to [46]. For NSPA-PM and NSPA-ST, we set the momentum parameters as $\beta = \gamma = 0.1$ and the proximal parameter as $\mu = 0.2$. As illustrated in Figure 7, both NSPA-PM and NSPA-ST achieve stable convergence across all tested batch sizes and noise levels, whereas ASCPG fails to converge to a desired accuracy, highlighting the effectiveness of our methods. Moreover, two observations can be made from the empirical results. First, as the noise becomes heavier (smaller α), the convergence rate slows down, requiring a larger number of samples to reach an acceptable error tolerance. This behavior is consistent with our theoretical complexity bounds, which depend explicitly on the tail index. Second, as the batch size decreases, both NSPA-PM and NSPA-ST require more iterations to reach stability. These results confirm that the normalized update mechanism effectively controls extreme outliers induced by heavy-tailed noise, enabling reliable policy evaluation even with limited samples.

6 Conclusions and discussions

This paper has developed the NSPA methods and provided sample complexity analysis for nonsmooth NCO problems under heavy-tailed noise. To mitigate the distortion effect of heavy-tailed noise within the composite structure and its impact on convergence trajectories of the classical proximal method, we propose a generic framework for the NSPA methods. For NCO problem with a smooth outer function, two specific algorithms: NSPA-PM and NSPA-ST were proposed, with the use of batch sampling and momentum strategies. For nonsmooth convex outer function, the NSPA-B and NSPA-SP algorithms were developed based on batch sampling and the SPIDER estimator, respectively. For all four algorithms, we analyzed the associated sample complexity under both constant and decaying parameter sequences. It is noteworthy that the sample complexities of NSPA-PM and NSPA-ST match the optimal results for single-layer nonconvex stochastic optimization under heavy-tailed noise. For NCO problems where the outer function is nonsmooth and convex, the sample complexities obtained by NSPA-B and NSPA-SP are the first to be established. Finally, we validated the effectiveness of the proposed algorithms on the sparse phase retrieval problem and policy evaluation for Markov decision processes.

A natural question is whether momentum-based algorithms can be designed for the NCO problem in Case II to reduce the reliance on large batch sizes inherent in NSPA-B and NSPA-SP. However, establishing the corresponding convergence analysis presents profound technical challenges—a difficulty that even appears in the bounded variance setting [52]. Consequently, whether algorithms relying on small-batch sampling can be designed for the NCO problem in Case II remains an open question. Additionally, for NSPA-SP with decaying parameter sequences (see Theorem 8), the sample complexity $\mathcal{O}(\epsilon^{-\frac{p(3p-1)}{2(p-1)^2}})$ grows quadratically as $p \to 1$. Improving this aspect constitutes a direction for future research.

A Technical Lemmas

We introduce several technical lemmas for the analysis.

Lemma 7 *[51] For any $a \in \mathbb{R}$, it holds that* $\min\{|a|, a^2\} \geqslant |a| - \frac{1}{4}$.

Lemma 8 *Let $T \geqslant 1$ be an integer. For any exponent parameters $a \in (0,1)$ and $d > 1$, the following inequalities hold:*
(a) $T^{1-a} \leqslant \sum_{t=1}^{T} t^{-a} \leqslant \frac{T^{1-a}}{1-a}$; (b) $\sum_{t=1}^{T} t^{-1} \leqslant 2\log(T)$, *provided that* $T \geqslant 3$; (c) $\sum_{t=1}^{T} t^{-d} \leqslant \frac{d}{d-1}$.

Lemma 9 (Inversion Lemma) *Let $a \in (0,1)$ and $d \geqslant e$ be constants. For any $x \geqslant 1$, the inequality $x^a \geqslant d\log(x)$ holds provided that $x \geqslant (\frac{2d}{a}\log(\frac{d}{a}))^{1/a}$.*

Lemma 10 *[24, 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 $\mathbb{E}[\|\sum_{j=1}^n M_j\|^p] \leqslant 2\sum_{j=1}^n \mathbb{E}[\|M_j\|^p]$.*

The following technical lemma generalizes Lemma 9 of [24] by relaxing the exponent restriction from $a \in [0,1]$ to any $a > 0$ satisfying $a - d < 1$ for $d \in (0,1)$. This extension accommodates a broader parameter regime, which is crucial for our convergence analysis under decaying parameter sequences.

Lemma 11 *Let $d \in (0,1)$ and $a > 0$ such that $a - d < 1$. For any integer $t \geqslant 1$, there exists a constant $D > 0$ (depending only on a and d) such that $\sum_{r=1}^t r^{-a} \prod_{k=r+1}^t (1 - k^{-d}) \leqslant D t^{d-a}$.*

Proof. By the inequality $1 - x \leqslant \exp(-x)$ for all $x \in \mathbb{R}$, we obtain $\prod_{k=r+1}^t \left(1 - k^{-d}\right) \leqslant \exp\left(-\sum_{k=r+1}^t k^{-d}\right)$. Since x^{-d} is monotonically decreasing, it holds that

$$\sum_{k=r+1}^{t} k^{-d} \geqslant \int_{r+1}^{t+1} x^{-d} dx = \frac{(t+1)^{1-d} - (r+1)^{1-d}}{1-d}.$$

Thus, we obtain

$$\sum_{r=1}^{t} r^{-a} \prod_{k=r+1}^{t} \left(1 - k^{-d}\right) \leqslant \sum_{r=1}^{t} r^{-a} \exp\left(\frac{(r+1)^{1-d} - (t+1)^{1-d}}{1-d}\right).$$

When the outer sum is approximated by an integral (which introduces at most a bounded constant factor), the asymptotic behavior is governed by the continuous integral $I(t) = \int_1^t x^{-a} \exp\left(-\frac{t^{1-d} - x^{1-d}}{1-d}\right) dx$. To determine the exact asymptotic order, we evaluate the limit $\lim_{t\to\infty} I(t)/t^{d-a}$ via L'Hôpital's rule. Rewriting the ratio as $\int_1^t x^{-a} \exp(\frac{x^{1-d}}{1-d}) dx / (t^{d-a} \exp(\frac{t^{1-d}}{1-d}))$, both the numerator and denominator tend to infinity. Differentiating them with respect to t yields

$$\lim_{t\to\infty} \frac{t^{-a} \exp\left(\frac{t^{1-d}}{1-d}\right)}{(t^{-a} + (d-a)t^{d-a-1}) \exp\left(\frac{t^{1-d}}{1-d}\right)} = \lim_{t\to\infty} \frac{1}{1 + (d-a)t^{d-1}} = 1,$$

where we utilize the fact that $d - 1 < 0$, ensuring $t^{d-1} \to 0$ as $t \to \infty$. This confirms that the integral scales exactly as $\mathcal{O}(t^{d-a})$. Therefore, there exists a constant $D > 0$ such that the bound Dt^{d-a} holds for all $t \geqslant 1$. □

B Proofs in Section 3

This subsection provides the missing proofs in Section 3. Before presenting the proof of Lemma 1, we first establish a descent property for Algorithm 2 in the following lemma. Let approximate proximal gradient mapping $\widetilde{\mathcal{G}}(x) = (x - \tilde{x})/\mu$, where $\tilde{x}$ denotes the tentative iterate constructed from the point x via its stochastic estimates.

Lemma 12 *Suppose Assumptions 3 and 5 hold, and $\mu = \frac{1}{2L}$. Then for any $t \in [T]$, it holds that*

$$\Psi(x_{t+1}) \leqslant \Psi(x_t) + \frac{\rho_t}{2L}\|\nabla_t^{\tilde{g}} - \nabla\Phi(x_t)\| - \frac{\rho_t}{8L}\left(\|\widetilde{\mathcal{G}}(x_t)\| - \frac{1}{4}\rho_t\right).$$

Proof. According to the L-smoothness of $\Phi(x)$, we obtain

$$\begin{aligned}\Psi(x_{t+1}) &= \Phi(x_{t+1}) + r(x_{t+1}) \\ &\leqslant \Phi(x_t) + \langle \nabla\Phi(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 + r(x_{t+1}) \\ &= \underbrace{\Phi(x_t) + \langle \nabla_t^{\tilde{g}}, x_{t+1} - x_t\rangle + \frac{1}{2\mu\kappa_t}\|x_{t+1} - x_t\|^2 + r(x_{t+1})}_{T_1} \\ &\quad + \underbrace{\langle \nabla\Phi(x_t) - \nabla_t^{\tilde{g}}, x_{t+1} - x_t\rangle}_{T_2} + \underbrace{\frac{\mu\kappa_t L - 1}{2\mu\kappa_t}\|x_{t+1} - x_t\|^2}_{T_3}.\end{aligned} \tag{14}$$

For term T_1, the normalization step (6) (i.e., $x_{t+1} = x_t + \kappa_t(\tilde{x}_{t+1} - x_t)$) leads to

$$\begin{aligned}T_1 &= \Phi(x_t) + \kappa_t\langle \nabla_t^{\tilde{g}}, \tilde{x}_{t+1} - x_t\rangle + \frac{\kappa_t}{2\mu}\|\tilde{x}_{t+1} - x_t\|^2 + r(x_t + \kappa_t(\tilde{x}_{t+1} - x_t)) \\ &\leqslant \Phi(x_t) + (1-\kappa_t)r(x_t) + \kappa_t\left(r\left(\tilde{x}_{t+1}\right) + \langle \nabla_t^{\tilde{g}}, \tilde{x}_{t+1} - x_t\rangle + \frac{1}{2\mu}\|\tilde{x}_{t+1} - x_t\|^2\right) \\ &\leqslant \Phi(x_t) + r(x_t) = \Psi(x_t),\end{aligned}$$

where the first inequality is by the convexity of r and the last inequality is due to the minimality of $\tilde{x}_{t+1}$, i.e., $r\left(\tilde{x}_{t+1}\right) + \langle \nabla_t^{\tilde{g}}, \tilde{x}_{t+1} - x_t\rangle + \frac{1}{2\mu}\|\tilde{x}_{t+1} - x_t\|^2 \leqslant r(x_t)$. By Cauchy-Schwarz inequality, we have

$$T_2 \leqslant \|\nabla_t^{\tilde{g}} - \nabla\Phi(x_t)\|\|x_{t+1} - x_t\| \leqslant \mu\rho_t\|\nabla_t^{\tilde{g}} - \nabla\Phi(x_t)\| = \frac{\rho_t}{2L}\|\nabla_t^{\tilde{g}} - \nabla\Phi(x_t)\|,$$

where we invoke the choice of parameter $\mu = \frac{1}{2L}$ in the last equality. Again from the normalization step (6) and $\mu = \frac{1}{2L}$, we obtain

$$T_3 = \frac{\mu\kappa_t^2 L - \kappa_t}{2\mu}\|\tilde{x}_{t+1} - x_t\|^2 = \frac{\mu^2\kappa_t^2 L - \mu\kappa_t}{2}\|\widetilde{\mathcal{G}}(x_t)\|^2 = \frac{\kappa_t^2 - 2\kappa_t}{8L}\|\widetilde{\mathcal{G}}(x_t)\|^2.$$

Applying the fact that $\kappa_t^2 - 2\kappa_t \leqslant -\kappa_t$ for any $\kappa_t \in (0, 1]$ yields

$$T_3 \leqslant -\frac{\kappa_t}{8L}\|\widetilde{\mathcal{G}}(x_t)\|^2 = -\frac{\rho_t^2}{8L}\min\left\{\|\widetilde{\mathcal{G}}(x_t)\|\rho_t^{-1}, \|\widetilde{\mathcal{G}}(x_t)\|^2\rho_t^{-2}\right\} \leqslant -\frac{\rho_t}{8L}\left(\|\widetilde{\mathcal{G}}(x_t)\| - \frac{1}{4}\rho_t\right),$$

where we use Lemma 7 with $z = \|\widetilde{\mathcal{G}}(x_t)\|\rho_t^{-1}$ in the last inequality. Finally, substituting this bound into (14) yields the desired result. ▫

B.1 Proof of Lemma 1

Proof. Note that $\|\mathcal{G}(x_t)\| = \frac{1}{\mu}\|x_t - \hat{x}_{t+1}\| \leqslant \frac{1}{\mu}\|x_t - \tilde{x}_{t+1}\| + \frac{1}{\mu}\|\tilde{x}_{t+1} - \hat{x}_{t+1}\| = \|\widetilde{\mathcal{G}}(x_t)\| + \frac{1}{\mu}\|\tilde{x}_{t+1} - \hat{x}_{t+1}\|$. By the non-expansiveness of the proximal operator, we have

$$\frac{1}{\mu}\|\tilde{x}_{t+1} - \hat{x}_{t+1}\| = \frac{1}{\mu}\|\mathrm{prox}_r^{\mu}(x_t - \mu\nabla_t^{\tilde{g}}) - \mathrm{prox}_r^{\mu}(x_t - \mu\nabla\Phi(x_t))\| \leqslant \|\nabla_t^{\tilde{g}} - \nabla\Phi(x_t)\|.$$

Additionally, Lemma 12 implies

$$\rho_t\|\widetilde{\mathcal{G}}(x_t)\| \leqslant 8L(\Psi(x_t) - \Psi(x_{t+1})) + 4\rho_t\|\nabla_t^{\tilde{g}} - \nabla\Phi(x_t)\| + \frac{\rho_t^2}{4}.$$

Then, it holds that

$$\begin{aligned}\rho_t\|\mathcal{G}(x_t)\| &\leqslant 8L(\Psi(x_t)-\Psi(x_{t+1}))+5\rho_t\|\nabla_t^{\tilde{g}}-\nabla\Phi(x_t)\|+\frac{\rho_t^2}{4}\\ &\leqslant 8L(\Psi(x_t)-\Psi(x_{t+1}))+5\rho_t\left(\ell_f\|\nabla\tilde{g}_t-\nabla G(x_t)\|+\ell_G L_f\|\tilde{g}_t-G(x_t)\|\right)+\frac{\rho_t^2}{4},\end{aligned}$$

where the last inequality is due to Assumptions 3 and 5. Taking the expectations on the above inequality and adding it for $t=1,\ldots,T$ gives

$$\begin{aligned}\sum_{t=1}^{T}\rho_t\mathbb{E}[\|\mathcal{G}(x_t)\|] \leqslant 5\sum_{t=1}^{T}\rho_t\left(\ell_f\mathbb{E}[\|\nabla\tilde{g}_t-\nabla G(x_t)\|]+\ell_G L_f\mathbb{E}[\|\tilde{g}_t-G(x_t)\|]\right)\\ +8L(\Psi(x_1)-\Psi^*)+\sum_{t=1}^{T}\frac{\rho_t^2}{4}.\end{aligned}\tag{15}$$

Recall that the output $\bar{x}$ is a random iterate sampled from $\{x_t\}_{t=1}^T$ with weights proportional to $\{\rho_t\}_{t=1}^T$. Consequently, we have $\mathbb{E}[\|\mathcal{G}(\bar{x})\|] = \frac{1}{\sum_{k=1}^T\rho_k}\sum_{t=1}^T\rho_t\mathbb{E}[\|\mathcal{G}(x_t)\|]$. Furthermore, substituting (15) into this equation yields the desired result. □

B.2 Proof of Lemma 2

The proof strategy of estimation error bound follows the framework established in [24, Lemma 17]. However, to accommodate both the momentum update and batch sampling in our stochastic estimator, we extend this analysis by applying Lemma 10 twice, yielding guarantees that hold for general cases.

Proof. Recalling the definition of $\tilde{g}_t$ in (8), it follows directly that

$$\begin{aligned}\tilde{g}_t-G(x_t) &= (1-\beta_t)\tilde{g}_{t-1}+\frac{\beta_t}{B_{t,1}}\sum_{i=1}^{B_{t,1}}g(x_t;\xi_t^{(i)})-G(x_t)\\ &= (1-\beta_t)(\tilde{g}_{t-1}-G(x_{t-1}))+\frac{\beta_t}{B_{t,1}}\sum_{i=1}^{B_{t,1}}\Big(g(x_t;\xi_t^{(i)})-G(x_t)\Big)+(1-\beta_t)(G(x_{t-1})-G(x_t)).\end{aligned}$$

Recursively applying this relation down to $t=1$ yields

$$\begin{aligned}\tilde{g}_t-G(x_t) &= \bar{\beta}_{2:t}(\tilde{g}_1-G(x_1))+\sum_{r=2}^{t}\bar{\beta}_{(r+1):t}\nu_r+\sum_{r=2}^{t}\bar{\beta}_{r:t}(G(x_{r-1})-G(x_r))\\ &= \sum_{r=1}^{t}\bar{\beta}_{(r+1):t}\nu_r+\sum_{r=2}^{t}\bar{\beta}_{r:t}(G(x_{r-1})-G(x_r)),\end{aligned}$$

where we denote $\nu_r=\frac{\beta_r}{B_{r,1}}\sum_{i=1}^{B_{r,1}}(g(x_r;\xi_r^{(i)})-G(x_r))$ and the second equality is due to $\tilde{g}_0=0$ and $\beta_1=1$. Then, it holds that

$$\begin{aligned}\mathbb{E}[\|\tilde{g}_t-G(x_t)\|] &\leqslant \mathbb{E}\left[\left\|\sum_{r=1}^{t}\bar{\beta}_{(r+1):t}\nu_r\right\|\right]+\sum_{r=2}^{t}\bar{\beta}_{r:t}\mathbb{E}[\|G(x_{r-1})-G(x_r)\|]\\ &\leqslant \mathbb{E}\left[\left\|\sum_{r=1}^{t}\bar{\beta}_{(r+1):t}\nu_r\right\|^p\right]^{\frac{1}{p}}+\mu\ell_G\sum_{r=2}^{t}\bar{\beta}_{r:t}\rho_{r-1},\end{aligned}\tag{16}$$

where the second inequality follows from applying Jensen's inequality to the first term and the Lipschitz continuity of G to the second. Next, we apply Lemma 10 to control the first term on the right side. Let $\mathcal{F}_t = \sigma(\xi_1, \hat{\xi}_1, \ldots, \xi_t, \hat{\xi}_t)$ be the associated filtration. We denote $M_r := \bar{\beta}_{(r+1):t}\nu_r$ and check that for any $r \in [t]$,

$$\begin{aligned}\mathbb{E}[M_r|\mathcal{F}_{r-1}] &= \bar{\beta}_{(r+1):t}\frac{\beta_r}{B_{r,1}}\sum_{i=1}^{B_{r,1}}\mathbb{E}\left[g(x_r;\xi_r^{(i)}) - G(x_r)|\mathcal{F}_{r-1}\right] \\ &= \bar{\beta}_{(r+1):t}\frac{\beta_r}{B_{r,1}}\sum_{i=1}^{B_{r,1}}\mathbb{E}\left[g(x_r;\xi_r^{(i)}) - G(x_r)|x_r\right] = 0,\end{aligned}$$

where the second equality holds because x_r is $\mathcal{F}_{r-1}$-measurable and ξ_r is independent of $\mathcal{F}_{r-1}$. By the convexity of $\|\cdot\|^p$ with $1 < p \leqslant 2$, we obtain

$$\mathbb{E}[\|M_r\|^p] \leqslant \frac{\bar{\beta}^p_{(r+1):t}\beta_r^p}{B_{r,1}}\sum_{i=1}^{B_{r,1}}\mathbb{E}\left[\mathbb{E}\left[\left\|g(x_r;\xi_r^{(i)}) - G(x_r)\right\|^p|x_r\right]\right] \leqslant \bar{\beta}^p_{(r+1):t}\beta_t^p V_g^p \leqslant +\infty,$$

where the second inequality follows from Assumption 1. Hence, we apply Lemma 10 to obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{r=1}^t \bar{\beta}_{(r+1):t}\nu_r\right\|^p\right]^{\frac{1}{p}} &\leqslant \mathbb{E}\left[2\sum_{r=1}^t\|\bar{\beta}_{(r+1):t}\nu_r\|^p\right]^{\frac{1}{p}} \\ &= \left(2\sum_{r=1}^t\frac{\bar{\beta}^p_{(r+1):t}\beta_r^p}{B_{r,1}^p}\mathbb{E}\left[\left\|\sum_{i=1}^{B_{r,1}}\left(g(x_r;\xi_r^{(i)}) - G(x_r)\right)\right\|^p\right]\right)^{\frac{1}{p}}. \qquad (17)\end{aligned}$$

To bound the approximation error $\mathbb{E}[\|\sum_{i=1}^{B_{r,1}}(g(x_r;\xi_r^{(i)}) - G(x_r))\|^p]$, Lemma 10 is invoked once again. Let $M_r^i := g(x_r;\xi_r^{(i)}) - G(x_r)$ with $i = 1, \ldots, B_{r,1}$. By the independent of the mini-batch and the measurability of x_r, we obtain $\mathbb{E}[M_r^i|M_r^1, \ldots, M_r^{i-1}] = 0$. Moreover, the bounded p-th moment property implies $\mathbb{E}[\|M_r^i\|^p] \leqslant \mathbb{E}[\mathbb{E}[\|g(x_r;\xi_r^{(i)}) - G(x_r)\|^p|x_r]] \leqslant V_g^p \leqslant +\infty$. Consequently, applying Lemma 10 yields

$$\mathbb{E}\left[\left\|\sum_{i=1}^{B_{r,1}}\left(g(x_r;\xi_r^{(i)}) - G(x_r)\right)\right\|^p\right] \leqslant 2\sum_{i=1}^{B_{r,1}}\mathbb{E}\left[\left\|g(x_r;\xi_r^{(i)}) - G(x_r)\right\|^p\right] \leqslant 2B_{r,1}V_g^p.$$

Plugging this bound into (17) gives

$$\mathbb{E}\left[\left\|\sum_{r=1}^t \bar{\beta}_{(r+1):t}\nu_r\right\|^p\right]^{\frac{1}{p}} \leqslant 2V_g\left(\sum_{r=1}^t\bar{\beta}^p_{(r+1):t}\beta_r^p B_{r,1}^{1-p}\right)^{\frac{1}{p}}. \qquad (18)$$

Finally, combining (18) with (16) establishes the bound for $\mathbb{E}[\|\tilde{g}_t - G(x_t)\|]$. The upper bound for $\mathbb{E}[\|\nabla\tilde{g}_t - \nabla G(x_t)\|]$ follows a similar derivation and is therefore omitted for brevity. □

B.3 Proof of Theorem 1

Proof. Since the parameters selected in this theorem are t-independent, the bound on $\mathbb{E}[\|\tilde{g}_t - G(x_t)\|]$ from Lemma 2 simplifies as

$$\mathbb{E}[\|\tilde{g}_t - G(x_t)\|] \leqslant 2\beta V_g b_1^{\frac{1-p}{p}}\left(\sum_{r=1}^t(1-\beta)^{p(t-r)}\right)^{\frac{1}{p}} + \mu\ell_G\rho\sum_{r=2}^t(1-\beta)^{t-r+1}$$

$$\leqslant \frac{2\beta V_g b_1^{\frac{1-p}{p}}}{(1-(1-\beta)^p)^{1/p}} + \frac{\mu \ell_G \rho}{\beta} \leqslant 2V_g \beta^{\frac{p-1}{p}} b_1^{\frac{1-p}{p}} + \frac{\ell_G \rho}{2\beta L},$$

where we use the property that $\sum_{k=0}^{t} q^k \leqslant \frac{1}{1-q}$ for $q \in (0,1)$ in the second inequality, and the last inequality follows from $1-(1-\beta)^p \geqslant \beta$ for $\beta \in (0,1)$, along with the condition $\mu = \frac{1}{2L}$. Similarly, we can also obtain

$$\mathbb{E}[\|\nabla \tilde{g}_t - \nabla G(x_t)\|] \leqslant 2V_J \gamma^{\frac{p-1}{p}} b_2^{\frac{1-p}{p}} + \frac{L_G \rho}{2\gamma L}.$$

By substituting this bound into Lemma 1 with t-independent parameters, we derive

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{8L(\Psi(x_1)-\Psi^*)}{\rho T} + 10\ell_f V_J \gamma^{\frac{p-1}{p}} b_2^{\frac{1-p}{p}} + 10\ell_G L_f V_g \beta^{\frac{p-1}{p}} b_1^{\frac{1-p}{p}} + \frac{5\ell_f L_G \rho}{2\gamma L} + \frac{5L_f \ell_G^2 \rho}{2\beta L} + \frac{\rho}{4}.$$

With the choices of parameters ρ, T, β, γ, b_1, and b_2, it holds that $\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \epsilon$, and then the sample complexity is $\sum_{t=1}^{T}(B_{t,1}+B_{t,2}) = \mathcal{O}(\epsilon^{-c-1} \cdot (\epsilon^{c-\frac{2p-1}{p-1}} + \epsilon^{c-\frac{2p-1}{p-1}})) = \mathcal{O}(\epsilon^{-\frac{3p-2}{p-1}})$. ▫

B.4 Proof of Theorem 2

Proof. With the t-dependent parameters schedule chosen in this theorem, the estimate for $\mathbb{E}[\|\tilde{g}_t - G(x_t)\|]$ in Lemma 2 becomes

$$\begin{aligned}\mathbb{E}[\|\tilde{g}_t - G(x_t)\|] &\leqslant 2V_g \left(\sum_{r=1}^{t} r^{-\frac{p}{2}} \prod_{k=r+1}^{t} \left(1-k^{-\frac{1}{2}}\right)\right)^{\frac{1}{p}} + \mu \ell_G \sum_{r=2}^{t} (r-1)^{-\frac{3}{4}} \prod_{k=r}^{t} \left(1-k^{-\frac{1}{2}}\right) \\ &\leqslant 2DV_g t^{\frac{1-p}{2p}} + \frac{D\ell_G t^{-\frac{1}{4}}}{2L},\end{aligned}$$

where we use Lemma 11 with $(a,d) = (\frac{p}{2}, \frac{1}{2})$ and $(a,d) = (\frac{3}{4}, \frac{1}{2})$ in the last inequality. Similarly, the estimate for $\mathbb{E}[\|\nabla \tilde{g}_t - \nabla G(x_t)\|]$ satisfies

$$\mathbb{E}[\|\nabla \tilde{g}_t - \nabla G(x_t)\|] \leqslant 2DV_J t^{\frac{1-p}{2p}} + \frac{DL_G t^{-\frac{1}{4}}}{2L}.$$

Combining these bounds with Lemma 1 and the selected t-dependent parameters leads to

$$\begin{aligned}\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant{}& \frac{8L(\Psi(x_1)-\Psi^*)}{\sum_{t=1}^{T} t^{-\frac{3}{4}}} + \frac{10D(\ell_f V_g + \ell_G L_f V_J)\sum_{t=1}^{T} t^{-\frac{5p-2}{4p}}}{\sum_{t=1}^{T} t^{-\frac{3}{4}}} \\ &+ \frac{5D(\ell_f \ell_G + \ell_G L_f L_G)\sum_{t=1}^{T} t^{-1}}{2L\sum_{t=1}^{T} t^{-\frac{3}{4}}} + \frac{\sum_{t=1}^{T} t^{-\frac{3}{2}}}{4\sum_{t=1}^{T} t^{-\frac{3}{4}}}.\end{aligned}$$

Noting that for $p \in (1,2]$, it holds that $3/4 < (5p-2)/(4p) \leqslant 1$ and $(2-p)/(4p) \geqslant 0$, which implies $\sum_{t=1}^{T} t^{-\frac{5p-2}{4p}} = \sum_{t=1}^{T} t^{-1} \cdot t^{\frac{2-p}{4p}} \leqslant T^{\frac{2-p}{4p}} \sum_{t=1}^{T} t^{-1}$. By invoking Lemma 8 with $a = \frac{3}{4}$ and $d = \frac{3}{2}$, we further have

$$\begin{aligned}\mathbb{E}[\|\mathcal{G}(\bar{x})\|] &\leqslant \frac{8L(\Psi(x_1)-\Psi^*)}{T^{\frac{1}{4}}} + \frac{20D(\ell_f V_g + \ell_G L_f V_J)\log(T)}{T^{\frac{p-1}{2p}}} + \frac{5D(\ell_f \ell_G + \ell_G L_f L_G)\log(T)}{LT^{\frac{1}{4}}} + \frac{3}{4T^{\frac{1}{4}}} \\ &\leqslant \frac{\Delta_1 \log(T)}{T^{\frac{p-1}{2p}}},\end{aligned}$$

where the last inequality is due to $\log(T) \geqslant 1$ with $T \geqslant 3$ and

$$\Delta_1 := 8L(\Psi(x_1) - \Psi^*) + 20D(\ell_f V_g + \ell_G L_f V_J) + \frac{5D(\ell_f \ell_G + \ell_G L_f L_G)}{L} + \frac{3}{4}.$$

Furthermore, by setting $T = \mathcal{O}((\log(\epsilon^{-1})/\epsilon)^{\frac{2p}{p-1}})$ and applying Lemma 9 with $(x, a, d) = (T, \frac{p-1}{2p}, \frac{\Delta_1}{\epsilon})$, NSPA-PM achieves an ϵ-stationary point. Since each iteration requires only two samples, the sample complexity to find an ϵ-stationary point is $\mathcal{O}((\log(\epsilon^{-1})/\epsilon)^{\frac{2p}{p-1}})$. By further hiding logarithmic factors with the $\tilde{\mathcal{O}}$ notation, we arrive at the desired result.

□

B.5 Proof of Lemma 3

Proof. For the update rule (9) in Algorithm 2, we obtain

$$\begin{aligned}
\tilde{g}_t - G(x_t) &= (1-\beta_t)\tilde{g}_{t-1} + \frac{\beta_t}{B_{t,1}} \sum_{i=1}^{B_{t,1}} g(x_t; \xi_t^{(i)}) + \frac{1-\beta_t}{B_{t,1}} \sum_{i=1}^{B_{t,1}} \Big(g(x_t; \xi_t^{(i)}) - g(x_{t-1}; \xi_t^{(i)}) \Big) - G(x_t) \\
&= (1-\beta_t)(\tilde{g}_{t-1} - G(x_{t-1})) + \frac{\beta_t}{B_{t,1}} \sum_{i=1}^{B_{t,1}} \Big(g(x_t; \xi_t^{(i)}) - G(x_t) \Big) \\
&\quad + \frac{1-\beta_t}{B_{t,1}} \sum_{i=1}^{B_{t,1}} \Big(G(x_{t-1}) - G(x_t) + g(x_t; \xi_t^{(i)}) - g(x_{t-1}; \xi_t^{(i)}) \Big).
\end{aligned}$$

We apply this relation recursively down to $t = 1$ with $\tilde{g}_0 = 0$ and $\beta_1 = 1$ to yield

$$\tilde{g}_t - G(x_t) = \sum_{r=1}^{t} \bar{\beta}_{(r+1):t} \nu_r + \sum_{r=2}^{t} \bar{\beta}_{r:t} \omega_r,$$

where we denote $\omega_r = \frac{1}{B_{r,1}} \sum_{i=1}^{B_{r,1}} (G(x_{r-1}) - G(x_r) + g(x_r; \xi_r^{(i)}) - g(x_{r-1}; \xi_r^{(i)}))$. Taking norm and expectation on both sides yields

$$\mathbb{E}[\|\tilde{g}_t - G(x_t)\|] \leqslant \mathbb{E}\left[\left\| \sum_{r=1}^{t} \bar{\beta}_{(r+1):t} \nu_r \right\|^p \right]^{\frac{1}{p}} + \mathbb{E}\left[\left\| \sum_{r=2}^{t} \bar{\beta}_{r:t} \omega_r \right\|^p \right]^{\frac{1}{p}}. \tag{19}$$

Having bounded the first term in Lemma 2 (i.e. (18)), we now focus on the second term. Let $M_r := \bar{\beta}_{r:t} \omega_r$. Then for any $r \in [t]$, following a derivation similar to that of Lemma 2, we can verify that

$$\begin{aligned}
\mathbb{E}[M_r | \mathcal{F}_{r-1}] &= \frac{\bar{\beta}_{r:t}}{B_{r,1}} \sum_{i=1}^{B_{r,1}} \mathbb{E}\Big[g(x_r; \xi_r^{(i)}) - G(x_r) + G(x_{r-1}) - g(x_{r-1}; \xi_r^{(i)}) | \mathcal{F}_{r-1} \Big] \\
&= \frac{\bar{\beta}_{r:t}}{B_{r,1}} \sum_{i=1}^{B_{r,1}} \mathbb{E}\Big[g(x_r; \xi_r^{(i)}) - G(x_r) + G(x_{r-1}) - g(x_{r-1}; \xi_r^{(i)}) | x_{r-1}, x_r \Big] = 0,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}[\|M_r\|^p] \\
&\leqslant \frac{2\bar{\beta}_{r:t}^p}{B_{r,1}} \sum_{i=1}^{B_{r,1}} \left(\mathbb{E}\left[\mathbb{E}\left[\left\| g(x_{r-1}; \xi_{r-1}^{(i)}) - G(x_{r-1}) \right\|^p | x_{r-1} \right] \right] + \mathbb{E}\left[\mathbb{E}\left[\left\| g(x_r; \xi_r^{(i)}) - G(x_r) \right\|^p | x_r \right] \right] \right)
\end{aligned}$$

$$\leqslant 4\bar{\beta}_{r:t}^p V_g^p \leqslant +\infty.$$

where we use the fact that $\|a+b\|^p \leqslant 2(\|a\|^p + \|b\|^p)$ in the first inequality. Then, Lemma 10 provides us with the following inequality

$$\mathbb{E}\left[\left\|\sum_{r=2}^{t} \bar{\beta}_{r:t}\omega_r\right\|^p\right]^{\frac{1}{p}} \leqslant \mathbb{E}\left[2\sum_{r=2}^{t}\|\bar{\beta}_{r:t}\omega_r\|^p\right]^{\frac{1}{p}} = \left(2\sum_{r=2}^{t}\frac{\bar{\beta}_{r:t}^p \mathbb{E}[\|\sum_{i=1}^{B_{r,1}} C_r^i\|^p]}{B_{r,1}^p}\right)^{\frac{1}{p}}, \tag{20}$$

where we denote $C_r^i = g(x_r;\xi_r^{(i)}) - G(x_r) + G(x_{r-1}) - g(x_{r-1};\xi_r^{(i)})$. Proceeding analogously to the proof of Lemma 2, we use Lemma 10 again to bound $\mathbb{E}[\|\sum_{i=1}^{B_{r,1}} C_r^i\|^p]$. Specifically, let $M_r^i := C_r^i$ with $i = 1,\ldots,B_{r,1}$. Noting that the samples are drawn independently and x_{r-1} and x_r are $\sigma(M_r^1,\ldots,M_r^{i-1})$-measurable, it holds that $\mathbb{E}[M_r^i|M_r^1,\ldots,M_r^{i-1}] = 0$. Next, using the bounded p-th moment property, we derive

$$\begin{aligned}\mathbb{E}[\|M_r^i\|^p] &\leqslant 2\mathbb{E}\left[\mathbb{E}\left[\left\|g(x_{r-1};\xi_r^{(i)}) - G(x_{r-1})\right\|^p |x_{r-1}\right]\right] + 2\mathbb{E}\left[\mathbb{E}\left[\left\|g(x_r;\xi_r^{(i)}) - G(x_r)\right\|^p |x_r\right]\right] \\ &\leqslant 4V_g^p \leqslant +\infty.\end{aligned}$$

Then, Lemma 10 implies

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=1}^{B_{r,1}} C_r^i\right\|^p\right] &\leqslant 2\sum_{i=1}^{B_{r,1}}\mathbb{E}[\|C_r^i\|^p] \\ &\leqslant 4\sum_{i=1}^{B_{r,1}}\left(\mathbb{E}[\|g(x_r;\xi_r^{(i)}) - g(x_{r-1};\xi_r^{(i)})\|^p] + \mathbb{E}[\|G(x_r) - G(x_{r-1})\|^p]\right) \\ &\leqslant 4B_{r,1}(\ell_g^p\|x_r - x_{r-1}\|^p + \ell_G^p\|x_r - x_{r-1}\|^p) \leqslant 4\mu^p\rho_{r-1}^p B_{r,1}(\ell_g^p + \ell_G^p)\end{aligned}$$

Substituting this back into (20) yields $\mathbb{E}\left[\|\sum_{r=2}^t \bar{\beta}_{r:t}\omega_r\|^p\right]^{\frac{1}{p}} \leqslant 8\mu(\ell_g+\ell_G)(\sum_{r=2}^t \bar{\beta}_{r:t}^p\rho_{r-1}^p B_{r,1}^{1-p})^{\frac{1}{p}}$ Finally, combining this with the upper bound on $\mathbb{E}[\|\sum_{r=1}^t \bar{\beta}_{(r+1):t}\nu_r\|^p]^{\frac{1}{p}}$ established in (18) and substituting into (19), we obtain the desired bound for $\mathbb{E}[\|\tilde{g}_t - G(x_t)\|]$. We omit the proof for $\mathbb{E}[\|\nabla\tilde{g}_t - \nabla G(x_t)\|]$ as it follows analogous arguments. □

B.6 Proof of Theorem 3

Proof. With t-independent parameters, the bounds on $\mathbb{E}[\|\tilde{g}_t - G(x_t)\|]$ and $\mathbb{E}[\|\nabla\tilde{g}_t - \nabla G(x_t)\|]$ generated by NSPA-ST become

$$\begin{aligned}\mathbb{E}[\|\tilde{g}_t - G(x_t)\|] &\leqslant 2\beta V_g b_1^{\frac{1-p}{p}}\left(\sum_{r=1}^{t}(1-\beta)^{p(t-r)}\right)^{\frac{1}{p}} + 8\mu(\ell_g+\ell_G)\rho b_1^{\frac{1-p}{p}}\left(\sum_{r=2}^{t}(1-\beta)^{p(t-r+1)}\right)^{\frac{1}{p}}, \\ \mathbb{E}[\|\nabla\tilde{g}_t - \nabla G(x_t)\|] &\leqslant 2\gamma V_J b_2^{\frac{1-p}{p}}\left(\sum_{r=1}^{t}(1-\gamma)^{p(t-r)}\right)^{\frac{1}{p}} + 8\mu(L_g+L_G)\rho b_2^{\frac{1-p}{p}}\left(\sum_{r=2}^{t}(1-\gamma)^{p(t-r+1)}\right)^{\frac{1}{p}}.\end{aligned}$$

By an argument similar to the proof of Theorem 1, we obtain

$$\mathbb{E}[\|\tilde{g}_t - G(x_t)\|] \leqslant 2\beta^{\frac{p-1}{p}} V_g b_1^{\frac{1-p}{p}} + \frac{4(\ell_g+\ell_G)\rho b_1^{\frac{1-p}{p}}}{L\beta^{\frac{1}{p}}},$$

$$\mathbb{E}[\|\nabla \tilde{g}_t - \nabla G(x_t)\|] \leqslant 2\gamma^{\frac{p-1}{p}} V_J b_2^{\frac{1-p}{p}} + \frac{4(L_g + L_G)\rho b_2^{\frac{1-p}{p}}}{L\gamma^{\frac{1}{p}}}.$$

Lemma 1 with above inequalities and the choices of parameters provides

$$\begin{aligned}\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{8L(\Psi(x_1) - \Psi^*)}{\rho T} + 10\ell_f V_J \gamma^{\frac{p-1}{p}} b_2^{\frac{1-p}{p}} + 10\ell_G L_f V_g \beta^{\frac{p-1}{p}} b_1^{\frac{1-p}{p}} \\ + \frac{20\ell_f(L_g + L_G)\rho b_2^{\frac{1-p}{p}}}{L\gamma^{\frac{1}{p}}} + \frac{20\ell_G L_f(\ell_g + \ell_G)\rho b_1^{\frac{1-p}{p}}}{L\beta^{\frac{1}{p}}} + \frac{\rho}{4} \leqslant \epsilon.\end{aligned}$$

Thus, its sample complexity is $\sum_{t=1}^{T}(B_{t,1}+B_{t,2}) = \mathcal{O}(\epsilon^{-c-1}(\epsilon^{c-\frac{p}{p-1}}+\epsilon^{c-\frac{p}{p-1}})) = \mathcal{O}(\epsilon^{-\frac{2p-1}{p-1}})$, which completes the proof. □

B.7 Proof of Theorem 4

Proof. In this case, the estimate for $\mathbb{E}[\|\tilde{g}_t - G(x_t)\|]$ in Lemma 3 satisfies

$$\begin{aligned}\mathbb{E}[\|\tilde{g}_t - G(x_t)\|] &\leqslant 2V_g \left(\sum_{r=1}^{t} r^{-\frac{2p}{3}} \prod_{k=r+1}^{t}\left(1 - k^{-\frac{2}{3}}\right)\right)^{\frac{1}{p}} + 8\mu(\ell_g + \ell_G)\left(\sum_{r=2}^{t}(r-1)^{-\frac{2p}{3}} \prod_{k=r}^{t}\left(1 - k^{-\frac{2}{3}}\right)\right)^{\frac{1}{p}} \\ &\leqslant \left(2V_g + \frac{4(\ell_g + \ell_G)}{L}\right)\left(\sum_{r=1}^{t} r^{-\frac{2p}{3}} \prod_{k=r+1}^{t}\left(1 - k^{-\frac{2}{3}}\right)\right)^{\frac{1}{p}} \\ &\leqslant 2D\left(V_g + \frac{2(\ell_g + \ell_G)}{L}\right) t^{\frac{2-2p}{3p}},\end{aligned}$$

where we invoke Lemma 11 with $(a, d) = (\frac{2p}{3}, \frac{2}{3})$ in the last inequality. Similarly, the estimate for $\mathbb{E}[\|\nabla \tilde{g}_t - \nabla G(x_t)\|]$ satisfies $\mathbb{E}[\|\nabla \tilde{g}_t - \nabla G(x_t)\|] \leqslant 2D(V_g + 2(\ell_g + \ell_G)/L)t^{\frac{2-2p}{3p}}$. Substituting these bounds and the specific t-dependent parameters into Lemma 1 results in

$$\begin{aligned}\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{8L(\Psi(x_1) - \Psi^*)}{\sum_{t=1}^{T} t^{-\frac{2}{3}}} + \frac{\sum_{t=1}^{T} t^{-\frac{4}{3}}}{4\sum_{t=1}^{T} t^{-\frac{2}{3}}} \\ + \frac{10D\left(\ell_f V_g + \ell_G L_f V_J + 2(\ell_f + \ell_G L_f)(\ell_g + \ell_G)L^{-1}\right)\sum_{t=1}^{T} t^{-\frac{4p-2}{3p}}}{\sum_{t=1}^{T} t^{-\frac{2}{3}}}.\end{aligned}$$

Following a similar argument as in the proof of Theorem 2, we exploit the non-negativity of $(2-p)/(3p)$ to yield $\sum_{t=1}^{T} t^{-\frac{4p-2}{3p}} = \sum_{t=1}^{T} t^{-1} \cdot t^{\frac{2-p}{3p}} \leqslant T^{\frac{2-p}{3p}} \sum_{t=1}^{T} t^{-1}$. Based on above two inequalities and applying Lemma 8 with $a = \frac{2}{3}$ and $d = \frac{4}{3}$, we deduce

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{8L(\Psi(x_1) - \Psi^*)}{T^{\frac{1}{3}}} + \frac{1}{T^{\frac{1}{3}}} + \frac{20D\left(\ell_f V_g + \ell_G L_f V_J + 2(\ell_f + \ell_G L_f)(\ell_g + \ell_G)L^{-1}\right)\log(T)}{T^{\frac{2p-2}{3p}}}.$$

Let $\Delta_2 = 8L(\Psi(x_1) - \Psi^*) + 20D\left(\ell_f V_g + \ell_G L_f V_J + 2(\ell_f + \ell_G L_f)(\ell_g + \ell_G)L^{-1}\right) + 1$, we have $\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \Delta_2 \log(T)/T^{\frac{2p-2}{3p}}$ with $T \geqslant 3$. By a similar argument, with the setting $\mathcal{O}((\log(\epsilon^{-1})/\epsilon)^{\frac{3p}{2p-2}})$ and Lemma 9 with $(x, a, d) = (T, \frac{2p-2}{3p}, \frac{\Delta_2}{\epsilon})$, it suffices to obtain an ϵ-stationary point. The sample complexity is thus $\mathcal{O}((\log(\epsilon^{-1})/\epsilon)^{-\frac{3p}{2p-2}})$. Hiding logarithmic factors yields the desired result. □

C Proofs in Section 4

This subsection presents the missing proofs in Section 4. Before presenting the proof of Lemma 4, we first provide the descent property tailed for Algorithm 3.

Lemma 13 *Suppose Assumption 3 holds and $\mu = \frac{1}{2\ell_f L_G}$. Then for any $t \in [T]$ and $j \in [\tau_t - 1]$, it holds that*

$$\Psi(x_{t,j+1}) \leqslant \Psi(x_{t,j}) + 2\ell_f \|\tilde{g}_{t,j} - G(x_{t,j})\| + \frac{\rho_t}{2L_G} \|\nabla G(x_{t,j}) - \nabla \tilde{g}_{t,j}\| - \frac{\rho_t}{8\ell_f L_G} \left(\|\widetilde{\mathcal{G}}(x_{t,j})\| - \frac{1}{4}\rho_t \right).$$

Proof. By leveraging the ℓ_f-Lipschitz continuity of f and the L_G-smoothness of G, we observe that for any $y, z \in \mathbb{R}^n$,

$$\begin{aligned}
\Phi(y) = f(G(y)) &= f\left(G(z) + \nabla G(z)^\top (y - z)\right) + f(G(y)) - f\left(G(z) + \nabla G(z)^\top (y - z)\right) \\
&\leqslant f\left(G(z) + \nabla G(z)^\top (y - z)\right) + \ell_f \|G(y) - G(z) - \nabla G(z)^\top (y - z)\| \\
&\leqslant f\left(G(z) + \nabla G(z)^\top (y - z)\right) + \frac{\ell_f L_G}{2} \|y - z\|^2.
\end{aligned}$$

Then, we obtain

$$\begin{aligned}
\Psi(x_{t,j+1}) &= \Phi(x_{t,j+1}) + r(x_{t,j+1}) \\
&\leqslant f\left(G(x_{t,j}) + \nabla G(x_{t,j})^\top (x_{t,j+1} - x_{t,j})\right) + \frac{\ell_f L_G}{2} \|x_{t,j+1} - x_{t,j}\|^2 + r(x_{t,j+1}) \\
&= \underbrace{f\left(\tilde{g}_{t,j} + \nabla \tilde{g}_{t,j}^\top (x_{t,j+1} - x_{t,j})\right) + r(x_{t,j+1}) + \frac{1}{2\mu\kappa_{t,j+1}} \|x_{t,j+1} - x_{t,j}\|^2}_{T_4} \\
&\quad + \underbrace{f\left(G(x_{t,j}) + \nabla G(x_{t,j})^\top (x_{t,j+1} - x_{t,j})\right) - f\left(\tilde{g}_{t,j} + \nabla G(x_{t,j})^\top (x_{t,j+1} - x_{t,j})\right)}_{T_5} \\
&\quad + \underbrace{f\left(\tilde{g}_{t,j} + \nabla G(x_{t,j})^\top (x_{t,j+1} - x_{t,j})\right) - f\left(\tilde{g}_{t,j} + \nabla \tilde{g}_{t,j}^\top (x_{t,j+1} - x_{t,j})\right)}_{T_6} \\
&\quad + \underbrace{\frac{\ell_f L_G \mu \kappa_{t,j+1} - 1}{2\mu\kappa_{t,j+1}} \|x_{t,j+1} - x_{t,j}\|^2}_{T_7}.
\end{aligned} \tag{21}$$

According to the normalization step (6), it holds that

$$\begin{aligned}
T_4 &= f\left(\tilde{g}_{t,j} + \kappa_{t,j+1} \nabla \tilde{g}_{t,j}^\top (\tilde{x}_{t,j+1} - x_{t,j})\right) + r\left(x_{t,j} + \kappa_{t,j+1}(\tilde{x}_{t,j+1} - x_{t,j})\right) + \frac{\kappa_{t,j+1}}{2\mu} \|\tilde{x}_{t,j+1} - x_{t,j}\|^2 \\
&= f\left((1 - \kappa_{t,j+1})\tilde{g}_{t,j} + \kappa_{t,j+1}\left(\tilde{g}_{t,j} + \nabla \tilde{g}_{t,j}^\top (\tilde{x}_{t,j+1} - x_{t,j})\right)\right) \\
&\quad + r\left(x_{t,j} + \kappa_{t,j+1}(\tilde{x}_{t,j+1} - x_{t,j})\right) + \frac{\kappa_{t,j+1}}{2\mu} \|\tilde{x}_{t,j+1} - x_{t,j}\|^2 \\
&\leqslant (1 - \kappa_{t,j+1})\left(f\left(\tilde{g}_{t,j}\right) + r(x_{t,j})\right) \\
&\quad + \kappa_{t,j+1} \left(f\left(\tilde{g}_{t,j} + \nabla \tilde{g}_{t,j}^\top (\tilde{x}_{t,j+1} - x_{t,j})\right) + r\left(\tilde{x}_{t,j+1}\right) + \frac{1}{2\mu} \|\tilde{x}_{t,j+1} - x_{t,j}\|^2 \right),
\end{aligned}$$

where the inequality is by the convexity of f and r. The minimality of $\tilde{x}_{t,j+1}$, i.e.,

$$f\left(\tilde{g}_{t,j} + \nabla \tilde{g}_{t,j}^\top (\tilde{x}_{t,j+1} - x_{t,j})\right) + r\left(\tilde{x}_{t,j+1}\right) + \frac{1}{2\mu} \|\tilde{x}_{t,j+1} - x_{t,j}\|^2 \leqslant f\left(\tilde{g}_{t,j}\right) + r(x_{t,j}),$$

leads to

$$T_4 \leqslant f\left(\tilde{g}_{t,j}\right) + r(x_{t,j}) \leqslant f\left(G(x_{t,j})\right) + r(x_{t,j}) + f\left(\tilde{g}_{t,j}\right) - f\left(G(x_{t,j})\right) \\ \leqslant \Psi(x_{t,j}) + \ell_f \|\tilde{g}_{t,j} - G(x_{t,j})\|,$$

By the Lipschitz continuity of f, we obtain $T_5 \leqslant \ell_f \left\|\tilde{g}_{t,j} - G(x_{t,j})\right\|$, and

$$T_6 \leqslant \ell_f \| \left(\nabla G(x_{t,j}) - \nabla \tilde{g}_{t,j}\right)^\top \left(x_{t,j+1} - x_{t,j}\right) \| \leqslant \frac{\rho_t}{2L_G} \left\|\nabla G(x_{t,j}) - \nabla \tilde{g}_{t,j}\right\|,$$

where the equality is due to the inequality $\|x_{t,j+1} - x_{t,j}\| \leqslant \mu\rho_t$ and the parameter condition $\mu = \frac{1}{2\ell_f L_G}$. The term T_7 can be bounded by following an argument analogous to that used for term T_3 in Lemma 12, which yields

$$T_7 \leqslant -\frac{\rho_t}{8\ell_f L_G} \left(\|\widetilde{\mathcal{G}}(x_{t,j})\| - \frac{1}{4}\rho_t \right).$$

Substituting these bounds into (21) gives the desired result. □

Instead of relying on the non-expansiveness of the proximal operator as in the smooth case, we bound $\|\tilde{x}_{t,j+1} - \hat{x}_{t,j+1}\|$ by adapting the arguments from [52, Lemma 2] with some modifications. For any $y, z \in \mathbb{R}^d$, we introduce the following notation:

$$\widehat{F}(y; z) := f\left(G(z) + \nabla G(z)^\top (y - z)\right) + r(y), \qquad \widehat{F}_\mu(y; z) := \widehat{F}(y; z) + \frac{1}{2\mu}\|y - z\|^2,$$
$$\widetilde{F}(y; z) := f\left(\tilde{g} + \nabla \tilde{g}^\top (y - z)\right) + r(y), \qquad \widetilde{F}_\mu(y; z) := \widetilde{F}(y; z) + \frac{1}{2\mu}\|y - z\|^2,$$

where $\tilde{g}$ and $\nabla\tilde{g}$ are the stochastic estimates of $G(z)$ and $\nabla G(z)$, respectively. The bound for $\|\tilde{x}_{t,j+1} - \hat{x}_{t,j+1}\|$ is established in the following lemma.

Lemma 14 *Suppose Assumption 3 holds. Then for any $t \in [T]$ and $j \in [\tau_t - 1]$, it holds that*

$$\|\tilde{x}_{t,j+1} - \hat{x}_{t,j+1}\| \leqslant \sqrt{2\mu\ell_f \|\tilde{g}_{t,j} - G(x_{t,j})\|} + \sqrt{\frac{\mu\ell_f}{L_G}} \|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\| \\ + \sqrt{\frac{\mu^3 \ell_f L_G}{2}} \|\widetilde{\mathcal{G}}(x_{t,j})\| + \sqrt{\frac{\mu^3 \ell_f L_G}{2}} \left\|\mathcal{G}(x_{t,j})\right\|.$$

Proof. Since both f and r are convex, it follows that $\widehat{F}(x; x_{t,j})$ and $\widetilde{F}(x; x_{t,j})$ are convex and thus $\widehat{F}_\mu(x; x_{t,j})$ and $\widetilde{F}_\mu(x; x_{t,j})$ are $\frac{1}{\mu}$-strongly convex. By definitions (3) and (5), we obtain

$$\widehat{F}_\mu(\hat{x}_{t,j+1}; x_{t,j}) \leqslant \widehat{F}_\mu(\tilde{x}_{t,j+1}; x_{t,j}) - \frac{1}{2\mu}\|\tilde{x}_{t,j+1} - \hat{x}_{t,j+1}\|^2,$$
$$\widetilde{F}_\mu(\tilde{x}_{t,j+1}; x_{t,j}) \leqslant \widetilde{F}_\mu(\hat{x}_{t,j+1}; x_{t,j}) - \frac{1}{2\mu}\|\tilde{x}_{t,j+1} - \hat{x}_{t,j+1}\|^2.$$

Summing these two inequalities gives

$$\frac{1}{\mu}\|\tilde{x}_{t,j+1} - \hat{x}_{t,j+1}\|^2 \leqslant \underbrace{\widetilde{F}(\hat{x}_{t,j+1}; x_{t,j}) - \widehat{F}(\hat{x}_{t,j+1}; x_{t,j})}_{T_8} + \underbrace{\widehat{F}(\tilde{x}_{t,j+1}; x_{t,j}) - \widetilde{F}(\tilde{x}_{t,j+1}; x_{t,j})}_{T_9}. \tag{22}$$

For term T_8, we have

$$T_8 = f\left(\tilde{g}_{t,j} + \nabla\tilde{g}_{t,j}^\top (\hat{x}_{t,j+1} - x_{t,j})\right) - f\left(\tilde{g}_{t,j} + \nabla G(x_{t,j})^\top (\hat{x}_{t,j+1} - x_{t,j})\right)$$

$$
\begin{aligned}
&+ f\left(\tilde{g}_{t,j} + \nabla G(x_{t,j})^\top(\hat{x}_{t,j+1} - x_{t,j})\right) - f\left(G(x_{t,j}) + \nabla G(x_{t,j})^\top(\hat{x}_{t,j+1} - x_{t,j})\right) \\
&\leqslant \ell_f \|\left(\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\right)^\top(\hat{x}_{t,j+1} - x_{t,j})\| + \ell_f\|\tilde{g}_{t,j} - G(x_{t,j})\| \\
&\leqslant \ell_f\left(\|\tilde{g}_{t,j} - G(x_{t,j})\| + \frac{1}{2L_G}\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\|^2 + \frac{L_G}{2}\|\hat{x}_{t,j+1} - x_{t,j}\|^2\right),
\end{aligned}
$$

where the first inequality follows the Lipschitz continuity of f and the second inequality is due to Young's inequality. Similar argument leads to

$$
T_9 \leqslant \ell_f\left(\|\tilde{g}_{t,j} - G(x_{t,j})\| + \frac{1}{2L_G}\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\|^2 + \frac{L_G}{2}\|\tilde{x}_{t,j+1} - x_{t,j}\|^2\right).
$$

Plugging these two bounds into (22) yields

$$
\begin{aligned}
&\frac{1}{\mu}\|\tilde{x}_{t,j+1} - \hat{x}_{t,j+1}\|^2 \\
&\leqslant 2\ell_f\|\tilde{g}_{t,j} - G(x_{t,j})\| + \frac{\ell_f}{L_G}\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\|^2 + \frac{\ell_f L_G}{2}\|\tilde{x}_{t,j+1} - x_{t,j}\|^2 + \frac{\ell_f L_G}{2}\|\hat{x}_{t,j+1} - x_{t,j}\|^2 \\
&= 2\ell_f\|\tilde{g}_{t,j} - G(x_{t,j})\| + \frac{\ell_f}{L_G}\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\|^2 + \frac{\mu^2\ell_f L_G}{2}\|\widetilde{\mathcal{G}}(x_{t,j})\|^2 + \frac{\mu^2\ell_f L_G}{2}\|\mathcal{G}(x_{t,j})\|^2.
\end{aligned}
$$

It follows from the fact $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ with $a, b \geqslant 0$ that

$$
\begin{aligned}
\|\tilde{x}_{t,j+1} - \hat{x}_{t,j+1}\| \leqslant &\sqrt{2\mu\ell_f\|\tilde{g}_{t,j} - G(x_{t,j})\|} + \sqrt{\frac{\mu\ell_f}{L_G}}\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\| \\
&+ \sqrt{\frac{\mu^3\ell_f L_G}{2}}\|\widetilde{\mathcal{G}}(x_{t,j})\| + \sqrt{\frac{\mu^3\ell_f L_G}{2}}\|\mathcal{G}(x_{t,j})\|,
\end{aligned}
$$

which completes the proof. □

C.1 Proof of Lemma 4

Proof. By combining Lemma 14 and the fact $\|\mathcal{G}(x_{t,j})\| = \frac{1}{\mu}\|x_{t,j} - \hat{x}_{t,j+1}\| \leqslant \|\widetilde{\mathcal{G}}(x_{t,j})\| + \frac{1}{\mu}\|\tilde{x}_{t,j+1} - \hat{x}_{t,j+1}\|$, we obtain

$$
\rho_t\|\mathcal{G}(x_{t,j})\| \leqslant \sqrt{\frac{2\rho_t^2\ell_f}{u^2\mu}}\|\tilde{g}_{t,j} - G(x_{t,j})\|^{\frac{1}{2}} + \sqrt{\frac{\rho_t^2\ell_f}{u^2\mu L_G}}\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\| + \frac{v\rho_t}{u}\|\widetilde{\mathcal{G}}(x_{t,j})\|, \tag{23}
$$

where $u := (1 - \sqrt{\mu\ell_f L_G/2})$ and $v := (1 + \sqrt{\mu\ell_f L_G/2})$. Lemma 13 implies

$$
\rho_t\|\widetilde{\mathcal{G}}(x_{t,j})\| \leqslant 8\ell_f L_G(\Psi(x_{t,j}) - \Psi(x_{t,j+1})) + 16\ell_f^2 L_G\|\tilde{g}_{t,j} - G(x_{t,j})\| + 4\rho_t\ell_f\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\| + \frac{\rho_t^2}{4}.
$$

Substituting above inequality into (23) yields

$$
\begin{aligned}
\rho_t\|\mathcal{G}(x_{t,j})\| \leqslant &\frac{8v\ell_f L_G}{u}(\Psi(x_{t,j}) - \Psi(x_{t,j+1})) + \sqrt{\frac{2\rho_t^2\ell_f}{u^2\mu}}\|\tilde{g}_{t,j} - G(x_{t,j})\|^{\frac{1}{2}} \\
&+ \frac{16v\ell_f^2 L_G}{u}\|\tilde{g}_{t,j} - G(x_{t,j})\| + \left(\sqrt{\frac{\rho_t^2\ell_f}{u^2\mu L_G}} + \frac{4v\rho_t\ell_f}{u}\right)\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\| + \frac{v\rho_t^2}{4u}
\end{aligned}
$$

$$= 24\ell_f L_G(\Psi(x_{t,j}) - \Psi(x_{t,j+1})) + 4\rho_t \ell_f \sqrt{L_G}\|\tilde{g}_{t,j} - G(x_{t,j})\|^{\frac{1}{2}} + 48\ell_f^2 L_G\|\tilde{g}_{t,j} - G(x_{t,j})\| + (2\sqrt{2}+12)\ell_f\rho_t\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\| + \frac{3\rho_t^2}{4},$$

where the equality holds by the choice $\mu = \frac{1}{2\ell_f L_G}$, such that $u = \frac{1}{2}$ and $\frac{v}{u} = 3$. By taking the expectations of the above inequality and summing over $t = 1, \ldots, T$, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\rho_t\sum_{j=0}^{\tau_t-1}\mathbb{E}[\|\mathcal{G}(x_{t,j})\|] \\
&\leqslant 24\ell_f L_G(\Psi(x_{1,0}) - \Psi^*) + 4\ell_f\sqrt{L_G}\sum_{t=1}^{T}\rho_t\sum_{j=0}^{\tau_t-1}\mathbb{E}[\|\tilde{g}_{t,j} - G(x_{t,j})\|^{\frac{1}{2}}] + \frac{3\sum_{t=1}^{T}\tau_t\rho_t^2}{4} \\
&\quad + 48\ell_f^2 L_G\sum_{t=1}^{T}\sum_{j=0}^{\tau_t-1}\mathbb{E}[\|\tilde{g}_{t,j} - G(x_{t,j})\|] + 16\ell_f\sum_{t=1}^{T}\rho_t\sum_{j=0}^{\tau_t-1}\mathbb{E}[\|\nabla\tilde{g}_{t,j} - \nabla G(x_{t,j})\|].
\end{aligned}$$

Furthermore, using the inequality $\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \sum_{t=1}^{T}\rho_t\sum_{j=0}^{\tau_t-1}\mathbb{E}[\|\mathcal{G}(x_{t,j})\|]/\sum_{k=1}^{T}\tau_k\rho_k$ leads to the desired result. □

C.2 Proof of Theorem 5

Proof. By applying Jensen's inequality and Lemma 5, we have

$$\mathbb{E}[\|\tilde{g}_{t,0} - G(x_{t,0})\|^{\frac{1}{2}}] \leqslant \mathbb{E}[\|\tilde{g}_{t,0} - G(x_{t,0})\|]^{\frac{1}{2}} \leqslant \sqrt{2V_g}B_{t,1}^{-\frac{p-1}{2p}}. \tag{24}$$

Substituting the bound derived above and the results established in Lemma 5 into Lemma 4 (with $\tau_t = 1$) yields

$$\begin{aligned}
\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant{}& \frac{24\ell_f L_G(\Psi(x_{1,0}) - \Psi^*)}{\sum_{k=1}^{T}\rho_k} + \frac{4\ell_f\sqrt{2L_G V_g}\sum_{t=1}^{T}\rho_t B_{t,1}^{-\frac{p-1}{2p}}}{\sum_{k=1}^{T}\rho_k} + \frac{96\ell_f^2 L_G V_g\sum_{t=1}^{T}B_{t,1}^{-\frac{p-1}{p}}}{\sum_{k=1}^{T}\rho_k} \\
&+ \frac{32\ell_f V_J\sum_{t=1}^{T}\rho_t B_{t,2}^{-\frac{p-1}{p}}}{\sum_{k=1}^{T}\rho_k} + \frac{3\sum_{t=1}^{T}\rho_t^2}{4\sum_{k=1}^{T}\rho_k}.
\end{aligned} \tag{25}$$

Noting that the parameters ρ_t, $B_{t,1}$, and $B_{t,2}$ are independent of t, we then obtain

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{24\ell_f L_G(\Psi(x_{1,0}) - \Psi^*)}{\rho T} + \frac{4\ell_f\sqrt{2L_G V_g}}{b_1^{\frac{p-1}{2p}}} + \frac{96\ell_f^2 L_G V_g}{\rho b_1^{\frac{p-1}{p}}} + \frac{32\ell_f V_J}{b_2^{\frac{p-1}{p}}} + \frac{3\rho}{4} \leqslant \epsilon,$$

where the last inequality follows from the choices of ρ, T, b_1, and b_2. Hence, the sample complexities of $\tilde{g}_t$ and $\nabla\tilde{g}_t$ to find an ϵ-stationary point are

$$\begin{aligned}
\sum_{t=1}^{T}B_{t,1} &= \mathcal{O}\left(\ell_f L_G(\Psi(x_{1,0}) - \Psi^*)\epsilon^{-2}\cdot\left(\ell_f\sqrt{L_G V_g}\epsilon^{-1}\right)^{\frac{2p}{p-1}}\right) = \mathcal{O}\left(\epsilon^{-\frac{4p-2}{p-1}}\right), \\
\sum_{t=1}^{T}B_{t,2} &= \mathcal{O}\left(\ell_f L_G(\Psi(x_{1,0}) - \Psi^*)\epsilon^{-2}\cdot\left(\ell_f V_J\epsilon^{-1}\right)^{\frac{p}{p-1}}\right) = \mathcal{O}\left(\epsilon^{-\frac{3p-2}{p-1}}\right).
\end{aligned}$$

This completes the proof. □

C.3 Proof of Theorem 6

Proof. The upper bound on $\mathbb{E}[\|\mathcal{G}(\bar{x})\|]$ established in (25) still holds in this setting. Substituting the chosen parameters ρ_t, $B_{t,1}$, and $B_{t,2}$ yields

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{24\ell_f L_G(\Psi(x_{1,0}) - \Psi^*)}{\sum_{t=1}^T t^{-\frac{1}{2}}} + \frac{4\ell_f\sqrt{2L_G V_g}\left(bT^2\right)^{-\frac{p-1}{2p}}\sum_{t=1}^T t^{-\frac{1}{2}}}{\sum_{t=1}^T t^{-\frac{1}{2}}} + \frac{96\ell_f^2 L_G V_g T\left(bT^2\right)^{-\frac{p-1}{p}}}{\sum_{t=1}^T t^{-\frac{1}{2}}} + \frac{32\ell_f V_J (bT)^{-\frac{p-1}{p}}\sum_{t=1}^T t^{-\frac{1}{2}}}{\sum_{t=1}^T t^{-\frac{1}{2}}} + \frac{3\sum_{t=1}^T t^{-1}}{4\sum_{t=1}^T t^{-\frac{1}{2}}}.$$

By applying Lemma 8 (a) with $a = \frac{1}{2}$ and Lemma 8 (b), we further obtain

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{24\ell_f L_G(\Psi(x_{1,0}) - \Psi^*)}{T^{\frac{1}{2}}} + \frac{4\ell_f\sqrt{2L_G V_g}b^{-\frac{p-1}{2p}}}{T^{\frac{p-1}{p}}} + \frac{96\ell_f^2 L_G V_g b^{-\frac{p-1}{p}}}{T^{\frac{5p-4}{2p}}} + \frac{32\ell_f V_J b^{-\frac{p-1}{p}}}{T^{\frac{p-1}{p}}} + \frac{3\log(T)}{2T^{\frac{1}{2}}} \leqslant \frac{\Delta_3\log(T)}{T^{\frac{p-1}{p}}},$$

where $\Delta_3 := 24\ell_f L_G(\Psi(x_{1,0}) - \Psi^*) + 4\ell_f\sqrt{2L_G V_g}b^{-\frac{p-1}{2p}} + 96\ell_f^2 L_G V_g b^{-\frac{p-1}{p}} + 32\ell_f V_J b^{-\frac{p-1}{p}} + 3/2$. The condition $\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \epsilon$ is established by the setting $T = \mathcal{O}((\log(\epsilon^{-1})/\epsilon)^{\frac{p}{p-1}})$ and applying Lemma 9 with $(x, a, d) = (T, \frac{p-1}{p}, \frac{\Delta_3}{\epsilon})$. Furthermore, the sample complexities of $\tilde{g}_t$ and $\nabla\tilde{g}_t$ to find an ϵ-stationary point are

$$\sum_{t=1}^T B_{t,1} = bT^3 = \mathcal{O}\left(\left(\frac{\log(\epsilon^{-1})}{\epsilon}\right)^{\frac{3p}{p-1}}\right), \quad \sum_{t=1}^T B_{t,2} = bT^2 = \mathcal{O}\left(\left(\frac{\log(\epsilon^{-1})}{\epsilon}\right)^{\frac{2p}{p-1}}\right).$$

Hiding logarithmic factors completes the proof. □

C.4 Proof of Theorem 7

Proof. From Lemma 6, we use Jensen's inequality to obtain

$$\mathbb{E}[\|\tilde{g}_{t,j} - G(x_{t,j})\|^{\frac{1}{2}}] \leqslant \mathbb{E}[\|\tilde{g}_{t,j} - G(x_{t,j})\|]^{\frac{1}{2}} \leqslant \sqrt{8\mu\rho_t\tau_t^{\frac{1}{p}}(\ell_g + \ell_G)}S_{t,1}^{-\frac{p-1}{2p}} + \sqrt{2V_g}B_{t,1}^{-\frac{p-1}{2p}}. \tag{26}$$

Combining the results of Lemma 6 and (26) with Lemma 4 yields

$$\begin{aligned}\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant{}& \frac{24\ell_f L_G(\Psi(x_{1,0}) - \Psi^*)}{\sum_{k=1}^T \tau_k\rho_k} + \frac{3\sum_{t=1}^T \tau_t\rho_t^2}{4\sum_{k=1}^T \tau_k\rho_k} \\ &+ \frac{4\ell_f\sqrt{L_G}\sum_{t=1}^T\left(\sqrt{8\mu(\ell_g + \ell_G)}\rho_t^{\frac{3}{2}}\tau_t^{\frac{2p+1}{2p}}S_{t,1}^{-\frac{p-1}{2p}} + \sqrt{2V_g}\rho_t\tau_t B_{t,1}^{-\frac{p-1}{2p}}\right)}{\sum_{k=1}^T \tau_k\rho_k} \\ &+ \frac{48\ell_f^2 L_G\sum_{t=1}^T\left(8\mu(\ell_g + \ell_G)\rho_t\tau_t^{\frac{p+1}{p}}S_{t,1}^{-\frac{p-1}{p}} + 2V_g\tau_t B_{t,1}^{-\frac{p-1}{p}}\right)}{\sum_{k=1}^T \tau_k\rho_k} \\ &+ \frac{16\ell_f\sum_{t=1}^T\left(8\mu(L_g + L_G)\rho_t^2\tau_t^{\frac{p+1}{p}}S_{t,2}^{-\frac{p-1}{p}} + 2V_J\rho_t\tau_t B_{t,2}^{-\frac{p-1}{p}}\right)}{\sum_{k=1}^T \tau_k\rho_k}.\end{aligned} \tag{27}$$

Given that ρ_t, τ_t, $B_{t,1}$, $B_{t,2}$, $S_{t,1}$, and $S_{t,2}$ are t-independent and $\mu = \frac{1}{2\ell_f L_G}$, we obtain

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{24\ell_f L_G(\Psi(x_{1,0}) - \Psi^*)}{\tau\rho T} + \frac{8\sqrt{\ell_f(\ell_g+\ell_G)}\rho^{\frac{1}{2}}\tau^{\frac{1}{2p}}}{s_1^{\frac{p-1}{2p}}} + \frac{4\ell_f\sqrt{2L_G V_g}}{b_1^{\frac{p-1}{2p}}}$$
$$+ \frac{192\ell_f(\ell_g+\ell_G)\tau^{\frac{1}{p}}}{s_1^{\frac{p-1}{p}}} + \frac{96\ell_f^2 L_G V_g}{\rho b_1^{\frac{p-1}{p}}} + \frac{64(L_g+L_G)\rho\tau^{\frac{1}{p}}}{L_G s_2^{\frac{p-1}{p}}} + \frac{32\ell_f V_J}{b_2^{\frac{p-1}{p}}} + \frac{3\rho}{4}.$$

The choices of T, τ, ρ, b_1, b_2, s_1, and s_2 ensure that $\mathbb{E}[\|\bar{x}\|] \leqslant \epsilon$. Therefore, the sample complexities of $\tilde{g}_t$ and $\nabla\tilde{g}_t$ to find an ϵ-stationary point are

$$\sum_{t=1}^{T}(B_{t,1}+\tau_t S_{t,1}) = \mathcal{O}\Bigg(\ell_f L_G(\Psi(x_{1,0})-\Psi^*)\epsilon^{-1}\cdot\left(\ell_f\sqrt{L_G V_g}\epsilon^{-1}\right)^{\frac{2p}{p-1}}$$
$$+ \ell_f L_G(\Psi(x_{1,0})-\Psi^*)\epsilon^{-2}\cdot\left(\sqrt{\ell_f(\ell_g+\ell_G)}\epsilon^{-\frac{p+1}{2p}}\right)^{\frac{2p}{p-1}}\Bigg) = \mathcal{O}\left(\epsilon^{-\frac{3p-1}{p-1}}\right),$$

$$\sum_{t=1}^{T}(B_{t,2}+\tau_t S_{t,2}) = \mathcal{O}\Bigg(\ell_f L_G(\Psi(x_{1,0})-\Psi^*)\epsilon^{-1}\cdot\left(\ell_f V_J\epsilon^{-1}\right)^{\frac{p}{p-1}}$$
$$+ \ell_f L_G(\Psi(x_{1,0})-\Psi^*)\epsilon^{-2}\cdot\left((L_g+L_G)L_G^{-1}\epsilon^{-\frac{1}{p}}\right)^{\frac{p}{p-1}}\Bigg) = \mathcal{O}\left(\epsilon^{-\frac{2p-1}{p-1}}\right).$$

This completes the proof. □

C.5 Proof of Theorem 8

Proof. Combining (27) with the choices of μ, ρ_t, τ_t, $B_{t,1}$, $B_{t,2}$, $S_{t,1}$, and $S_{t,2}$ gives

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{24\ell_f L_G(\Psi(x_{1,0})-\Psi^*)}{T} + \frac{8\sqrt{\ell_f(\ell_g+\ell_G)}s^{-\frac{p-1}{2p}}\sum_{t=1}^{T}t^{\frac{1-p}{2p}}}{T^{\frac{3p+1}{2p}}} + \frac{4\ell_f\sqrt{2L_G V_g}b^{-\frac{p-1}{2p}}}{T}$$
$$+ \frac{192\ell_f(\ell_g+\ell_G)\sum_{t=1}^{T}t^{\frac{1}{p}}s^{-\frac{p-1}{p}}}{T^{\frac{2p+1}{p}}} + \frac{96\ell_f^2 L_G V_g\sum_{t=1}^{T}tb^{-\frac{p-1}{p}}}{T^3} + \frac{3\sum_{t=1}^{T}t^{-1}}{4T}$$
$$+ \frac{64(L_g+L_G)s^{-\frac{p-1}{p}}\sum_{t=1}^{T}t^{\frac{1-p}{p}}}{L_G T^{\frac{2p-1}{p}}} + \frac{32\ell_f V_J b^{-\frac{p-1}{p}}}{T^{\frac{2p-2}{p}}}.$$

By applying Lemma 8 (a) with $a = \frac{p-1}{2p}$ and $a = \frac{p-1}{p}$, and Lemma 8 (b), we obtain

$$\mathbb{E}[\|\mathcal{G}(\bar{x})\|] \leqslant \frac{24\ell_f L_G(\Psi(x_{1,0})-\Psi^*)}{T} + \frac{16p\sqrt{\ell_f(\ell_g+\ell_G)}s^{-\frac{p-1}{2p}}}{(p+1)T} + \frac{4\ell_f\sqrt{2L_G V_g}b^{-\frac{p-1}{2p}}}{T}$$
$$+ \frac{192\ell_f(\ell_g+\ell_G)s^{-\frac{p-1}{p}}}{T} + \frac{96\ell_f^2 L_G V_g b^{-\frac{p-1}{p}}}{T} + \frac{3\log(T)}{2T}$$
$$+ \frac{64p\ell_f(L_g+L_G)s^{-\frac{p-1}{p}}}{L_G T^{\frac{2p-2}{p}}} + \frac{32\ell_f V_J b^{-\frac{p-1}{p}}}{T^{\frac{2p-2}{p}}} \leqslant \frac{\Delta_4\log(T)}{T^{\frac{2p-2}{p}}},$$

where

$$\Delta_4 := 24\ell_f L_G(\Psi(x_{1,0}) - \Psi^*) + \frac{16p\sqrt{\ell_f(\ell_g+\ell_G)}s^{-\frac{p-1}{2p}}}{p+1} + 4\ell_f\sqrt{2L_G V_g}b^{-\frac{p-1}{2p}}$$
$$+ 192\ell_f(\ell_g+\ell_G)s^{-\frac{p-1}{p}} + 96\ell_f^2 L_G V_g b^{-\frac{p-1}{p}} + \frac{3}{2} + \frac{64p\ell_f(L_g+L_G)s^{-\frac{p-1}{p}}}{L_G} + 32\ell_f V_J b^{-\frac{p-1}{p}}.$$

Furthermore, setting $\mathcal{O}((\log(\epsilon^{-1})/\epsilon)^{\frac{p}{2p-2}})$ and applying Lemma 9 with $(x,a,d) = (T, \frac{2p-2}{p}, \frac{\Delta_4}{\epsilon})$ guarantees an ϵ-stationary point. Hence, the sample complexities of $\tilde{g}_t$ and $\nabla\tilde{g}_t$ to find an ϵ-stationary point are

$$\sum_{t=1}^{T}(B_{t,1}+\tau_t S_{t,1}) = bT^{\frac{3p-1}{p-1}} + sT^{\frac{p+1}{p-1}}\sum_{t=1}^{T}t^{-1} = \mathcal{O}\left(T^{\frac{3p-1}{p-1}} + T^{\frac{p+1}{p-1}}\cdot\log(T)\right)$$
$$= \mathcal{O}\left(\left(\frac{\log(\epsilon^{-1})}{\epsilon}\right)^{\frac{p(3p-1)}{2(p-1)^2}}\right),$$
$$\sum_{t=1}^{T}(B_{t,2}+\tau_t S_{t,1}) = bT^3 + sT\sum_{t=1}^{T}t^{-1} = \mathcal{O}\left(T^3 + T^2\cdot\log(T)\right) = \mathcal{O}\left(\left(\frac{\log(\epsilon^{-1})}{\epsilon}\right)^{\frac{3p}{2p-2}}\right).$$

This completes the proof. □

D Additional experimental results

Table 3: Relative error under different noise intensities (constant parameter set).

Noise Intensity	Tailed Index	NSPA-PM	NSPA-ST	NSPA-B	NSPA-SP
$\sigma = 2$	$\alpha = 1.8$	8.1847×10^{-2}	7.9937×10^{-2}	4.5969×10^{-2}	4.7215×10^{-2}
	$\alpha = 1.5$	9.2694×10^{-2}	8.4137×10^{-2}	4.7396×10^{-2}	4.8779×10^{-2}
	$\alpha = 1.2$	1.0454×10^{-1}	1.0017×10^{-1}	5.3555×10^{-2}	5.4806×10^{-2}
$\sigma = 1$	$\alpha = 1.8$	4.7849×10^{-2}	4.3555×10^{-2}	2.4479×10^{-2}	2.6121×10^{-2}
	$\alpha = 1.5$	5.1436×10^{-2}	4.8362×10^{-2}	2.4304×10^{-2}	2.6944×10^{-2}
	$\alpha = 1.2$	5.9412×10^{-2}	5.4992×10^{-2}	2.8134×10^{-2}	3.0505×10^{-2}
$\sigma = 0.5$	$\alpha = 1.8$	2.9355×10^{-2}	2.7299×10^{-2}	1.2490×10^{-2}	1.4944×10^{-2}
	$\alpha = 1.5$	3.2494×10^{-2}	2.7263×10^{-2}	1.2973×10^{-2}	1.5177×10^{-2}
	$\alpha = 1.2$	3.4972×10^{-2}	3.3093×10^{-2}	1.5236×10^{-2}	1.8204×10^{-2}
$\sigma = 0.1$	$\alpha = 1.8$	2.0719×10^{-2}	1.6131×10^{-2}	2.8354×10^{-3}	6.7675×10^{-3}
	$\alpha = 1.5$	2.0620×10^{-2}	1.7814×10^{-2}	2.8923×10^{-3}	6.5436×10^{-3}
	$\alpha = 1.2$	2.1588×10^{-2}	2.1320×10^{-2}	3.2478×10^{-3}	6.2898×10^{-3}
$\sigma = 0.01$	$\alpha = 1.8$	2.1780×10^{-2}	1.5756×10^{-2}	4.7331×10^{-4}	4.9581×10^{-3}
	$\alpha = 1.5$	1.9300×10^{-2}	1.7400×10^{-2}	4.3459×10^{-4}	4.7890×10^{-3}
	$\alpha = 1.2$	1.9465×10^{-2}	1.6839×10^{-2}	4.8348×10^{-4}	5.4690×10^{-3}

Table 4: Relative error under different noise intensities (decaying parameter set).

Noise Intensity	Tailed Index	NSPA-PM	NSPA-ST	NSPA-B	NSPA-SP
$\sigma = 2$	$\alpha = 1.8$	6.4254×10^{-2}	6.4234×10^{-2}	4.4326×10^{-2}	4.2005×10^{-2}
	$\alpha = 1.5$	7.0808×10^{-2}	7.0787×10^{-2}	4.5952×10^{-2}	4.4333×10^{-2}
	$\alpha = 1.2$	8.6138×10^{-2}	8.6103×10^{-2}	5.2217×10^{-2}	5.0700×10^{-2}
$\sigma = 1$	$\alpha = 1.8$	2.9652×10^{-2}	2.9762×10^{-2}	2.2899×10^{-2}	2.1193×10^{-2}
	$\alpha = 1.5$	3.2735×10^{-2}	3.2720×10^{-2}	2.3456×10^{-2}	2.2260×10^{-2}
	$\alpha = 1.2$	4.0548×10^{-2}	4.0537×10^{-2}	2.6533×10^{-2}	2.5274×10^{-2}
$\sigma = 0.5$	$\alpha = 1.8$	1.2526×10^{-2}	1.2442×10^{-2}	1.1250×10^{-2}	1.0448×10^{-2}
	$\alpha = 1.5$	1.3809×10^{-2}	1.4158×10^{-2}	1.1817×10^{-2}	1.0916×10^{-2}
	$\alpha = 1.2$	1.7618×10^{-2}	1.7442×10^{-2}	1.3557×10^{-2}	1.2484×10^{-2}
$\sigma = 0.1$	$\alpha = 1.8$	1.8449×10^{-3}	2.3046×10^{-3}	2.3213×10^{-3}	2.0454×10^{-3}
	$\alpha = 1.5$	1.9445×10^{-3}	2.2261×10^{-3}	2.3909×10^{-3}	2.1532×10^{-3}
	$\alpha = 1.2$	2.3248×10^{-3}	2.3844×10^{-3}	2.7998×10^{-3}	2.4717×10^{-3}
$\sigma = 0.01$	$\alpha = 1.8$	1.1145×10^{-3}	1.3342×10^{-3}	4.3166×10^{-4}	2.3908×10^{-4}
	$\alpha = 1.5$	1.1045×10^{-3}	1.4424×10^{-3}	4.6448×10^{-4}	2.4832×10^{-4}
	$\alpha = 1.2$	1.0818×10^{-3}	1.3888×10^{-3}	5.4002×10^{-4}	2.7878×10^{-4}

References

[1] Krishnakumar Balasubramanian, Saeed Ghadimi, and Anthony Nguyen. Stochastic multilevel composition optimization algorithms with level-independent convergence rates. *SIAM Journal on Optimization*, 32(2):519–544, 2022.

[2] Nail Bashirov, Alexander Gasnikov, and Aleksandr Lobanov. Zeroth-order methods for non-smooth stochastic problems under heavy-tailed noise. *Optimization Methods and Software*, pages 1–26, 2026.

[3] Amir Beck. *First-Order Methods in Optimization*. SIAM, 2017.

[4] Jian-Feng Cai, Meng Huang, Dong Li, and Yang Wang. Solving phase retrieval with random initial guess is nearly as good as by spectral initialization. *Applied and Computational Harmonic Analysis*, 58:60–84, 2022.

[5] Jian-Feng Cai, Yu Long, Ruixue Wen, and Jiaxi Ying. A fast and provable algorithm for sparse phase retrieval. *arXiv preprint arXiv:2309.02046*, 2023.

[6] Emmanuel J Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[7] Vasileios Charisopoulos, Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Composite optimization for robust blind deconvolution. *arXiv preprint arXiv:1901.01624*, 2019.

[8] 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.

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

[10] 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, 2021.

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

[12] Christoph Dann, Gerhard Neumann, and Jan Peters. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research*, 15(1):809–883, 2014.

[13] Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

[14] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558, 2019.

[15] 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.

[16] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. John Wiley & Sons, 1991.

[17] 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.

[18] Iuri Frosio and N Alberto Borghese. Statistical based impulsive noise removal in digital radiography. *IEEE Transactions on Medical Imaging*, 28(1):3–16, 2008.

[19] 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.

[20] Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

[21] Chunhao Han, Xiao Wang, Pengxiang Xu, and Jin Zhang. Non-convex stochastic compositional optimization under heavy-tailed noise. *https://optimization-online.org/?p=33308*, 2026.

[22] Chuan He and Zhaosong Lu. Accelerated stochastic first-order method for convex optimization under heavy-tailed noise. *arXiv preprint arXiv:2510.11676*, 2025.

[23] 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.

[24] 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.

[25] 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.

[26] Lingzi Jin and Xiao Wang. Stochastic nested primal-dual method for nonconvex constrained composition optimization. *Mathematics of Computation*, 94(351):305–358, 2025.

[27] 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.

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

[29] Yin Liu and Sam Davanloo Tajbakhsh. Stochastic composition optimization of functions without Lipschitz continuous gradient. *Journal of Optimization Theory and Applications*, 198(1):239–289, 2023.

[30] Zhuanghua Liu, Luo Luo, and Bryan Kian Hsiang Low. Gradient-free methods for nonconvex nonsmooth stochastic compositional optimization. In *Advances in Neural Information Processing Systems*, volume 37, pages 45438–45461, 2024.

[31] Zijian Liu. Clipped gradient methods for nonsmooth convex optimization under heavy-tailed noise: A refined analysis. *arXiv preprint arXiv:2512.23178*, 2025.

[32] 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.

[33] Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. *arXiv preprint arXiv:2303.12277*, 2023.

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

[35] 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.

[36] 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.

[37] R Tyrrell Rockafellar. *Convex Analysis*, volume 28. Princeton university press, 1997.

[38] R Tyrrell Rockafellar. Coherent approaches to risk in optimization under uncertainty. In *OR Tools and Applications: Glimpses of Future Technologies*, pages 38–61. Informs, 2007.

[39] Andrzej Ruszczyński. Advances in risk-averse optimization. In *Theory Driven by Influential Applications*, pages 168–190. INFORMS, 2013.

[40] Umut Şimşekli, Mert Gürbüzbalaban, Thanh Huy Nguyen, Gaël Richard, and Levent Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019.

[41] Umut Şimşekli, Levent Sagun, and Mert Gürbüzlalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[42] Tao Sun, Xinwang Liu, and Kun Yuan. Revisiting gradient normalization and clipping for nonconvex SGD under heavy-tailed noise: Necessity, sufficiency, and acceleration. *Journal of Machine Learning Research*, 26(237):1–42, 2025.

[43] Quoc Tran-Dinh, Nhan Pham, and Lam Nguyen. Stochastic Gauss-Newton algorithms for nonconvex compositional optimization. In *International Conference on Machine Learning*, pages 9572–9582. PMLR, 2020.

[44] Rasul Tutunov, Minne Li, Alexander I Cowen-Rivers, Jun Wang, and Haitham Bou-Ammar. Compositional adam: An adaptive compositional solver. *arXiv preprint arXiv:2002.03755*, 2020.

[45] 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.

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

[47] Yu Xia and Zhiqiang Xu. Sparse phase retrieval via PhaseLiftOff. *IEEE Transactions on Signal Processing*, 69:2129–2143, 2021.

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

[49] 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.

[50] 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.

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

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

[53] Zhe Zhang and Guanghui Lan. Optimal methods for convex nested stochastic composite optimization: Z. zhang, g. lan. *Mathematical Programming*, 212(1):1–48, 2025.