Journal of Machine Learning Research 27 (2026) 1-62 Submitted 12/25; Revised 8/26; Published 9/26

Stochastic Bilevel Optimization with Heavy-Tailed Noise

Zhuanghua Liu LIUZHUANGHUA9@GMAIL.COM
Department of Computer Science, National University of Singapore
21 Lower Kent Ridge Rd, Singapore

Luo Luo* LUOLUO@FUDAN.EDU.CN
School of Data Science, Fudan University
Shanghai, China
and
Shanghai Key Laboratory for Contemporary Applied Mathematics
Shanghai, China

Editor: Shiqian Ma

Abstract

This paper considers the smooth bilevel optimization in which the lower-level problem is strongly convex and the upper-level problem is possibly nonconvex. We focus on the stochastic setting where the algorithm can access the unbiased stochastic gradient evaluation with heavy-tailed noise, which is prevalent in many machine learning applications, such as training large language models and reinforcement learning. We propose a nested-loop normalized stochastic bilevel approximation (N^{2}SBA) for finding an ϵ-stationary point with the stochastic first-order oracle (SFO) complexity of $\tilde{\mathcal{O}}\big(\kappa^{\frac{7p-3}{p-1}}\sigma^{\frac{p}{p-1}}\epsilon^{-\frac{4p-2}{p-1}}\big)$, where κ is the condition number, $p \in (1,2]$ is the order of central moment for the noise, and σ is the noise level. Furthermore, we specialize our idea to solve the nonconvex-strongly-concave minimax optimization problem, achieving an ϵ-stationary point with the SFO complexity of $\tilde{\mathcal{O}}\big(\kappa^{\frac{2p-1}{p-1}}\sigma^{\frac{p}{p-1}}\epsilon^{-\frac{3p-2}{p-1}}\big)$. All the above upper bounds match the best-known results under the special case of the bounded variance setting, i.e., $p = 2$. We also conduct the numerical experiments to show the empirical superiority of proposed methods.

Keywords: bilevel optimization, stochastic optimization, heavy-tailed noise

1. Introduction

We consider the following smooth bilevel optimization problem

$$\begin{aligned} &\min_{x\in\mathbb{R}^{d_x}} \varphi(x) \coloneqq f(x, y^*(x)), \\ &\text{s.t. } y^*(x) \coloneqq \arg\min_{y\in\mathbb{R}^{d_y}} g(x,y), \end{aligned} \tag{1}$$

where both the upper-level function and the lower-level function are jointly continuously differentiable and have the form of

$$f(x,y) \coloneqq \mathbb{E}[F(x,y;\xi)] \qquad \text{and} \qquad g(x,y) \coloneqq \mathbb{E}[G(x,y;\zeta)] \tag{2}$$

∗. The corresponding author

©2026 Zhuanghua Liu and Luo Luo.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-3100.html.

where $\xi \sim \mathcal{D}_f$ and $\zeta \sim D_g$ are the stochastic indices. In particular, we focus on the nonconvex-strongly-convex setting, i.e., the lower-level function $g(x, y)$ is smooth and strongly convex with respect to y, while the upper-level function f is smooth but possibly nonconvex. This formulation is ubiquitous in a wide range of machine learning applications, including reinforcement learning (Sutton and Barto, 1998; Li, 2017; Hong et al., 2023), hyperparameter optimization (Franceschi et al., 2018; Pedregosa, 2016; Feurer and Hutter, 2019), meta learning (Finn et al., 2017; Fallah et al., 2020; Rajeswaran et al., 2019; Ji et al., 2022), and neural architecture search (Liu et al., 2019; Zoph and Le, 2016; Zhang et al., 2021).

There has been increasing interest in developing stochastic algorithms for large-scale bilevel optimization under the bounded variance assumption (Hao et al., 2024; Gong et al., 2024a,b; Ghadimi and Wang, 2018; Chen et al., 2021, 2022; Hong et al., 2023; Khanduri et al., 2021; Ji et al., 2021; Yang et al., 2021, 2022; Shen and Chen, 2023; Kwon et al., 2023; Chen et al., 2025; Chen and Zhang, 2025; Ji and Liang, 2023; Zhang et al., 2025a; Huang, 2024; Yao et al., 2025). The early methods construct the estimator for the hyper-gradient of the objective (Ghadimi and Wang, 2018; Hong et al., 2023; Ji et al., 2021; Huang et al., 2025; Yang et al., 2022; Wang et al., 2024; Gong et al., 2024a,b; Hao et al., 2024), which typically requires iterating with the (stochastic) Hessian-vector product (HVP) oracle to avoid directly computing the inverse Hessian. However, accessing the HVP oracle may be expensive in many real-world applications (Sow et al., 2022; Song et al., 2020; Finn et al., 2017; Nichol et al., 2018). To address this issue, the recent works introduce the penalized formulation for the bilevel problem to design the fully first-order methods, which efficiently achieves an approximate first-order stationary point without any second-order or HVP oracle calls (Kwon et al., 2023; Chen et al., 2025; Liu et al., 2022; Shen and Chen, 2023).

It is worth noting that the bounded variance assumption used in existing analyses for stochastic bilevel optimization is restrictive and may not hold in practice. The empirical studies on many real-world applications, such as image classification (Simsekli et al., 2019), large language model (LLM) training (Zhang et al., 2020; Ahn et al., 2024), and reinforcement learning (Garg et al., 2021) have shown that the stochastic gradient typically follows the heavy-tailed distributions. These observations motivate us to adopt the weaker p-th bounded central moment (p-BCM) assumption on the noise of the stochastic gradient, i.e., for all $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$, it holds that

$$\mathbb{E}[\|\nabla F(x, y; \xi) - \nabla f(x, y)\|^p] \leq \sigma_f^p \qquad \text{and} \qquad \mathbb{E}[\|\nabla G(x, y; \zeta) - \nabla g(x, y)\|^p] \leq \sigma_g^p$$

for some σ_f, $\sigma_g > 0$ and $p \in (1, 2]$ is the order of central moment for the noise. In the special case of $p = 2$, the p-BCM assumption reduces to the standard bounded variance assumption (Nemirovski et al., 2009; Ghadimi and Lan, 2013). In the case of $p \in (1, 2)$, the variance may be unbounded, leading to heavy-tailed noise distributions such as the Lévy α-stable distribution (Lévy, 1925; Mandelbrot, 1960). While the stochastic minimization problem under heavy-tailed has been extensively studied in recent years (Gorbunov et al., 2020, 2022; Sadiev et al., 2023; Gorbunov et al., 2024; Zhang et al., 2020; Nguyen et al., 2023; Cutkosky and Mehta, 2020; Liu and Zhou, 2023; Hübler et al., 2024; Liu and Zhou, 2025; Sun et al., 2024; He et al., 2025), the stochastic bilevel optimization remains unexplored beyond the bounded variance setting.

In this paper, we propose an efficient stochastic first-order method called nested-loop normalized stochastic bilevel approximation (N^2SBA) for nonconvex-strongly-convex bilevel

Table 1: We compare our proposed N^{2}SBA with existing methods by showing the upper complexity bounds for achieving an ϵ-stationary point in the stochastic nonconvex-strongly-convex bilevel optimization problem (1), where κ is the condition number and $\sigma := \max\{\sigma_f, \sigma_g\}$. The second column indicates the oracles used in the algorithm, i.e., the stochastic Hessian-vector product (SHVP) and the stochastic first-order oracle (SFO).

Methods	Oracle	Complexity	p
BSA (Ghadimi and Wang, 2018)	SFO/SHVP	$\mathcal{O}\left(\frac{\kappa^9\sigma^3}{\epsilon^6}\right)$	2
TTSA (Hong et al., 2023)[§]	SFO/SHVP	$\mathcal{O}\left(\frac{\mathrm{poly}(\kappa,\sigma)}{\epsilon^5}\right)$	2
stocBiO (Ji et al., 2021)	SFO/SHVP	$\mathcal{O}\left(\frac{\kappa^9\sigma^2}{\epsilon^4}\right)$	2
F^{2}SA (Kwon et al., 2023)[§]	SFO	$\tilde{\mathcal{O}}\left(\frac{\mathrm{poly}(\kappa,\sigma)}{\epsilon^7}\right)$	2
F^{2}BSA (Chen et al., 2025)[†]	SFO	$\tilde{\mathcal{O}}\left(\frac{\kappa^{11}\sigma^2}{\epsilon^6}\right)$	2
N^{2}SBA (Theorem 3)	SFO	$\tilde{\mathcal{O}}\left(\frac{\kappa^{\frac{7p-3}{p-1}}\sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{4p-2}{p-1}}}\right)$	$(1, 2]$

[§] We use the notation $\mathrm{poly}(\kappa,\sigma)$ in the complexity bounds of TTSA and F^{2}BSA since the dependency on the condition number and the noise level has not been explicitly presented by the analyses of Hong et al. (2023) and Kwon et al. (2023).

[†] The original analysis of Chen et al. (2025, Theorem 4.2) show that F^{2}SBA attains an ϵ-stationary point with the SFO complexity of $\tilde{\mathcal{O}}\left(\kappa^{12}\sigma^2\epsilon^{-6}\right)$. In fact, this result can be improved to $\tilde{\mathcal{O}}\left(\kappa^{11}\sigma^2\epsilon^{-6}\right)$. Please refer to Appendix E for details.

optimization with heavy-tailed noise. Our algorithm employs a nested-loop structure such that the outer loop can be regarded as inexact normalized stochastic gradient descent on the surrogate nonconvex function, while the inner loop solves the subproblems by clipped stochastic gradient descent. We show that our N^{2}SBA can achieve an ϵ-stationary point with the stochastic first-order oracle (SFO) complexity of $\tilde{\mathcal{O}}\big(\kappa^{\frac{7p-3}{p-1}}\sigma^{\frac{p}{p-1}}\epsilon^{-\frac{4p-2}{p-1}}\big)$ in expectation, where κ denotes the condition number and $\sigma = \max\{\sigma_f, \sigma_g\}$. We further establish the high probability convergence guarantee for our method, matching the complexity of the in-expectation result up to the logarithmic factor. For the special case of the bounded variance setting (i.e., $p = 2$), our result matches the state-of-the-art SFO complexity of $\tilde{\mathcal{O}}(\kappa^{11}\sigma^2\epsilon^{-6})$ established by Chen et al. (2025). In addition, we provide a lower bound with condition number dependence for stochastic bilevel optimization under heavy-tailed noise, which indicates such problem is indeed more difficult than the single level one. We compare the result of the proposed N^{2}SBA with related work in Table 1.

Table 2: We compare our proposed N^{2}SGDA with existing methods by showing the upper complexity bounds for achieving an ϵ-stationary point of $\Phi(x) \coloneqq \max_{y\in\mathbb{R}^{d_y}} f(x,y)$ for the stochastic nonconvex-strongly-concave minimax optimization problem (3). Here, we denote $\sigma \coloneqq \sigma_f$ for the ease of presentation.

Methods	Oracle	Complexity	p
SGDmax (Jin et al., 2020)	SFO	$\tilde{\mathcal{O}}\left(\frac{\kappa^3\sigma^2}{\epsilon^4}\right)$	2
SGDA (Lin et al., 2020)	SFO	$\mathcal{O}\left(\frac{\kappa^3\sigma^2}{\epsilon^4}\right)$	2
N^{2}SGDA (Theorem 11)	SFO	$\tilde{\mathcal{O}}\left(\frac{\kappa^{\frac{2p-1}{p-1}}\sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}}\right)$	(1,2]

We also study the stochastic nonconvex-strongly-concave minimax optimization problem

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} f(x,y) \coloneqq \mathbb{E}[F(x,y;\xi)], \tag{3}$$

which can be regarded as the special case of the bilevel problem (1) by taking $G = -F$ in equation (2). This formulation covers various machine learning applications, including generative adversarial networks (GANs) (Goodfellow et al., 2020), robust statistics (Xu et al., 2009; Shafieezadeh Abadeh et al., 2015), online learning (Cesa-Bianchi and Lugosi, 2006), and adversarial robustness (Madry et al., 2018). Most existing studies on stochastic minimax optimization with heavy-tailed noise focus on the convex–concave setting (Gorbunov et al., 2022; Sadiev et al., 2023; Gorbunov et al., 2023). Gorbunov et al. (2022) and Sadiev et al. (2023) investigated problems under the more general star-cocoercive condition, which is different from our nonconvex–strongly-concave setting. Following the idea of N^{2}SBA, we develop nested-loop normalized stochastic gradient descent ascent (N^{2}SGDA) for stochastic nonconvex-strongly-concave minimax optimization with heavy-tailed noise. We show that N^{2}SGDA can achieve an ϵ-stationary point of the function $\Phi(x) \coloneqq \max_{y\in\mathbb{R}^{d_y}} f(x,y)$ with the SFO complexity of $\tilde{\mathcal{O}}\big(\kappa^{\frac{2p-1}{p-1}}\sigma^{\frac{p}{p-1}}\epsilon^{-\frac{3p-2}{p-1}}\big)$ in expectation (with the high probability). This result is near-optimal to σ and ϵ, since it matches the lower bound of $\Omega\big(\sigma^{\frac{p}{p-1}}\epsilon^{-\frac{3p-2}{p-1}}\big)$ up to logarithmic factor for finding an ϵ-stationary point of stochastic nonconvex minimization problem under the p-BCM noise (Zhang et al., 2020; Liu and Zhou, 2025). We compare the result of proposed N^{2}SBA with related work in Table 2.

Paper Organization In Section 2, we introduce preliminaries for our problem setting. In Section 3, we propose N^{2}SBA for solving stochastic nonconvex-strongly-convex bilevel optimization problem with heavy-tailed noise and establish both in-expectation and high-probability theortical guarantees. In Section 4, we propose the N^{2}SGDA method for solving stochastic nonconvex-strongly-concave minimax optimization problem that obtains the near-optimal SFO complexity with respect to the accuracy and noise level. In Section 5, we conduct numerical experiments to validate the effectiveness of our methods. Finally, we conclude our work and outline several potential directions for future directions in Section 6.

2. Preliminaries

In this section, we formalize the problem setting for our stochastic bilevel optimization problem. We begin by introducing the standard assumptions on smoothness and lower boundedness, following (Ghadimi and Wang, 2018; Ji et al., 2021; Chen et al., 2024).

Assumption 2.1 *We suppose the bilevel optimization problem (1) holds that*

(a) the upper-level function $f(x, y)$ is C_f-Lipschitz in y, L_f-gradient Lipschitz, and twice continuously differentiable;

(b) the lower-level function $g(x, y)$ is L_g-gradient Lipschitz, ρ_g-Hessian Lipschitz, and μ-strongly convex in y;

(c) the objective $\varphi(x)$ is lower bounded, i.e., $\inf_{x\in\mathbb{R}^{d_x}} \varphi(x) > -\infty$.

Under Assumption 2.1, we define largest smoothness constant and the condition number of the problem (1) as follows.

Definition 2.1 *Under Assumption 2.1, we define the largest smoothness constant and the condition number of problem (1) as $\ell := \max\{C_f, L_f, L_g, \rho_g\}$ and $\kappa := \ell/\mu$, respectively.*

Assumption 2.1 also indicates the hyper-gradient $\nabla\varphi(x)$ has the following closed-form expression (Ghadimi and Wang, 2018).

Proposition 2.1 (Ghadimi and Wang (2018, Lemma 2.2)) *Under Assumption 2.1, the function φ in problem (1) is differentiable and the hyper-gradient $\nabla\varphi(x)$ has the closed-form expression*

$$\nabla\varphi(x) = \nabla_x f(x, y^*(x)) - \nabla^2_{xy} g(x, y^*(x))(\nabla^2_{yy} g(x, y^*(x)))^{-1}\nabla_y f(x, y^*(x)),$$

where $y^*(x) = \arg\min_{y\in\mathbb{R}^{d_y}} g(x, y)$ is uniquely defined. Furthermore, the hyper-gradient $\nabla\varphi(x)$ is L_φ-Lipschitz continuous with $L_\varphi = \mathcal{O}(\kappa^3\ell)$.

We target to find the ϵ-stationary point of φ, which is defined as follows.

Definition 2.2 *Given a differentiable function $\varphi(x) : \mathbb{R}^d \to \mathbb{R}$, we call the point $\hat{x} \in \mathbb{R}^d$ is an ϵ-stationary point of $\varphi(x)$ if it holds $\|\nabla\varphi(\hat{x})\| \le \epsilon$.*

We impose the p-th bounded central moment (p-BCM) assumption (Zhang et al., 2020) on the gradient noise for the upper-level and lower-level functions.

Assumption 2.2 *For given $x \in \mathbb{R}^{d_x}$, $y \in \mathbb{R}^{d_y}$, we suppose the algorithm can draw the stochastic indices $\xi \sim \mathcal{D}_f$ and $\zeta \sim \mathcal{D}_g$ and access the stochastic first-order oracle $\nabla F(x, y; \xi)$ and $\nabla G(x, y; \zeta)$ such that*

$$\mathbb{E}[\nabla F(x, y; \xi)] = \nabla f(x, y), \quad \mathbb{E}[\|\nabla F(x, y; \xi) - \nabla f(x, y)\|^p] \le \sigma_f^p,$$
$$\mathbb{E}[\nabla G(x, y; \zeta)] = \nabla g(x, y), \quad \mathbb{E}[\|\nabla G(x, y; \zeta) - \nabla g(x, y)\|^p] \le \sigma_g^p,$$

where $p \in (1, 2]$ and $\sigma_f, \sigma_g > 0$.

Additionally, we introduce the notation $\sigma := \max\{\sigma_f, \sigma_g\}$ for the stochastic bilevel optimization problem (1). We remark that Assumption 2.2 with $p = 2$ aligns with the classical bounded variance condition, while the variance of $\nabla F(x, y; \xi)$ and $\nabla G(x, y; \zeta)$ are possibly unbounded in the case of $p \in (1, 2)$ (Lévy, 1925; Mandelbrot, 1960).

3. Nested-loop Normalized Stochastic Bilevel Approximation

We propose nested-loop normalized stochastic bilevel approximation (N^{2}SBA) in Algorithm 1, where we define the clipping operator as

$$\mathrm{clip}(g,\tau)=\begin{cases}\min\left\{1,\dfrac{\tau}{\|g\|}\right\}g, & g\neq 0,\\ 0, & g=0,\end{cases}$$

for $g\in\mathbb{R}^{d_y}$ and $\tau>0$. The remainder of this section first introduces the intuition of our algorithm design, then provides the convergence guarantees both in expectation and with the high probability.

3.1 The Algorithm Design

We introduce the design of our N^{2}SBA by starting with the penalized reformulation (Kwon et al., 2023) for problem (1), i.e.,

$$\min_{x\in\mathbb{R}^{d_x}}\mathcal{L}_\lambda^*(x)\coloneqq\min_{y\in\mathbb{R}^{d_y}}\left\{f(x,y)+\lambda\left(g(x,y)-\min_{z\in\mathbb{R}^{d_y}}g(x,z)\right)\right\}. \tag{4}$$

It is worth noting that the gradient of $\mathcal{L}_\lambda^*$ has the closed-form expression that consists of the first-order information of f and g, i.e.,

$$\nabla\mathcal{L}_\lambda^*(x)=\nabla_x f(x,y_\lambda^*(x))+\lambda(\nabla_x g(x,y_\lambda^*(x))-\nabla_x g(x,y^*(x))),$$

where $y_\lambda^*(x)\coloneqq\arg\min_{y\in\mathbb{R}^{d_y}}f(x,y)+\lambda(g(x,y)-g^*(x))$ and $g^*(x)=g(x,y^*(x))$. In addition, taking $\lambda=\Theta(\kappa^3\ell\epsilon^{-1})$ guarantees that any ϵ-stationary point of the function $\mathcal{L}_\lambda^*$ corresponds to an $\mathcal{O}(\epsilon)$-stationary point of the objective φ (Kwon et al., 2023, Lemma 3.1). Therefore, it is natural to solve the bilevel optimization problem (1) by using first-order methods to address the surrogate problem (4). However, existing stochastic first-order bilevel optimization methods are based on the framework of stochastic gradient descent (Kwon et al., 2023; Chen et al., 2025; Liu et al., 2022; Shen and Chen, 2023), whose convergence guarantees require the assumption of bounded variance that does not hold in our p-BCM setting (Zhang et al., 2020, Remark 1).

In our proposed N^{2}SBA, we address the heavy-tailed noise by solving the surrogate problem (4) in the framework of inexact normalized stochastic gradient descent. Specifically, we first solve the sub-problems

$$\min_{y\in\mathbb{R}^{d_x}}f(x_t,y)+\lambda g(x_t,y)\qquad\text{and}\qquad\min_{z\in\mathbb{R}^{d_x}}g(x_t,z) \tag{5}$$

to approximate $y_\lambda^*(x_t)$ and $y^*(x_t)$ by y_t and z_t, respectively. In the views of $y_t\approx y_\lambda^*(x_t)$ and $z_t\approx y^*(x_t)$, we then estimate $\nabla\mathcal{L}_\lambda^*(x_t)$ by the mini-batch stochastic gradient of the function $\mathcal{L}_\lambda(x,y,z)\coloneqq f(x,y)+\lambda\left(g(x,y)-g(x,z)\right)$ with respect to x, i.e.,

$$g_{x,t}=\frac{1}{M}\sum_{i=1}^{M}\left(\nabla_x F(x_t,y_t;\xi_{t,i})+\lambda(\nabla_x G(x_t,y_t;\zeta_{t,i})-\nabla_x G(x_t,z_t;\zeta_{t,i}))\right),$$

Algorithm 1 Nested-loop Normalized Stochastic Bilevel Approximation (N^{2}SBA)

1: **Input:** initial points $x_0 \in \mathbb{R}^{d_x}$, $\hat{y}_{0,0} = \hat{z}_{0,0} \in \mathbb{R}^{d_y}$; penalized parameter $\lambda > 0$; step-sizes $\eta_x > 0, \{\eta_{y,t} > 0\}_{t=0}^{T-1}, \{\eta_{z,t} > 0\}_{t=0}^{T-1}$; sequences of clipping parameters $\{\tau_{t,k}\}_{k=0}^{K-1}$, $\{\tau'_{t,k}\}_{k=0}^{K-1}$; iteration numbers $T, K > 0$; batch size $M > 0$
2: **for** $t = 0, \ldots, T-1$
3: **for** $k = 0, \ldots, K-1$
4: draw $\xi'_t \sim \mathcal{D}_f$, $\zeta'_t \sim \mathcal{D}_g$, and $\zeta_t \sim \mathcal{D}_g$
5: $g_{y,t,k} = \nabla_y F(x_t, \hat{y}_{t,k}; \xi'_t) + \lambda \nabla_y G(x_t, \hat{y}_{t,k}; \zeta'_t)$
6: $\hat{y}_{t,k+1} = \hat{y}_{t,k} - \eta_{y,t} \mathrm{clip}(g_{y,t,k}, \tau'_{t,k})$
7: $g_{z,t,k} = \lambda \nabla_y G(x_t, \hat{z}_{t,k}; \zeta_t)$
8: $\hat{z}_{t,k+1} = \hat{z}_{t,k} - \eta_{z,t} \mathrm{clip}(g_{z,t,k}, \tau_{t,k})$
9: **end for**
10: $y_t = \hat{y}_{t,K}, \quad \hat{y}_{t+1,0} = \hat{y}_{t,K}$
11: $z_t = \hat{z}_{t,K}, \quad \hat{z}_{t+1,0} = \hat{z}_{t,K}$
12: draw $\xi_{t,i} \sim \mathcal{D}_f$ and $\zeta_{t,i} \sim \mathcal{D}_g$ for all $i \in [M]$
13: $g_{x,t} = \frac{1}{M} \sum_{i=1}^{M} (\nabla_x F(x_t, y_t; \xi_{t,i}) + \lambda(\nabla_x G(x_t, y_t; \zeta_{t,i}) - \nabla_x G(x_t, z_t; \zeta_{t,i})))$
14: $x_{t+1} = x_t - \eta_x \cdot \frac{g_{x,t}}{\|g_{x,t}\|}$
15: **end for**
16: draw $\hat{x}$ from $\{x_t\}_{t=1}^{T}$ uniformly
17: **Return:** $\hat{x}$

where $\xi_{t,i} \sim \mathcal{D}_f$ and $\zeta_{t,i} \sim \mathcal{D}_g$ for all $i \in [M]$. Note that the p-BCM condition on the stochastic gradients of f and g (Assumption 2.2) implies the stochastic gradients of the objectives in sub-problems (5) and the function $\mathcal{L}_\lambda$ also has the p-th bounded central moments, which typically requires the steps of clipping or normalization to guarantee the convergence of the first-order methods (Zhang et al., 2020; Sadiev et al., 2023; Hübler et al., 2024). Therefore, our N^{2}SBA (Algorithm 1) solves the strongly-convex sub-problems (5) by performing the clipped stochastic gradient descent (Line 3–9) and minimizing the nonconvex function $\mathcal{L}_\lambda^*$ by iterating with its inexact normalized stochastic gradient $g_{x,t}/\|g_{x,t}\|$ (Line 12–14).

3.2 Convergence Analysis

In this subsection, we present the complexity analysis of our N^{2}SBA for solving the stochastic nonconvex-strongly-convex bilevel optimization problem with heavy-tailed noise. We use notations $\mathcal{O}(\cdot)$, $\Theta(\cdot)$, $\tilde{\mathcal{O}}(\cdot)$, and $\tilde{\Theta}(\cdot)$ to simply the presentation of some parameter settings, and the corresponding explicit expressions can be found in appendix.

As discussed in Section 3.1, our objective is to identify an ϵ-stationary point of $\mathcal{L}_\lambda^*(x)$ with $\lambda = \Theta(\kappa^3 \ell \epsilon^{-1})$, which corresponds to an $\mathcal{O}(\epsilon)$-stationary point of $\varphi(x)$. Hence, we first

establish the in-expectation result for our algorithm, which starts from the following key lemma with respect to the function $\nabla\mathcal{L}_\lambda^*(x)$.

Lemma 1 *Under Assumptions 2.1 and 2.2, running N^{2}SBA (Algorithm 1) with $\lambda \geq 2L_f/\mu$ holds that*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}_\lambda^*(x_t)\|] \leq & \frac{\mathbb{E}[\mathcal{L}_\lambda^*(x_0)-\mathcal{L}_\lambda^*(x_T)]}{\eta_x T}+\frac{4\sigma_f+8\lambda\sigma_g}{M^{\frac{p-1}{p}}}+\sum_{t=0}^{T-1}\frac{4\lambda L_g\mathbb{E}[\|y_t-y_\lambda^*(x_t)\|]}{T}\\ &+\sum_{t=0}^{T-1}\frac{2\lambda L_g\mathbb{E}[\|z_t-y^*(x_t)\|]}{T}+\frac{D_3\eta_x}{2},\end{aligned} \tag{6}$$

where $D_3=\mathcal{O}(\ell\kappa^3)$.

In order to find $\mathcal{O}(\epsilon)$-stationary point of $\mathcal{L}_\lambda^*(x)$, we have to bound the terms $\mathbb{E}[\|y_t-y_\lambda^*(x_t)\|]$ and $\mathbb{E}[\|z_t-y^*(x_t)\|]$ on the right-hand side of inequality equation (6). Recall that the points $y_\lambda^*(x_t)$ and $y^*(x_t)$ are the solutions of the convex minimization problems in equation (5). Therefore, we introduce the following result for using the clipped stochastic gradient descent to solve the stochastic convex problem with heavy-tailed noise.

Lemma 2 *Assume the function $h(y)=\mathbb{E}[H(y;\nu)]$ is ℓ_h-smooth and μ_h-strongly convex with the stochastic index $\nu\sim\mathcal{D}_h$, and the stochastic gradient estimator $\nabla H(y;\nu)$ satisfies the conditions $\mathbb{E}[\nabla H(y;\nu)]=\nabla h(y)$ and $\mathbb{E}[\|\nabla H(y;\nu)-\nabla h(y)\|^p]\leq\sigma_h^p$ for some $\sigma_h>0$. We run the clipped stochastic gradient descent*

$$\begin{cases}\nu_k\sim\mathcal{D}_h\\ \hat{y}_{k+1}=\hat{y}_k-\eta_y\mathrm{clip}(\nabla H(\hat{y}_k;\nu_k),\tau_k)\end{cases} \tag{7}$$

for $k=0,1,\ldots,K$ with

$$\tau_k=\tilde{\Theta}\left(\frac{\hat{R}\exp(-\eta_y\mu_h(1+k/2))}{\eta_y}\right),\ \eta_y=\tilde{\Theta}\left(\min\left\{\frac{1}{\ell_h},\frac{\ln B_K}{\mu_h(K+1)}\right\}\right),\ B_K=\tilde{\mathcal{O}}\left(1+\frac{\mu_h^2K^{\frac{2(p-1)}{p}}\hat{R}^2}{\sigma_h^2}\right).$$

Then for all $\hat{R}^2\geq\|\hat{y}_0-y^*\|^2$, we have

$$\mathbb{E}[\|\hat{y}_K-y^*\|^2]\leq\tilde{\mathcal{O}}\left(\hat{R}^2\exp\left(-\frac{K\mu_h}{\ell_h}\right)+\frac{\sigma_h^2}{\mu_h^2K^{\frac{2(p-1)}{p}}}+\frac{\hat{R}^2}{K^2}\right).$$

By appropriate parameter settings, we can apply Lemma 2 to show that both $\mathbb{E}\left[\|y_t-y_\lambda^*(x_t)\|\right]$ and $\mathbb{E}\left[\|z_t-y^*(x_t)\|\right]$ are upper bounded by $\mathcal{O}\left(\epsilon^2/(\ell^2\kappa^3)\right)$ for all $t=0,\ldots,T-1$ (Lemma 24 in Appendix B.1.2). Combining with the result of Lemma 1, we derive the following in-expectation convergence of our N^{2}SBA for stochastic bilevel optimization problem.

Theorem 3 *Under Assumption 2.1 and 2.2, we run Algorithm 1 (N^{2}SBA) with*

$$\eta_x = \frac{\epsilon}{\ell\kappa^3}, \quad \eta_{y,t} = \tilde{\Theta}\left(\min\left\{\frac{1}{(L_f+\lambda L_g)}, \frac{1}{\lambda\mu K}\right\}\right), \quad \eta_{z,t} = \tilde{\Theta}\left(\min\left\{\frac{1}{\lambda L_g}, \frac{1}{\lambda\mu K}\right\}\right),$$
$$T = \mathcal{O}\left(\frac{\Delta\ell\kappa^3}{\epsilon^2}\right), \quad M = \mathcal{O}\left(\frac{\ell^{\frac{p}{p-1}}\kappa^{\frac{3p}{p-1}}\sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{2p}{p-1}}}\right), \quad K = \tilde{\mathcal{O}}\left(\frac{\ell^{\frac{p}{p-1}}\kappa^{\frac{4p}{p-1}}\sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{2p}{p-1}}} + \frac{\ell^2\kappa^3 R_y}{\epsilon^2}\right),$$
$$\tau_{t,k} = \tilde{\Theta}\left(\frac{\exp(-\eta_{z,t}\lambda\mu k)R_z}{\eta_{z,t}}\right), \ \tau'_{t,k} = \tilde{\Theta}\left(\frac{\exp(-\eta_{y,t}\lambda\mu k)R_y}{\eta_{y,t}}\right), \ \lambda = \max\left\{\frac{\kappa}{R_0}, \frac{\ell\kappa^2}{\Delta}, \frac{\ell\kappa^3}{\epsilon}\right\},$$

where we set

$$R_y^2 \coloneqq \mathcal{O}\left(R_0^2 + \frac{\epsilon^4}{\ell^4\kappa^6} + \frac{\epsilon^2}{\ell^2\kappa^4}\right), \quad R_z^2 \coloneqq \mathcal{O}\left(R_0^2 + \frac{\epsilon^4}{\ell^4\kappa^6} + \frac{\epsilon^2}{\ell^2\kappa^4}\right)$$
$$R_0^2 \geq \|\hat{y}_{0,0} - y^*(x_0)\|^2, \quad \Delta = \varphi(x_0) - \inf_{x\in\mathbb{R}^{d_x}}\varphi(x).$$

Then we have $\mathbb{E}[\|\nabla\varphi(\hat{x})\|] = \mathcal{O}(\epsilon)$. In particular, the overall SFO calls is upper bounded by

$$\tilde{\mathcal{O}}\left(\frac{\Delta\ell^{\frac{2p-1}{p-1}}\kappa^{\frac{7p-3}{p-1}}\sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{4p-2}{p-1}}} + \frac{\Delta\ell^3\kappa^6 R_0}{\epsilon^4}\right).$$

It is worth noting that our analysis is carried out under the general p-BCM noise condition (Assumption 2.2). In the special case of $p = 2$, it reduces to the standard bounded-variance setting, then Theorem 3 yields an SFO complexity of $\tilde{\mathcal{O}}(\kappa^{11}\sigma^2\epsilon^{-6})$ which aligns with the result of the best-known fully first-order methods (Chen et al., 2025).

The above in-expectation results guarantee the small gradient norm by taking the average on the outputs of running N^{2}SBA sufficient times, while the results of the high-probability guarantee for a single run are more desirable in practice. The following theorem establishes a high-probability guarantee for our N^{2}SBA algorithm.

Theorem 4 *Under Assumption 2.1 and 2.2, we run Algorithm 1 (N^{2}SBA) with*

$$\eta_x = \sqrt{\frac{\Delta}{\ell\kappa^3 T}}, \quad \eta_{y,t} = \tilde{\Theta}\left(\min\left\{\frac{1}{L_f+\lambda L_g}, \frac{1}{\lambda\mu K}\right\}\right), \eta_{z,t} = \tilde{\Theta}\left(\min\left\{\frac{1}{\lambda L_g}, \frac{1}{\lambda\mu K}\right\}\right),$$
$$M = \mathcal{O}\left(\frac{\ell^{\frac{p}{p-1}}\sigma^{\frac{p}{p-1}}\kappa^{\frac{3p}{p-1}}}{\epsilon^{\frac{2p}{p-1}}}\right), \quad K = \tilde{\mathcal{O}}\left(\frac{\ell^{\frac{p}{p-1}}\kappa^{\frac{4p}{p-1}}\sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{2p}{p-1}}}\right), \quad \lambda = \max\left\{\frac{\kappa}{\sqrt{R}}, \frac{\ell\kappa^2}{\Delta}, \frac{\ell\kappa^3}{\epsilon}\right\},$$
$$\tau_{t,k} = \tilde{\Theta}\left(\frac{\exp(-\eta_{z,t}\lambda\mu k)R_z}{\eta_{z,t}}\right), \quad \tau'_{t,k} = \tilde{\Theta}\left(\frac{\exp(-\eta_{y,t}\lambda\mu k)R_y}{120\eta_{y,t}}\right), \quad T = \tilde{\mathcal{O}}\left(\frac{\Delta\ell\kappa^3}{\epsilon^2}\right),$$

where R_y^2, R_z^2, R_0^2, and Δ follow the settings of Theorem 3. Then we have $\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\varphi(x_t)\| = \mathcal{O}(\epsilon)$ with probability at least $1-\delta$ for all $\delta \in (0,1)$. In particular, the overall SFO calls is upper bounded by[1]

$$\tilde{\mathcal{O}}\left(\frac{\Delta\ell^{\frac{2p-1}{p-1}}\kappa^{\frac{7p-3}{p-1}}\sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{4p-2}{p-1}}}\right).$$

1. The notations $\tilde{\Theta}(\cdot)$ and $\tilde{\mathcal{O}}(\cdot)$ in parameters setting hide the logarithmic factors including $\log(1/\delta)$.

We remark that our SFO complexity for the high-probability result in Theorem 4 matches that of the in-expectation counterpart in Theorem 3, up to the logarithmic factor. It also nearly matches the best-known in-expectation SFO complexity bounds under the bounded variance assumption (i.e., Assumption 2.2 with $p = 2$) (Chen et al., 2025).

3.3 The Lower Complexity Bound

In this subsection, we provide a lower bound for solving smooth nonconvex-strongly-convex bilevel optimization problems. We use $\mathcal{F}^{\text{nc-sc}}(C_f, L_f, L_g, \rho_g, \mu, \Delta)$ to denote the set of all bilevel optimization problems satisfying Assumption 2.1 and

$$\varphi(x_0) - \inf_{x\in\mathbb{R}^{d_x}} \varphi(x) \le \Delta,$$

where $x_0 \in \mathbb{R}^{d_x}$ is the initial point of the algorithm.

In our lower bound analyses, our construction focus on bilevel problems with quadratic lower-level functions, which are formally defined as follows.

Definition 5 *Given $C_f, L_f, L_g > 0$, $\mu \in (0, L_g]$ and $\Delta > 0$, we denote bilevel problems with quadratic lower-level functions as $\mathcal{F}^{\text{nc-scq}}(C_f, L_f, L_g, \mu, \Delta)$, which is a subset of nonconvex-strongly-convex problems $\mathcal{F}^{\text{nc-sc}}(C_f, L_f, L_g, 0, \mu, \Delta)$. In particular, the lower-level function of a problem in $\mathcal{F}^{\text{nc-scq}}(C_f, L_f, L_g, \mu, \Delta)$ takes the form of*

$$g(x, y) = \frac{1}{2} y^\top H y + x^\top J y + b^\top y,$$

where $\mu I \preceq H \preceq L_g I$ and $-L_g I \preceq J \preceq L_g I$.

We then introduce the class of stochastic first-order oracle algorithms as follows.

Definition 6 *A stochastic first-order algorithm A consists of a distribution $\mathbb{P}_r$ over a measurable set $\mathcal{R}$ and a sequence of measurable mappings $\{A_t\}_{t\in\mathbb{N}}$ such that A_t takes in the first $t-1$ oracle responses and the random seed $r \in \mathcal{R}$ to produce the t-th query w_t, such that $w_0 = (0, 0)$ and*

$$w_t = A_t(r, \mathbb{O}^{\text{sfo}}(w_0), \ldots, \mathbb{O}^{\text{sfo}}(w_{t-1})), \quad \forall t \in \mathbb{N}_+, \tag{8}$$

where $\mathbb{O}^{\text{sfo}}(x, y) = (\hat{\nabla} F(x, y; \xi), \hat{\nabla} G(x, y; \zeta))$ satisfying the p-BCM condition in Assumption 2.2. We denote $\mathcal{A}^{\text{sfo}}$ as the set of all randomized algorithms that follow the protocol (8).

We follow the construction of Chen and Zhang (2025) to establish the following lower bound for stochastic bilevel optimization under heavy-tailed noise.

Theorem 7 *Given $L_f, L_g, \mu, \sigma, \Delta, C_f > 0$ such that $\mu \in (0, L_g]$ and $C_f \ge \sqrt{13 L_f \Delta}$, the bilevel problem instance $(f, g) \in \mathcal{F}^{\text{nc-scq}}(C_f, L_f, L_g, \mu, \Delta)$ equipped with a stochastic first-order oracle (SFO) satisfying the p-BCM noise in Assumption 2.2, such that finding an expected ϵ-stationary point of the hyper-objective $\varphi(x) = f(x, z^*(x))$ with $\epsilon \le \min(\sigma\kappa_g/92, \kappa_g\sqrt{L_f\Delta}/312)$, any stochastic first-order algorithm $A \in \mathcal{A}^{\text{sfo}}$ requires at least*

$$\Omega\left(\frac{\Delta L_f \sigma^{\frac{p}{p-1}} \kappa_g^{\frac{3p-2}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}}\right) \tag{9}$$

SFO calls, where $\kappa_g = L_g/\mu$ is the condition number of the lower level problem.

Note that the optimal SFO complexity for finding an ϵ-stationary point in the single-level stochastic optimization problem with heavy-tailed noise is $\mathcal{O}(\Delta L_f \sigma^{\frac{p}{p-1}} \epsilon^{-\frac{3p-2}{p-1}})$ (Liu and Zhou, 2025). Therefore, the additional dependence on κ_g in our SFO lower bound (9) indicates the bilevel problem is indeed more challenging. In the special case of $p = 2$, the lower bound (9) reduces to $\Omega(\Delta L_f \sigma^2 \kappa_g^4 \epsilon^{-4})$, matching the state-of-the-art SFO lower bound for stochastic bilevel optimization problems under the bounded-variance setting (Chen and Zhang, 2025).

4. Application to Nonconvex-Strongly-Concave Minimax Optimization

In this section, we consider the stochastic nonconvex-strongly-concave minimax problem

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} f(x,y) \coloneqq \mathbb{E}[F(x,y;\xi)], \tag{10}$$

where $\xi \sim \mathcal{D}_f$ is the stochastic index and the stochastic gradient estimate $\nabla F(x,y;\xi)$ satisfies the p-BCM assumption. Concretely, we impose the following assumptions on problem (10).

Assumption 4.1 *We suppose minimax optimization problem (10) holds that*

(a) the objective function $f(x,y)$ is ℓ-smooth and μ-strongly concave in y;

(b) the primal function $\Phi(\cdot) = \max_{y\in\mathbb{R}^{d_y}} f(\cdot,y)$ is lower bounded, i.e., $\inf_{x\in\mathbb{R}^{d_x}} \Phi(x) > -\infty$;

(c) for given $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$, the algorithm can draw the stochastic indices $\xi \sim \mathcal{D}_f$ and access the stochastic first-order oracle $\nabla F(x,y;\xi)$ such that $\mathbb{E}[\nabla F(x,y;\xi)] = \nabla f(x,y)$ and $\mathbb{E}[\|\nabla F(x,y;\xi) - \nabla f(x,y)\|^p] \le \sigma^p$, where $p \in (1,2]$ and $\sigma > 0$.

It is worth noting that minimax problem (10) is a special case of bilevel problem (1) by taking $G = -F$. Therefore, we can directly solve minimax problem (10) by applying N^{2}SBA (Algorithm 1) with $\lambda = 0$ and omitting the update on the variable $\hat{z}_{t,k}$. We present this strategy in Algorithm 2, which is called Nested-loop Normalized Stochastic Gradient Descent Ascent (N^{2}SGDA).

It is natural to applying the general theoretical results in Section 3 to analyze our N^{2}SGDA for minimax optimization, while this cannot achieve the tight complexity bounds. Therefore, we need to specialize the structure of minimax problem(10) to refine the complexity bounds. We first consider the properties of function $\Phi(\cdot)$.

Lemma 8 (Lin et al. (2020, Lemma 4.3)) *Under Assumption 4.1, the function $\Phi(\cdot)$ is $(\ell+\kappa\ell)$-smooth and has the form of $\nabla\Phi(\cdot) = \nabla_x f(\cdot, y^*(\cdot))$, where $\kappa \coloneqq \ell/\mu$ is the condition number and $y^*(\cdot) = \arg\max_{y\in\mathbb{R}^{d_y}} f(\cdot,y)$. In addition, the mapping $y^*(\cdot)$ is κ-Lipschitz.*

We then establish the descent lemma for function $\Phi(\cdot)$ as follows.

Lemma 9 *Under Assumption 4.1, running N^{2}SGDA (Algorithm 2) holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t)\|] \le \frac{\mathbb{E}[\Phi(x_0) - \Phi(x_T)]}{\eta_x T} + \frac{4\sigma}{M^{\frac{p-1}{p}}} + \frac{2\ell\sum_{t=0}^{T-1}\mathbb{E}[\|y_t - y^*(x_t)\|]}{T} + \eta_x(\kappa+1)\ell.$$

Algorithm 2 Nested-loop Normalized Stochastic Gradient Descent Ascent (N^2SGDA)

1: **Input:** initial points $x_0 \in \mathbb{R}^{d_x}, \hat{y}_{0,0} \in \mathbb{R}^{d_y}$, stepsizes $\eta_x > 0, \{\eta_{y,t} > 0\}_{t=0}^{T-1}$, sequences of clipping parameters $\{\tau_{t,k}\}_{k=0}^{K-1}$, iteration numbers $T, K > 0$, batch size $M > 0$
2: **for** $t = 0, \ldots, T-1$
3: **for** $k = 0, \ldots, K-1$
4: draw $\xi_t' \sim \mathcal{D}_f$
5: $g_{y,t,k} = \nabla_y F(x_t, \hat{y}_{t,k}; \xi_t')$
6: $\hat{y}_{t,k+1} = \hat{y}_{t,k} + \eta_{y,t}\mathrm{clip}(g_{y,t,k}, \tau_{t,k})$
7: **end for**
8: $y_t = \hat{y}_{t,K}, \quad \hat{y}_{t+1,0} = \hat{y}_{t,K}$
9: draw $\xi_{t,i} \sim \mathcal{D}_f$ for all $i \in [M]$
10: $g_{x,t} = \dfrac{1}{M}\displaystyle\sum_{i=1}^{M} \nabla_x F(x_t, y_t; \xi_{t,i})$
11: $x_{t+1} = x_t - \eta_x \cdot \dfrac{g_{x,t}}{\|g_{x,t}\|}$
12: **end for**
13: draw $\hat{x}$ from $\{x_t\}_{t=1}^{T}$ uniformly
14: **Return:** $\hat{x}$

Similar to the analysis for the bilevel problem, the iterations (line 4-6) on $\hat{y}_{t,k}$ can be regarded as solving the strongly convex problem $\min_{y\in\mathbb{R}^d} -f(x_t, y)$. Hence, we combine Lemmas 2 and 9 to achieve the in-expectation guarantee for N^2SGDA as follows.

Theorem 10 *Under Assumption 4.1, we run Algorithm 2 (N^2SGDA) with*

$$\eta_x = \frac{\epsilon}{\kappa\ell}, \quad \eta_{y,t} = \tilde{\Theta}\left(\min\left\{\frac{1}{\ell}, \frac{1}{\mu K}\right\}\right), \quad \tau_{t,k} = \tilde{\Theta}\left(\frac{\exp(-\eta_{y,t}\mu k)R_y}{\eta_{y,t}}\right),$$
$$T = \mathcal{O}\left(\frac{\Delta_\Phi \kappa \ell}{\epsilon^2}\right), \quad M = \left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}, \quad K = \tilde{\mathcal{O}}\left(\kappa + \left(\frac{\ell^2\sigma^2}{\mu^2\epsilon^2}\right)^{\frac{p}{2(p-1)}} + \frac{R_y\ell}{\epsilon}\right),$$

where we set $\Delta_\Phi = \Phi(x_0) - \min_{x\in\mathbb{R}^{d_x}} \Phi(x)$, $R_y^2 = \mathcal{O}(R_0^2 + \epsilon^2/\ell^2)$, and $R_0^2 \geq \|\hat{y}_{0,0} - y^*(x_0)\|^2$. Then we have $\mathbb{E}[\|\nabla\Phi(\hat{x})\|] \leq \epsilon$. In particular, the overall SFO complexity is upper bounded by

$$\tilde{\mathcal{O}}\left(\frac{\Delta_\Phi \kappa^{\frac{2p-1}{p-1}} \ell \sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}} + \frac{\Delta_\Phi \kappa^2 \ell}{\epsilon^2} + \frac{R_0 \Delta_\Phi \kappa \ell^2}{\epsilon^3}\right).$$

Note that the upper bound of N^2SGDA (Theorem 10) is significantly better than that of N^2SBA (Theorem 3). Moreover, the dependency on σ and ϵ in the result of Theorem 10 matches the lower bound of $\Omega\big(\sigma^{\frac{p}{p-1}}\epsilon^{-\frac{3p-2}{p-1}}\big)$ for finding the ϵ-stationary point of the general nonconvex function under the p-BCM noise (Liu and Zhou, 2025). This also implies our

upper bound is near-optimal to σ and ϵ. Specifically, we consider the instance of minimax problem in which the objective function takes the form

$$f(x,y) = g(x) + h(y)$$

where $g(x)$ is the hard instance for nonconvex optimization under the p-BCM (Liu and Zhou, 2025) and $h(y)$ is a smooth and μ-strongly concave function. Consequently, identifying an ϵ-stationary point of function $\Phi(x)$ can be reduced to finding an $\mathcal{O}(\epsilon)$-stationary point of the nonconvex function $g(x)$, which requires at least the SFO complexity of $\Omega\big(\sigma^{\frac{p}{p-1}}\epsilon^{-\frac{3p-2}{p-1}}\big)$ (Liu and Zhou, 2025). However, the optimality with respect to ℓ and μ is still an open problem even in the special case of $p = 2$ (Li et al., 2021).

We now provide the high-probability guarantee for our N^{2}SGDA method.

Theorem 11 *Under Assumption 4.1, we run Algorithm 2 (N^{2}SGDA) with*

$$\eta_x = \sqrt{\frac{\Delta_\Phi}{(\kappa+1)\ell T}}, \quad \eta_{y,t} = \tilde{\Theta}\left(\min\left\{\frac{1}{\ell}, \frac{1}{\mu K}\right\}\right), \quad \tau_{t,k} = \tilde{\Theta}\left(\frac{\exp(-\eta_{y,t}\mu k)R_y}{\eta_{y,t}}\right),$$
$$T = \tilde{\mathcal{O}}\left(\frac{\Delta_\Phi \kappa \ell}{\epsilon^2}\right), \quad M = \left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}, \quad K = \tilde{\mathcal{O}}\left(\kappa + \left(\frac{\ell^2\sigma^2}{\mu^2\epsilon^2}\right)^{\frac{p}{2(p-1)}}\right),$$

where we set $\Delta_\Phi = \Phi(x_0) - \min_{x\in\mathbb{R}^{d_x}}\Phi(x)$, $R_y^2 = R_0^2 + 10\epsilon^2/\ell^2$, and $R_0^2 \geq \|\hat{y}_{0,0} - y^*(x_0)\|^2$. Then we have $\sum_{t=0}^{T-1}\|\nabla\Phi(x_t)\|_2/T \leq \epsilon$ with probability at least $1-\delta$. In particular, the overall SFO complexity is upper bounded by

$$\tilde{\mathcal{O}}\left(\frac{\Delta_\Phi \kappa^{\frac{2p-1}{p-1}}\ell\sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}} + \frac{\Delta_\Phi\kappa^2\ell}{\epsilon^2}\right).$$

We remark that the SFO complexity in Theorem 11 matches its in-expectation counterpart in Theorem 10, up to a multiplicative logarithmic factor. In addition, its dependency on δ and ϵ also nearly matches the lower bound.

5. Experiments

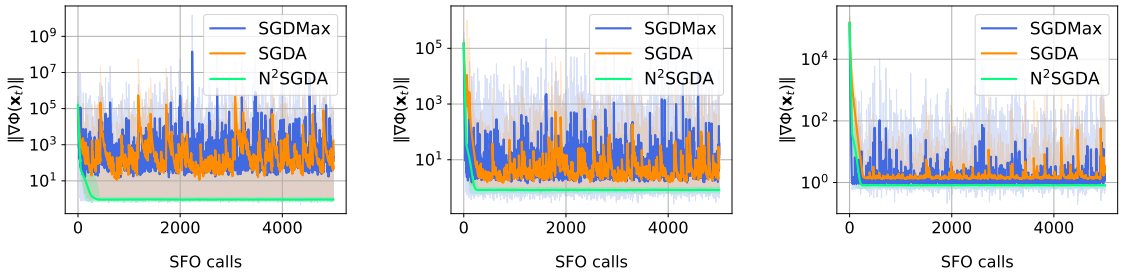
In this section, we conduct numerical experiments to validate the superiority of proposed algorithms for solving stochastic minimax and bilevel optimization problems.

5.1 A Synthetic Problem

We consider the synthetic nonconvex-strongly-concave minimax optimization problem

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}} f(x,y) = m_1\left[\|x\|_2^2 + \sin(3\sqrt{\|x\|^2+1})\right] + x^\top K y - m_2\|y\|^2, \tag{11}$$

where $K \in \mathbb{R}^{d\times d}$ and $m_1, m_2 > 0$. Following the setup of Laguel et al. (2023, 2024), we set $d_x = 30$ and $m_1 = m_2 = 1$. For the interaction matrix, we set $K = 10\tilde{K}/\|\tilde{K}\|$, where $\tilde{K} = (M + M^\top)/2$ and $M \in \mathbb{R}^{d\times d}$ has entries independently drawn from $\mathcal{N}(0,1)$.



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

Figure 1: We compare results of the number of SFO calls against $\|\nabla\Phi(x)\|$ on a synthetic nonconvex–strongly-concave minimax problem (11) with Pareto noise for the shape parameter $\alpha \in \{1.2, 1.5, 1.8\}$. The shaded region represents the standard deviation of the method.

In experiments, we employ the stochastic gradient estimate as $\nabla F(x, y; \xi) = \nabla f(x, y) + \xi$ for given $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}$, where $\xi = \hat{\xi} - \mathbb{E}[\hat{\xi}]$ and $\hat{\xi}$ is drawn from the Pareto distribution with the shape parameter $\alpha \in (1, 2]$. This construction ensures that the estimate $F(x, y; \xi)$ satisfies our Assumption 4.1.

We compare our N^{2}SGDA (Algorithm 2) against baseline methods including stochastic gradient descent ascent (SGDA) (Lin et al., 2020) and stochastic gradient descent with a max-oracle (SGDmax) (Jin et al., 2020; Nouiehed et al., 2019). The stepsizes for all algorithms are tuned from $\{1 \times 10^{-6}, 3 \times 10^{-6}, 1 \times 10^{-5} \ldots, 1\}$. For N^{2}SGDA and SGDmax, we fix the number of inner iterations be 10. We set the total iteration count to $T = 5{,}000$ for N^{2}SGDA, Clipped-SGDmax, and SGDmax, and $T = 50{,}000$ for SGDA, so that all methods incur the same total number of stochastic first-order oracle calls. We perform 20 independent runs for each algorithm and present the averaged results.

We present the performance of all methods under Pareto noise with $\alpha \in \{1.2, 1.5, 1.8\}$ in Figure 1. We observe the proposed N^{2}SGDA converges faster than baseline methods SGDmax and SGDA, and its advantage is especially evident in the cases of $\alpha = 1.2$ and $\alpha = 1.5$. Furthermore, the baselines SGDmax and SGDA exhibit pronounced oscillations in their trajectories, while the proposed N^{2}SGDA demonstrates markedly more stable convergence behavior under the heavy-tailed noise.

5.2 Learnable Regularization

We consider the problem of learnable regularization in logistic regression (Grazzi et al., 2020; Ji et al., 2021; Chen et al., 2025), which is formulated by the bilevel optimization problem,

$$\begin{aligned} &\min_{x \in \mathbb{R}^p} \varphi(x) \coloneqq \left\{ \frac{1}{|\mathcal{D}^{\mathrm{val}}|} \sum_{(a_i, b_i) \in \mathcal{D}^{\mathrm{val}}} \ell\left(\langle a_i, y^*(x) \rangle, b_i \right) \right\}, \\ &\text{s.t} \quad y^*(x) = \arg\min_{y \in \mathbb{R}^{p \times c}} \left\{ \frac{1}{|\mathcal{D}^{\mathrm{tr}}|} \sum_{(a_i, b_i) \in \mathcal{D}^{\mathrm{tr}}} \ell\left(\langle a_i, y \rangle, b_i \right) + \mathrm{tr}(y^\top \Sigma(x)\, y) \right\}, \end{aligned} \tag{12}$$

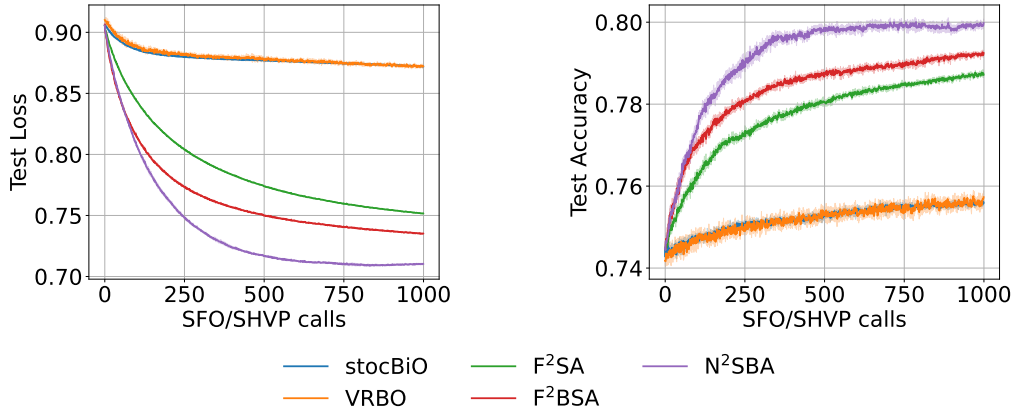


Figure 2: We present the results of stochastic oracle calls (SFO/SHVP) against test loss and test accuracy on bilevel problem (12) for the problem of learnable regularization in logistic regression. The shaded region represents the standard deviation.

where $\mathcal{D}^{\text{tr}}$ is the training dataset and $\mathcal{D}^{\text{val}}$ is the validation dataset, $a_i \in \mathbb{R}^p$ is the data feature, $b_i \in \{0, 1, \ldots, c-1\}$ is the corresponding class label, c is the number of classes, $\Sigma(x) := \text{diag}(\exp(x_1), \ldots, \exp(x_p))$ specifies the regularizers on each feature, and $\ell(\cdot, \cdot)$ denotes the cross entropy function, i.e.,

$$\ell(Z, y) = -\sum_{i=1}^{c} \mathbb{I}(i = y) \log \left(\frac{\exp(Z_i)}{\sum_{j=1}^{c} \exp(Z_j)} \right),$$

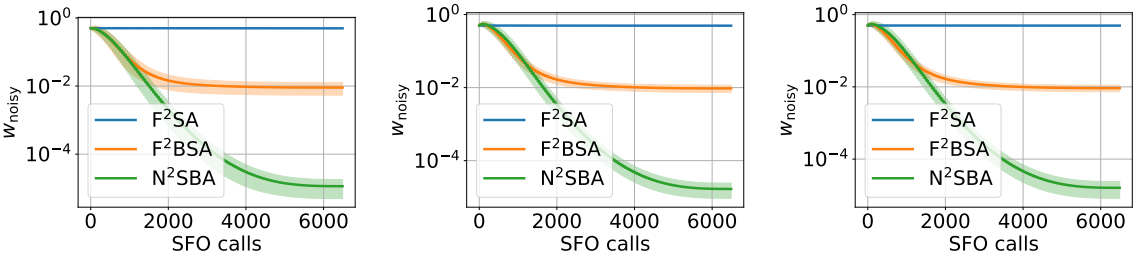
We conduct our experiments on dataset "20 newsgroup" that consists of 18,000 documents with 20 classes. Each document is represented as a p-dimensional vector (p = 101,631), where each of its coordinates corresponds to the term frequency–inverse document frequency of the word.

We compare our N^{2}SBA method with the stochastic Hessian-vector product-based methods stocBiO (Ji et al., 2021) and VRBO (Yang et al., 2021), as well as the stochastic first-order methods F^{2}SA (Kwon et al., 2023) and F^{2}BSA (Chen et al., 2025). For all algorithms, we set the number of inner iterations be 10 and the number of outer iterations be 1,000. In addition, we tune the Lagrangian parameter λ from $\{10^1, 10^2, 10^3, 10^4\}$ for F^{2}SA, F^{2}BSA, and N^{2}SBA. The stepsizes of all algorithms are tuned from $\{10^{-5}, 10^{-4}., \ldots, 10^3\}$. We run the experiment with 10 different random seeds and present the empirical results in Figure 2. Clearly, our proposed N^{2}SBA method outperforms all baseline methods.

5.3 Data Hyper-Cleaning

We consider the application of data hyper-cleaning on multiple data sources, which is formulated by the bilevel optimization problem

$$\begin{aligned} &\min_{x \in \mathbb{R}^m} \varphi(x) := \ell_{\text{val}}(y^*(x)), \\ &\text{s.t} \quad y^*(x) = \arg\min_{y \in \mathbb{R}^p} \sum_{i=1}^{m} \sigma(x_i) \ell_{\text{tr}}^i(y), \end{aligned} \tag{13}$$



(a) $p = 0.5$ (b) $p = 0.9$ (c) $p = 0.99$

Figure 3: We present the results of SFO calls against the weights on garbage for the data hyper-cleaning problem (13) under different corruption ratios $q \in \{0.5, 0.9, 0.99\}$. The shaded region represents the standard deviation of the method.

where $\ell_{\text{val}}(\cdot)$ is the validation loss, $\ell_{\text{tr}}^i(\cdot)$ is the training loss on the ith data source, m is the number of data sources, and $\sigma(x_i) = \exp(x_i)/\sum_{j=1}^m \exp(x_j)$ is the softmax function. We perform our experiment on training the GPT-2 model with 124M parameters (Radford et al., 2019) on dataset "Alpaca" dataset (Taori et al., 2023), which contains 52,000 instructions paired with demonstration outputs generated by OpenAI's "text-davinci-003" engine from $m = 10,000$ data sources. Following the setup of Chen et al. (2025), we split the dataset into training and validation sets in an 8:2 ratio, then introduce corruption into the training set at a proportion of $q \in (0, 1)$, i.e., implement the corruption by replacing the demonstration outputs with empty strings.

We run our experiment on a system of 8 GPUs. We compare our proposed method N^{2}SBA with F^{2}SA (Kwon et al., 2023) and F^{2}BSA (Chen et al., 2025). Note that we do not include the SHVP-based methods since they are difficult to implement in multi-GPU training (Tarzanagh et al., 2022; Chen et al., 2023, 2025). For all algorithms, we use a batch size of 64 and set a fixed penalty parameter $\lambda = 10^3$. We run our experiment with 10 different random seeds and present the empirical results for different corruption ratios $q = 0.5, 0.9, 0.99$ in Figure 5.3. We observe that our proposed N^{2}SBA methods converge significantly faster than F^{2}SA and F^{2}BSA, which aligns with the empirical findings in the single-level nonconvex problem Zhang et al. (2020) that the performance of SGD-based methods deteriorates in the presence of heavy-tailed noise, whereas the strategies of gradient clipping and normalization can mitigate this effect and thereby improve the convergence.

6. Conclusion and Future Work

In this work, we propose a stochastic first-order method called N^{2}SBA for solving stochastic nonconvex–strongly-convex bilevel optimization problems in the presence of heavy-tailed noise. We establish both in-expectation and high-probability theoretical guarantees for N^{2}SBA under the p-BCM gradient noise assumption. We prove that the stochastic first-order (SFO) complexity of our method matches that of the best-known stochastic first-order bilevel optimization algorithms under the specific bounded-variance assumption.

Furthermore, we introduce N^2SGDA for addressing stochastic nonconvex–strongly-concave minimax optimization problems with heavy-tailed noise, achieving the SFO complexity with the nearly optimal dependence on noise level and accuracy.

In future work, it is interesting to incorporate acceleration techniques (Sadiev et al., 2023; He et al., 2025; Fradin et al., 2026) to further improve the dependence on condition numbers in upper complexity bounds. Another promising direction is to simplify our algorithms for bilevel and minimax optimization problems, e.g., establish the single-loop methods. In addition, it would be valuable to extend our methodology to constrained bilevel and minimax optimization problems.

Acknowledgment

We greatly appreciate the insightful discussion with Lesi Chen. This work is supported by the National Natural Science Foundation of China (No. 12571557) and the Shanghai Basic Research Program (23JC1401000). The computations in this research were performed using the CFFF platform of Fudan University

Appendix A. Supporting Lemmas

In this section, we revisit some technical lemmas that are essential for establishing the convergence guarantee of the proposed algorithms.

Lemma 12 (Hübler et al. (2024, Lemma 7)) *For all $a, b \in \mathbb{R}^d$ such that $b \neq 0$, it holds that $a^\top b / \|b\| \geq \|a\| - 2\|a - b\|$.*

Lemma 13 (Hübler et al. (2024, Lemma 10)) *Let $p \in (1, 2]$ and $X_1, \ldots, X_n \in \mathbb{R}^d$ be a martingale difference sequence, i.e., $\mathbb{E}[X_j \mid X_{j-1}, \ldots, X_1] = 0$ a.s. and $\mathbb{E}[\|X_j\|^p] < +\infty$ for all $j = 1, \ldots, n$. Define $S_n \coloneqq \sum_{j=1}^n X_j$, then we have $\mathbb{E}[\|S_n\|^p] \leq 2\sum_{j=1}^n \mathbb{E}[\|X_j\|^p]$.*

Lemma 14 (Durrett (2019, Example 4.1.7)) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X, Y be independent random variables mapping to measurable spaces (E_1, Σ_1) and (E_2, Σ_2) respectively. Furthermore, let $h : E_1 \times E_2 \to \mathbb{R}^d$ be a (Lebesgue-)measurable function with $\mathbb{E}[\|h(X, Y)\|] < \infty$. Then we have $\mathbb{E}[h(X, Y) \mid X] \overset{a.s.}{=} g(X)$, where $g(\cdot) \coloneqq \mathbb{E}[h(\cdot, Y)]$.*

The following lemma provides an upper bound on the norm of the difference between a minibatch gradient estimator and its expectation.

Lemma 15 *Assume that $\nabla f(x, y) = \mathbb{E}[\nabla F(x, y; \xi)]$ and $\mathbb{E}[\|\nabla F(x, y; \xi) - \nabla f(x, y)\|^p] \leq \sigma^p$. Let $g = \frac{1}{M}\sum_{i=1}^M \nabla F(x, y; \xi_i)$, then it holds that $\mathbb{E}[\|g - \nabla f(x, y)\|] \leq 2\sigma / M^{\frac{p-1}{p}}$.*

Proof Recall that the minibatch gradient estimator is defined as

$$g = \frac{1}{M}\sum_{i=1}^M \nabla F(x, y; \xi_i).$$

We start by controlling the expected deviation of g from $\nabla f(x, y)$ using Lemma 13. We define $X_j(x, y) \coloneqq \nabla F(x, y; \xi_j) - \nabla f(x, y)$ for all $j \in [M]$. Now note that $X_1(x, y), \ldots, X_M(x, y)$ are

independent random variables with mean zero and hence a Martingale Difference Sequence (MDS). Furthermore, note that $\mathbb{E}[\|X_j(x,y)\|^p] \leq \sigma^p$ by the assumption. Therefore, we can apply the Lemma 13 to get

$$l(x,y) \coloneqq \mathbb{E}\left[\left\|\sum_{j=1}^{M} X_j(x,y)\right\|^p\right] \leq 2\sum_{j=1}^{M}\mathbb{E}[\|X_j(x,y)\|^p] \leq 2M\sigma^p.$$

Next we calculate

$$\begin{aligned}\mathbb{E}[\|g-\nabla f(x,y)\| \mid x,y] =& \mathbb{E}\left[\left\|\frac{1}{M}\sum_{j=1}^{M}\left(\nabla F(x,y;\xi_j)-\nabla f(x,y)\right)\right\| \;\middle|\; x,y\right] \\ \leq& \frac{1}{M}\mathbb{E}\left[\left\|\sum_{j=1}^{M}\left(\nabla F(x,y;\xi_j)-\nabla f(x,y)\right)\right\|^p \;\middle|\; x,y\right]^{1/p},\end{aligned}$$

where we applied Jensen's inequality in the last inequality. Next we define

$$Y=(\xi_1,\ldots,\xi_M) \quad \text{and} \quad h(x,y,Y)=\left\|\sum_{j=1}^{M}\left(\nabla F(x,y,\xi_j)-\nabla f(x,y)\right)\right\|^p,$$

and note that (x,y) and Y are independent. Hence, we may apply Lemma 14 which yields

$$\mathbb{E}[\|g-\nabla f(x,y)\|] \leq \frac{1}{M}\mathbb{E}[h(x,y,Y) \mid x,y]^{1/p} = \frac{1}{M}l(x,y)^{1/p} \leq \frac{2\sigma}{M^{\frac{p-1}{p}}}.$$

■

The clip operator enjoys the following properties.

Lemma 16 (Sadiev et al. (2023, Lemma 5.1)) *Let X be a random vector in $\mathbb{R}^d$ and define $\tilde{X} = \operatorname{clip}(X,\tau) \coloneqq \min(1,\tau/\|X\|)X$, then we have $\|\tilde{X}-\mathbb{E}[\tilde{X}]\| \leq 2\tau$. Moreover, if for some $\sigma \geq 0$ and $p \in (1,2]$ we have $\mathbb{E}[X]=x \in \mathbb{R}^d$, $\mathbb{E}[\|X-x\|^p] \leq \sigma^p$, and $\|x\| \leq \tau/2$, then we have*

$$\left\|\mathbb{E}[\tilde{X}]-x\right\| \leq \frac{2^p\sigma^p}{\tau^{p-1}}, \quad \mathbb{E}\left[\left\|\tilde{X}-x\right\|^2\right] \leq 18\tau^{2-p}\sigma^p \quad \text{and} \quad \mathbb{E}\left[\left\|\tilde{X}-\mathbb{E}[\tilde{X}]\right\|^2\right] \leq 18\tau^{2-p}\sigma^p.$$

Finally, we present some technical lemmas for high-probability convergence analysis.

Lemma 17 (Li and Orabona (2020, Lemma 1)) *Let $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ be a filtration and $\{\Theta_t\}_{t\in\mathbb{N}}$ be a Martingale Difference Sequence with respect to $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$. Furthermore, for each $t \in \mathbb{N}_+$, let σ_t be $\mathcal{F}_{t-1}$-measurable and assume that $\mathbb{E}[\exp(\Theta_t^2/\sigma_t^2) \mid \mathcal{F}_{t-1}] \leq \mathrm{e}$. Then for all $T \in \mathbb{N}$, $\chi > 0$ and $\hat{\delta} \in (0,1)$, it holds that*

$$\mathbb{P}\left(\sum_{t=1}^{T}\Theta_t \leq \frac{3\chi}{4}\sum_{t=1}^{T}\sigma_t^2 + \frac{1}{\chi}\ln\left(\frac{1}{\hat{\delta}}\right)\right) \geq 1-\hat{\delta}.$$

Our proof also uses the following Bernstein inequality for martingale differences (Bennett, 1962; Dzhaparidze and Van Zanten, 2001; Freedman, 1975).

Lemma 18 (Freedman (1975, Lemma 1.6)) *Let the sequence of random variables $\{X_i\}_{i\geq 1}$ form a martingale difference sequence, i.e., $\mathbb{E}[X_i \mid X_{i-1}, \ldots, X_1] = 0$ for all $i \geq 1$. Assume that conditional variances $\sigma_i^2 \coloneqq \mathbb{E}[X_i^2 \mid X_{i-1}, \ldots, X_1]$ exist and are bounded and assume also that there exists a deterministic constant $c > 0$ such that $|X_i| \leq c$ almost surely for all $i \geq 1$. Then for all $b > 0$, $G > 0$ and $n > 1$, it holds that*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i\right| > b \quad \text{and} \quad \sum_{i=1}^n \sigma_i^2 \leq G\right\} \leq 2\exp\left(-\frac{b^2}{2G + 2cb/3}\right). \tag{14}$$

Appendix B. Proofs for Results of Stochastic Bilevel Optimization

In this section, we establish the convergence guarantee of the N^{2}SBA algorithm for solving the nonconvex–strongly-convex bilevel optimization problem. We begin by presenting key properties of the bilevel objective function $\varphi(x)$ and the auxiliary function $\mathcal{L}_\lambda^*(x)$.

Lemma 19 (Kwon et al. (2023, page 4)) *Under Assumption 2.1, for all $\lambda \geq 2L_f/\mu$, the function $\mathcal{L}_\lambda(x, y) \coloneqq f(x, y) + \lambda(g(x, y) - g^*(x))$ is $(\lambda\mu/2)$-strongly convex in y for given $x \in \mathbb{R}^{d_x}$.*

Lemma 20 (Chen et al. (2025, Lemma B.2–B.5)) *Under Assumption 2.1, for $\lambda \geq 2L_f/\mu$, it holds that*

$$\begin{aligned} &\|y_\lambda^*(x) - y^*(x)\| \leq C_f/(\lambda\mu), \quad &&|\mathcal{L}_\lambda^*(x) - \varphi(x)| \leq D_0/\lambda, \\ &\|\nabla\mathcal{L}_\lambda^*(x) - \nabla\varphi(x)\| \leq D_1/\lambda, \quad &&\|\nabla y^*(x) - \nabla y_\lambda^*(x)\| \leq D_2/\lambda, \end{aligned}$$

where we define

$$D_0 \coloneqq \left(C_f + \frac{C_f L_g}{2\mu}\right)\frac{C_f}{\mu} = \mathcal{O}(\ell\kappa^2), \quad D_1 \coloneqq \left(L_f + \frac{\rho_g L_g}{\mu} + \frac{C_f L_g \rho_g}{2\mu^2} + \frac{C_f\rho_g}{2\mu}\right)\frac{C_f}{\mu} = \mathcal{O}(\ell\kappa^3),$$
$$D_2 \coloneqq \left(\frac{1}{\mu} + \frac{2L_g}{\mu^2}\right)\left(L_f + \frac{C_f\rho_g}{\mu}\right) = \mathcal{O}(\kappa^3).$$

Lemma 21 (Chen et al. (2025, Lemma B.6)) *Under Assumption 2.1, for $\lambda \geq 2L_f/\mu$, it holds that $\|\nabla y^*(x)\| \leq L_g/\mu$ and $\|\nabla y_\lambda^*(x)\| \leq 4L_g/\mu$.*

Lemma 22 (Chen et al. (2025, Lemma B.7)) *Under Assumption 2.1, for $\lambda \geq 2L_f/\mu$, it holds that $\nabla\mathcal{L}_\lambda^*(x)$ is D_3-Lipschitz, where*

$$D_3 \coloneqq L_f + \frac{4L_f L_g}{\mu} + \frac{C_f\rho_g}{\mu} + \frac{C_f L_g \rho_g}{\mu^2} + L_g D_2 = \mathcal{O}(\ell\kappa^3).$$

B.1 The In-Expectation Convergence Guarantee

In this subsection, we present the proofs for the in-expectation convergence guarantee of our N^{2}SBA shown in Section 3.2.

B.1.1 The Proof of Lemma 1

Proof Lemma 22 implies that $\mathcal{L}_\lambda$ is D_3-smooth, then we have

$$\begin{aligned}\mathcal{L}_\lambda^*(x_{t+1}) \leq & \mathcal{L}_\lambda^*(x_t) + \langle \nabla \mathcal{L}_\lambda^*(x_t), x_{t+1} - x_t \rangle + \frac{D_3}{2} \|x_{t+1} - x_t\|^2 \\ = & \mathcal{L}_\lambda^*(x_t) - \eta_x \left\langle \nabla \mathcal{L}_\lambda^*(x_t), \frac{g_{x,t}}{\|g_{x,t}\|} \right\rangle + \frac{D_3 \eta_x^2}{2} \\ \leq & \mathcal{L}_\lambda^*(x_t) - \eta_x \|\nabla \mathcal{L}_\lambda^*(x_t)\| + 2\eta_x \|\nabla \mathcal{L}_\lambda^*(x_t) - g_{x,t}\| + \frac{D_3 \eta_x^2}{2},\end{aligned} \tag{15}$$

where the last inequality is due to Lemma 12. Taking the average on equation (15) over $t = 0, 1, \ldots, T-1$, we obtain

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \mathcal{L}_\lambda^*(x_t)\| \leq \frac{\mathcal{L}_\lambda^*(x_0) - \mathcal{L}_\lambda^*(x_T)}{\eta_x T} + \sum_{t=0}^{T-1} \frac{2\|\nabla \mathcal{L}_\lambda^*(x_t) - g_{x,t}\|}{T} + \frac{D_3 \eta_x}{2}. \tag{16}$$

Based on the facts (Kwon et al., 2023, Lemma 3.1)

$$\nabla \mathcal{L}_\lambda^*(x) = \nabla_x f(x, y_\lambda^*(x)) + \lambda(\nabla_x g(x, y_\lambda^*(x)) - \nabla_x g(x, y^*(x)))$$

and

$$g_{x,t} = \frac{1}{M} \sum_{i=1}^{M} \left(\nabla_x F(x_t, y_t; \xi_{t,i}) + \lambda(\nabla_x G(x_t, y_t; \zeta_{t,i}) - \nabla_x G(x_t, z_t; \zeta_{t,i})) \right),$$

we apply the triangle inequality to obtain

$$\begin{aligned}& \|g_{x,t} - \nabla \mathcal{L}_\lambda^*(x_t)\| \\ \leq & \left\| \frac{1}{M} \sum_{i=1}^{M} \nabla_x F(x_t, y_t; \xi_{t,i}) - \nabla_x f(x_t, y_t) \right\| + \lambda \left\| \frac{1}{M} \sum_{i=1}^{M} \nabla_x G(x_t, y_t; \zeta_{t,i}) - \nabla_x g(x_t, y_t) \right\| \\ & + \lambda \left\| \frac{1}{M} \sum_{i=1}^{M} \nabla_x G(x_t, z_t; \zeta_{t,i}) - \nabla_x g(x_t, z_t) \right\| + \|\nabla_x f(x_t, y_t) - \nabla_x f(x_t, y_\lambda^*(x_t))\| \\ & + \lambda \|\nabla_x g(x_t, y_t) - \nabla_x g(x_t, y_\lambda^*(x_t))\| + \lambda \|\nabla_x g(x_t, z_t) - \nabla_x g(x_t, y^*(x_t))\| \\ \leq & \left\| \frac{1}{M} \sum_{i=1}^{M} \nabla F(x_t, y_t; \xi_{t,i}) - \nabla f(x_t, y_t) \right\| + \lambda \left\| \frac{1}{M} \sum_{i=1}^{M} \nabla G(x_t, y_t; \zeta_{t,i}) - \nabla g(x_t, y_t) \right\| \\ & + \lambda \left\| \frac{1}{M} \sum_{i=1}^{M} \nabla G(x_t, z_t; \zeta_{t,i}) - \nabla g(x_t, z_t) \right\| + \|\nabla f(x_t, y_t) - \nabla f(x_t, y_\lambda^*(x_t))\| \\ & + \lambda \|\nabla g(x_t, y_t) - \nabla g(x_t, y_\lambda^*(x_t))\| + \lambda \|\nabla g(x_t, z_t) - \nabla g(x_t, y^*(x_t))\|.\end{aligned}$$

Taking the expectation on both sides of the above inequality, we have

$$\begin{aligned}
&\mathbb{E}[\|g_{x,t} - \nabla\mathcal{L}_\lambda^*(x_t)\|] \\
\leq&\frac{2\sigma_f}{M^{\frac{p-1}{p}}} + \frac{4\lambda\sigma_g}{M^{\frac{p-1}{p}}} + (L_f + \lambda L_g)\mathbb{E}[\|y_t - y_\lambda^*(x_t)\|] + \lambda L_g\mathbb{E}[\|z_t - y^*(x_t)\|] \\
\leq&\frac{2\sigma_f}{M^{\frac{p-1}{p}}} + \frac{4\lambda\sigma_g}{M^{\frac{p-1}{p}}} + \left(\frac{\lambda\mu}{2} + \lambda L_g\right)\mathbb{E}[\|y_t - y_\lambda^*(x_t)\|] + \lambda L_g\mathbb{E}[\|z_t - y^*(x_t)\|] \\
\leq&\frac{2\sigma_f + 4\lambda\sigma_g}{M^{\frac{p-1}{p}}} + 2\lambda L_g\mathbb{E}[\|y_t - y_\lambda^*(x_t)\|] + \lambda L_g\mathbb{E}[\|z_t - y^*(x_t)\|],
\end{aligned} \tag{17}$$

where the first inequality follows from Lemma 15, the second inequality is due to the setting of $\lambda \geq 2L_f/\mu$, and the last inequality follows the fact $L_g \geq \mu$. Combing the results of (16) and (17), we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}_\lambda^*(x_t)\|] \leq&\frac{\mathbb{E}[\mathcal{L}_\lambda^*(x_0) - \mathcal{L}_\lambda^*(x_T)]}{\eta_x T} + \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}_\lambda^*(x_t) - g_{x,t}\|] + \frac{D_3\eta_x}{2} \\
\leq&\frac{\mathbb{E}[\mathcal{L}_\lambda^*(x_0) - \mathcal{L}_\lambda^*(x_T)]}{\eta_x T} + \frac{4\sigma_f + 8\lambda\sigma_g}{M^{\frac{p-1}{p}}} + \frac{4\lambda L_g}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|y_t - y_\lambda^*(x_t)\|] \\
&+ \frac{2\lambda L_g}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|z_t - y^*(x_t)\|] + \frac{D_3\eta_x}{2},
\end{aligned} \tag{18}$$

which finishes the proof. ■

B.1.2 The Proof of Theorem 3

We first provide an upper bound on the p-th bounded central moment of the stochastic gradient estimator of $\mathcal{L}_\lambda(x, y)$ as follows.

Lemma 23 *Under Assumption 2.2, it holds that*

$$\mathbb{E}[\|\nabla_y F(x, y; \xi) + \lambda\nabla_y G(x, y; \zeta) - \nabla_y f(x, y) - \lambda\nabla_y g(x, y)\|^p] \leq (2\sigma_f + 2\lambda\sigma_g)^p.$$

Proof Based on Assumption 2.2, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla_y F(x, y; \xi) + \lambda\nabla_y G(x, y; \zeta) - \nabla_y f(x, y) - \lambda\nabla_y g(x, y)\|^p] \\
\leq&\mathbb{E}[(\|\nabla_y F(x, y; \xi) - \nabla_y f(x, y)\| + \lambda\|\nabla_y G(x, y; \zeta) - \nabla_y g(x, y)\|)^p] \\
\leq&2^{p-1}\mathbb{E}[\|\nabla F(x, y; \xi) - \nabla f(x, y)\|^p] + 2^{p-1}\lambda^p\mathbb{E}[\|\nabla G(x, y; \zeta) - \nabla g(x, y)\|^p] \\
\leq&2^{p-1}\sigma_f^p + 2^{p-1}\lambda^p\sigma_g^p \\
\leq&2^{p-1}(\sigma_f + \lambda\sigma_g)^p \\
\leq&(2\sigma_f + 2\lambda\sigma_g)^p,
\end{aligned} \tag{19}$$

where the first inequality is due to the triangle inequality, and the second inequality follows from the convexity of $\|\cdot\|^p$. The third inequality is due to Assumption 2.2, and the fourth inequality follows from $a^p + b^p \leq (a+b)^p$ for any positive number a and b. ■

We then provide lemmas to upper bound the terms $\mathbb{E}[\|y_t - y_\lambda^*(x_t)\|]$ and $\mathbb{E}[\|z_t - y^*(x_t)\|]$.

Lemma 24 *Following the setting of Theorem 3, for all $t = 0, \ldots, T-1$, we have*

$$\mathbb{E}\left[\|y_t - y_\lambda^*(x_t)\|\right] = \mathcal{O}\left(\frac{\epsilon^2}{\ell^2\kappa^3}\right) \quad \textit{and} \quad \mathbb{E}\left[\|z_t - y^*(x_t)\|\right] = \mathcal{O}\left(\frac{\epsilon^2}{\ell^2\kappa^3}\right)$$

Proof We specify the parameter settings for N^{2}SBA as

$$\lambda = \max\left\{\frac{\kappa}{R_0}, \frac{\ell\kappa^2}{\Delta}, \frac{\ell\kappa^3}{\epsilon}\right\}, \quad \eta_{y,t} = \min\left\{\frac{1}{400(L_f + \lambda L_g)}, \frac{2\ln B_{y,t}}{\lambda\mu K}\right\}$$
$$\eta_{z,t} = \min\left\{\frac{1}{400\lambda L_g}, \frac{\ln(B_{z,t})}{\lambda\mu K}\right\}, \quad K = \tilde{\mathcal{O}}\left(\frac{\ell^{\frac{p}{p-1}}\kappa^{\frac{4p}{p-1}}\sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{2p}{p-1}}} + \frac{\ell^2\kappa^3 R_y}{\epsilon^2}\right), \quad \eta_x = \frac{\epsilon}{\ell\kappa^3},$$
$$\tau_{t,k} = \frac{\exp(-\eta_{z,t}\lambda\mu(1/2 + k/4))R_z}{120\eta_{z,t}}, \quad \tau'_{t,k} = \frac{\exp(-\eta_{y,t}\lambda\mu(1 + k/2))R_y}{120\eta_{y,t}},$$

where we set

$$B_{y,t} = \max\left\{2, \frac{\lambda^2\mu^2 K^{\frac{2(p-1)}{p}} R_y^2}{5400^{\frac{2}{p}}(\sigma_f + \lambda\sigma_g)^2(\ln(B_{y,t}))^2}\right\}, \quad B_{z,t} = \max\left\{2, \frac{\mu^2 K^{\frac{2(p-1)}{p}} R_z^2}{5400^{\frac{2}{p}}\sigma_g^2(\ln(B_{z,t}))^2}\right\},$$
$$R_y^2 = R_z^2 \coloneqq \mathcal{O}\left(R_0^2 + \frac{\epsilon^4}{\ell^4\kappa^6} + \frac{\epsilon^2}{\ell^2\kappa^4}\right), \ R_0 \geq \|\hat{y}_{0,0} - y^*(x_0)\|, \ \Delta = \varphi(x_0) - \inf_{x\in\mathbb{R}^{d_x}}\varphi(x).$$

We first consider the term $\mathbb{E}[\|y_t - y_\lambda^*(x_t)\|]$. Recall that Algorithm 1 set $y_t = \hat{y}_{t,K}$ in line 10, where $\hat{y}_{t,K}$ can be regarded as the output of applying clipped stochastic gradient descent (lines 5–6) to minimize the function $\mathcal{L}_\lambda(x_t, y) = f(x_t, y) + \lambda(g(x_t, y) - g^*(x_t))$ with respect to y. The updates use the stochastic gradient estimate $\nabla_y F(x_t, y; \xi'_t) + \lambda\nabla_y G(x_t, y; \zeta'_t)$ for $\mathcal{L}_\lambda(x_t, y)$. Since the setting of λ and Lemma 19 implies that the function $\mathcal{L}_\lambda(x_t, y)$ is $(\lambda\mu/2)$-strongly convex in y, we apply Lemma 2 with $h(y) = \mathcal{L}_\lambda(x_t, y)$, $\sigma_h = 2(\sigma_f + \lambda\sigma_g)$, $\ell_h = 2\lambda L_g$, and $\mu_h = \lambda\mu/2$ to obtain the following result

$$\begin{aligned}&\mathbb{E}\left[\|\hat{y}_{t,K} - y_\lambda^*(x_t)\|^2\right] \\ =&\mathcal{O}\left(\hat{R}_{y,t}^2 \exp\left(-\frac{K\mu}{L_g\ln(KL_g/\mu)}\right) + \frac{(\sigma_f + \lambda\sigma_g)^2(\ln(B_K))^2(\ln(KL_g/\mu))^{\frac{2(p-1)}{p}}}{\lambda^2\mu^2 K^{\frac{2(p-1)}{p}}} + \frac{\hat{R}_{y,t}^2}{K^2}\right),\end{aligned} \tag{20}$$

for all $\hat{R}_{y,t}^2 \geq \mathbb{E}[\|\hat{y}_{t,0} - y_\lambda^*(x_t)\|^2]$. Here, the choice of $\sigma_h = 2(\sigma_f + \lambda\sigma_g)$ that corresponds to the upper bound for the p-th central moment shown in Lemma 23.

We then use induction to show each $\mathbb{E}[\|\hat{y}_{t,0} - y_\lambda^*(x_t)\|^2]$ has a uniform upper bound, i.e.,

$$\mathbb{E}\left[\|\hat{y}_{t,0} - y_\lambda^*(x_t)\|^2\right] = R_y^2 \coloneqq \mathcal{O}\left(R_0^2 + \frac{\epsilon^4}{\ell^4\kappa^6} + \frac{\epsilon^2}{\ell^2\kappa^4}\right) \tag{21}$$

holds for all $t = 0, \ldots, T-1$. For the induction base $t = 0$, it follows that

$$\mathbb{E}[\|\hat{y}_{0,0} - y_\lambda^*(x_0)\|^2] \leq 2\mathbb{E}[\|\hat{y}_{0,0} - y^*(x_0)\|^2] + 2\|y^*(x_0) - y_\lambda^*(x_0)\|^2 \leq 2R_0^2 + 2R_0^2 = 4R_0^2, \quad (22)$$

where the first step follows the Young's inequality and the second step is due to the result $\|y_\lambda^*(x) - y^*(x)\| \leq C_f/(\lambda\mu)$ from Lemma 20 and the settings of $R_0 \geq \|\hat{y}_{0,0} - y^*(x_0)\|$ and $\lambda = \max\{\kappa R_0^{-1}, \ell\kappa^2\Delta^{-1}, \ell\kappa^3\epsilon^{-1}\}$.

For the induction step, we suppose equation (21) holds for all $t = 0, \ldots, T'-1$. In the case of $t = T'$, we have

$$\begin{aligned}
\left\|\hat{y}_{T',0} - y_\lambda^*(x_{T'})\right\|^2 \leq & 2\left\|\hat{y}_{T'-1,K} - y_\lambda^*(x_{T'-1})\right\|^2 + 2\|y_\lambda^*(x_{T'}) - y_\lambda^*(x_{T'-1})\|^2 \\
\leq & 2\left\|\hat{y}_{T'-1,K} - y_\lambda^*(x_{T'-1})\right\|^2 + \frac{32L_g^2}{\mu^2}\|x_{T'} - x_{T'-1}\|^2 \\
\leq & 2\left\|\hat{y}_{T'-1,K} - y_\lambda^*(x_{T'-1})\right\|^2 + \frac{32L_g^2\eta_x^2}{\mu^2},
\end{aligned} \quad (23)$$

where the first inequality follows from Young's inequality, the second inequality is due to Lemma 21, and the last inequality is based on line 14 of Algorithm 1. Following the derivation of equation (20), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{y}_{T'-1,K} - y_\lambda^*(x_{T'-1})\|^2\right] \\
=&\mathcal{O}\left(\hat{R}_{y,T'-1}^2 \exp\left(-\frac{\mu K}{L_g \ln(KL_g/\mu)}\right) + \frac{(\sigma_f + \lambda\sigma_g)^2(\ln(B_{y,T'-1}))^2(\ln(KL_g/\mu))^{\frac{2(p-1)}{p}}}{\lambda^2\mu^2 K^{\frac{2(p-1)}{p}}} + \frac{\hat{R}_{y,T'-1}^2}{K^2}\right)
\end{aligned} \quad (24)$$

for all $\hat{R}_{y,T'-1}^2 \geq \mathbb{E}[\left\|\hat{y}_{T'-1,0} - y_\lambda^*(x_{T'-1})\right\|^2]$. In addition, the induction hypothesis (21) implies that we can apply equation (24) by taking $\hat{R}_{y,T'-1}^2 = R_y^2$ to achieve

$$\begin{aligned}
&\mathbb{E}\left[\left\|\hat{y}_{T'-1,K} - y_\lambda^*(x_{T'-1})\right\|^2\right] \\
=&\mathcal{O}\left(R_y^2 \exp\left(-\frac{\mu K}{L_g \ln(KL_g/\mu)}\right) + \frac{(\sigma_f + \lambda\sigma_g)^2(\ln(B_{y,T'-1}))^2(\ln(KL_g/\mu))^{\frac{2(p-1)}{p}}}{\lambda^2\mu^2 K^{\frac{2(p-1)}{p}}} + \frac{R_y^2}{K^2}\right).
\end{aligned}$$

Combining above inequality with equation (23), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{y}_{T',0} - y_\lambda^*(x_{T'})\|^2\right] \\
\leq&\mathcal{O}\left(R_y^2 \exp\left(-\frac{\mu K}{L_g \ln(KL_g/\mu)}\right) + \frac{(\sigma_f + \lambda\sigma_g)^2(\ln(B_{y,T'-1}))^2(\ln(KL_g/\mu))^{\frac{2(p-1)}{p}}}{\lambda^2\mu^2 K^{\frac{2(p-1)}{p}}} + \frac{R_y^2}{K^2}\right) + \frac{32L_g^2\eta_x^2}{\mu^2} \\
\leq&\mathcal{O}\left(\frac{\epsilon^4}{\ell^4\kappa^6}\right) + \frac{32L_g^2\eta_x^2}{\mu^2} = \mathcal{O}\left(\frac{\epsilon^4}{\ell^4\kappa^6} + \frac{\epsilon^2}{\ell^2\kappa^4}\right) \leq R_y^2,
\end{aligned}$$

where the second inequality is based on the setting of K, the third inequality is based on the setting of η_x, and the last inequality is based on the definition of R_y^2. This completes the induction.

Therefore, we can apply Lemma 2 on $\|\hat{y}_{t,K} - y_\lambda^*(x_t)\|^2$ with $\hat{R}_{y,t} = R_y$ to obtain

$$\begin{aligned}
&\mathbb{E}\left[\|y_t - y_\lambda^*(x_t)\|^2\right] = \mathbb{E}\left[\|\hat{y}_{t,K} - y_\lambda^*(x_t)\|^2\right] \\
\leq &\mathcal{O}\left(R_y^2 \exp\left(-\frac{K\mu}{L_g \ln(KL_g/\mu)}\right) + \frac{(\sigma_f + \lambda\sigma_g)^2(\ln(B_K))^2\big(\ln\left(KL_g/\mu\right)\big)^{\frac{2(p-1)}{p}}}{\lambda^2\mu^2 K^{\frac{2(p-1)}{p}}} + \frac{R_y^2}{K^2}\right) \\
= &\mathcal{O}\left(\frac{\epsilon^4}{\ell^4\kappa^6}\right)
\end{aligned}$$

for all $t = 0, \ldots, T-1$. Consequently, we use Jensen's inequality to achieve

$$\mathbb{E}\left[\|y_t - y_\lambda^*(x_t)\|\right] \leq \sqrt{\mathbb{E}\left[\left\|y_t - y_\lambda^*(x_t)\right\|^2\right]} = \mathcal{O}\left(\frac{\epsilon^2}{\ell^2\kappa^3}\right).$$

Next, we consider the term $\mathbb{E}[\|z_t - y^*(x_t)\|]$. Recall that Algorithm 1 set $z_t = \hat{z}_{t,K}$ in line 11, where $\hat{z}_{t,K}$ can be regarded as the output of applying clipped stochastic gradient descent (lines 7–8) to minimize the function $\lambda g(x_t, y)$ with respect to y. The updates use the stochastic gradient estimate $\lambda\nabla_y G(x_t, y; \zeta_t)$ for $\lambda g(x_t, y)$. Since the function $\lambda g(x_t, y)$ is $\lambda\mu$-strongly convex in y, we apply Lemma 2 with $h(y) = \lambda g(x_t, y)$, $\sigma_h = \lambda\sigma_g$, $\ell_h = \lambda L_g$, and $\mu_h = \lambda\mu$ to obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{z}_{t,K} - y^*(x_t)\|^2\right] \\
= &\mathcal{O}\left(\hat{R}_{z,t}^2 \exp\left(-\frac{\mu K}{L_g \ln(KL_g/\mu)}\right) + \frac{\sigma_g^2(\ln(B_{z,t}))^2(\ln(KL_g/\mu))^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}} + \frac{\hat{R}_{z,t}^2}{K^2}\right).
\end{aligned} \tag{25}$$

for all $\hat{R}_{z,t}^2 \geq \mathbb{E}[\|\hat{z}_{t,0} - y^*(x_t)\|^2]$.

Similar to the derivation of equation (21), we use induction to show each $\mathbb{E}[\|\hat{z}_{t,0} - y^*(x_t)\|^2]$ has a uniform upper bound, i.e.,

$$\mathbb{E}\left[\|\hat{z}_{t,0} - y^*(x_t)\|^2\right] \leq R_z^2 \coloneqq \mathcal{O}\left(R_0^2 + \frac{\epsilon^4}{\ell^4\kappa^6} + \frac{\epsilon^2}{\ell^2\kappa^4}\right) \tag{26}$$

holds for all $t = 0, \ldots, T-1$. For the induction base $t = 0$, it follows that

$$\mathbb{E}[\|\hat{z}_{0,0} - y^*(x_0)\|^2] = \mathbb{E}[\|\hat{y}_{0,0} - y^*(x_0)\|^2] \leq R_0^2 \leq R_z^2, \tag{27}$$

where the first step follows the definition of R_0 and the second step is due to the setting of R_0. For the induction step, we suppose equation (26) holds for all $t = 0, \ldots, T'-1$. In the case of $t = T'$, we have

$$\begin{aligned}
\left\|\hat{z}_{T',0} - y^*(x_{T'})\right\|^2 \leq &2\left\|\hat{z}_{T'-1,K} - y^*(x_{T'-1})\right\|^2 + 2\left\|y^*(x_{T'}) - y^*(x_{T'-1})\right\|^2 \\
\leq &2\left\|\hat{z}_{T'-1,K} - y^*(x_{T'-1})\right\|^2 + \frac{2L_g^2}{\mu^2}\left\|x_{T'} - x_{T'-1}\right\|^2 \\
= &2\left\|\hat{z}_{T'-1,K} - y^*(x_{T'-1})\right\|^2 + \frac{2L_g^2\eta_x^2}{\mu^2},
\end{aligned} \tag{28}$$

where the first inequality follows from Young's inequality, the second inequality is due to Lemma 21, and the last inequality is based on line 14 of Algorithm 1. Following the derivation of equation (25), we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{z}_{T'-1,K}-y^*(x_{T'-1})\right\|^2\right]\\=&\mathcal{O}\left(\hat{R}^2_{z,T'-1}\exp\left(-\frac{\mu K}{L_g\ln(KL_g/\mu)}\right)+\frac{\sigma_g^2(\ln(B_{z,T'-1}))^2(\ln(KL_g/\mu))^{\frac{2(p-1)}{p}}}{\mu^2K^{\frac{2(p-1)}{p}}}+\frac{\hat{R}^2_{z,T'-1}}{K^2}\right)\end{aligned}\tag{29}$$

for all $\hat{R}^2_{z,T'-1}\geq\mathbb{E}[\left\|\hat{z}_{T'-1,0}-y^*(x_{T'-1})\right\|^2]$. In addition, the induction hypothesis (26) implies we can apply equation (29) by taking $\hat{R}^2_{z,T'-1}=R_z^2$ to achieve

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{z}_{T'-1,K}-y^*(x_{T'-1})\right\|^2\right]\\=&\mathcal{O}\left(R_z^2\exp\left(-\frac{\mu K}{L_g\ln(KL_g/\mu)}\right)+\frac{\sigma_g^2(\ln(B_{z,T'-1}))^2(\ln(KL_g/\mu))^{\frac{2(p-1)}{p}}}{\mu^2K^{\frac{2(p-1)}{p}}}+\frac{R_z^2}{K^2}\right).\end{aligned}$$

Combining above inequality with equation (28), we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{z}_{T',0}-y^*(x_{T'})\right\|^2\right]\\\leq&\mathcal{O}\left(R_z^2\exp\left(-\frac{\mu K}{L_g\ln(KL_g/\mu)}\right)+\frac{\sigma_g^2(\ln(B_{z,T'-1}))^2(\ln(KL_g/\mu))^{\frac{2(p-1)}{p}}}{\mu^2K^{\frac{2(p-1)}{p}}}+\frac{R_z^2}{K^2}\right)+\frac{2L_g^2\eta_x^2}{\mu^2}\\\leq&\mathcal{O}\left(\frac{\epsilon^4}{\ell^4\kappa^6}\right)+\frac{2L_g^2\eta_x^2}{\mu^2}\leq\mathcal{O}\left(\frac{\epsilon^4}{\ell^4\kappa^6}+\frac{\epsilon^2}{\ell^2\kappa^4}\right)\leq R_z^2,\end{aligned}$$

where the second inequality is based on the setting of K, the third inequality is based on the setting of η_x, and the last inequality is based on the definition of R_z^2. This completes the induction.

Therefore, we can apply Lemma 2 on $\|\hat{z}_{t,K}-y^*(x_t)\|^2$ with $\hat{R}_{z,t}=R_z$ to obtain

$$\begin{aligned}&\mathbb{E}\left[\|z_t-y^*(x_t)\|^2\right]=\mathbb{E}\left[\|\hat{z}_{t,K}-y^*(x_t)\|^2\right]\\=&\mathcal{O}\left(R_z^2\exp\left(-\frac{\mu K}{L_g\ln(KL_g/\mu)}\right)+\frac{\sigma_g^2(\ln(B_{z,t}))^2(\ln(KL_g/\mu))^{\frac{2(p-1)}{p}}}{\mu^2K^{\frac{2(p-1)}{p}}}+\frac{R_z^2}{K^2}\right)=\mathcal{O}\left(\frac{\epsilon^4}{\ell^4\kappa^6}\right)\end{aligned}$$

for all $t=0,\ldots,T-1$. Consequently, we use Jensen's inequality to achieve

$$\mathbb{E}\left[\|z_t-y^*(x_t)\|\right]\leq\sqrt{\mathbb{E}\left[\|z_t-y^*(x_t)\|^2\right]}=\mathcal{O}\left(\frac{\epsilon^2}{\ell^2\kappa^3}\right).$$

■

We then provide the proof of Theorem 3.

Proof We follow the parameter settings in the proof of Lemma 24 and additionally set

$$T = \mathcal{O}\left(\frac{\Delta \ell \kappa^3}{\epsilon^2}\right) \qquad \text{and} \qquad M = \mathcal{O}\left(\frac{\ell^{\frac{p}{p-1}} \kappa^{\frac{3p}{p-1}} \sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{2p}{p-1}}}\right).$$

According to Lemma 20 and the settings of λ and D_1, we have

$$\|\nabla \mathcal{L}_\lambda^*(x) - \nabla \varphi(x)\| = \mathcal{O}(\epsilon) \tag{30}$$

for all $x \in \mathbb{R}^{d_x}$. Therefore, we only need to consider the task of finding an $\mathcal{O}(\epsilon)$-stationary point of $\mathcal{L}_\lambda^*$. Recall that Lemma 1 implies

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \mathcal{L}_\lambda^*(x_t)\|] \leq & \frac{\mathbb{E}[\mathcal{L}_\lambda^*(x_0) - \mathcal{L}_\lambda^*(x_T)]}{\eta_x T} + \frac{4\sigma_f + 8\lambda\sigma_g}{M^{\frac{p-1}{p}}} + \frac{4\lambda L_g}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|y_t - y_\lambda^*(x_t)\|] \\ & + \sum_{t=0}^{T-1} \frac{2\lambda L_g \mathbb{E}[\|z_t - y^*(x_t)\|]}{T} + \frac{D_3 \eta_x}{2}.\end{aligned} \tag{31}$$

In the remainder of this proof, we will show that each term on the right-hand side of the equation (31) can be upper bounded by $\mathcal{O}(\epsilon)$ in expectation, which means that the output $\hat{x}$ is an $\mathcal{O}(\epsilon)$-stationary point of $\mathcal{L}_\lambda^*$, also an $\mathcal{O}(\epsilon)$-stationary point of φ.

According to Lemma 20 and the settings of λ and D_0, we have

$$\mathcal{L}_\lambda^*(x_0) - \inf_{x \in \mathbb{R}^{d_x}} \mathcal{L}_\lambda^*(x) = \mathcal{O}(\Delta). \tag{32}$$

Based on the settings of η_x, T, and M, the fact $D_3 = \mathcal{O}(\ell\kappa^3)$ (see Lemma 22), and equation (32), we have

$$\frac{\mathbb{E}[\mathcal{L}_\lambda^*(x_0) - \mathcal{L}_\lambda^*(x_T)]}{\eta_x T} = \mathcal{O}(\epsilon), \quad \frac{4\sigma_f + 8\lambda\sigma_g}{M^{\frac{p-1}{p}}} = \mathcal{O}(\epsilon), \quad \text{and} \quad \frac{D_3\eta_x}{2} = \mathcal{O}(\epsilon). \tag{33}$$

According to Lemma 24, we have

$$\mathbb{E}\left[\|y_t - y_\lambda^*(x_t)\|\right] = \mathcal{O}\left(\frac{\epsilon^2}{\ell^2\kappa^3}\right) \quad \text{and} \quad \mathbb{E}\left[\|z_t - y^*(x_t)\|\right] = \mathcal{O}\left(\frac{\epsilon^2}{\ell^2\kappa^3}\right) \tag{34}$$

Combining results of (31), (33), and (34), we have

$$\mathbb{E}[\|\nabla \mathcal{L}_\lambda^*(\hat{x})\|] = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \mathcal{L}_\lambda^*(x_t)\|] = \mathcal{O}(\epsilon),$$

where the first step is due to that $\hat{x}$ is uniformly sampled from $\{x_t\}_{t=1}^T$. Therefore, equation (30) implies the point $\hat{x}$ is an $\mathcal{O}(\epsilon)$-stationary point of φ in expectation. In addition, the total SFO complexity is

$$TM + TK = \tilde{\mathcal{O}}\left(\frac{\Delta \ell^{\frac{2p-1}{p-1}} \kappa^{\frac{7p-3}{p-1}} \sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{4p-2}{p-1}}} + \frac{\Delta \ell^3 \kappa^6 R_0}{\epsilon^4}\right).$$

■

B.2 The High-Probability Convergence Guarantee

In this subsection, we present the proofs for the high-probability convergence guarantee of our N^{2}SBA shown in Section 3.2. We denote the natural filtration of our method by $\mathcal{F}_t \coloneqq \sigma(\{g_{x,s}\}_{s\leq t}, \{g_{y,s,k}, g_{z,s,k}\}_{s\leq t+1, k\leq K})$.

B.2.1 Technical Lemmas

We first provide a lemma to bound the gradient norm of $\mathcal{L}_\lambda^*$ as follows.

Lemma 25 *We denote $\delta \in (0,1)$ and $\tilde{\mathcal{L}}_\lambda(x,y,z) = f(x,y) + \lambda\left(g(x,y) - g(x,z)\right)$. Under Assumptions 2.1 and 2.2, running N^{2}SBA (Algorithm 1) with $\lambda \geq 2L_f/\mu$ holds*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\mathcal{L}_\lambda^*(x_t)\right\| \leq & \frac{2(\mathcal{L}_\lambda^*(x_0) - \mathcal{L}_\lambda^*(x_T))}{\eta_x T} + 4\sum_{t=0}^{T-1}\frac{\mathbb{E}\left[\left\|g_{x,t} - \nabla_x\tilde{\mathcal{L}}_\lambda(x_t,y_t,z_t)\right\| \mid \mathcal{F}_{t-1}\right]}{T} \\ & + 4\sum_{t=0}^{T-1}\frac{\left\|\nabla_x\tilde{\mathcal{L}}_\lambda(x_t,y_t,z_t) - \nabla\mathcal{L}_\lambda^*(x_t)\right\|}{T} + D_3\eta_x \\ & + 12\left(\frac{\left\|\nabla\mathcal{L}_\lambda^*(x_0)\right\|}{T} + \eta_x D_3\right)\ln\left(\frac{2}{\delta}\right)\end{aligned} \tag{35}$$

with probability at least $1-\delta/2$.

Proof We let

$$\nu_t \coloneqq \frac{\langle\nabla\mathcal{L}_\lambda^*(x_t), g_{x,t}\rangle}{\left\|\nabla\mathcal{L}_\lambda^*(x_t)\right\|\left\|g_{x,t}\right\|}.$$

According to equation (15), we have

$$\begin{aligned}\mathcal{L}_\lambda^*(x_{t+1}) \leq & \mathcal{L}_\lambda^*(x_t) - \eta_x\left\langle\nabla\mathcal{L}_\lambda^*(x_t), \frac{g_{x,t}}{\left\|g_{x,t}\right\|}\right\rangle + \frac{D_3\eta_x^2}{2} \\ = & \mathcal{L}_\lambda^*(x_t) - \eta_x\left\|\nabla\mathcal{L}_\lambda^*(x_t)\right\|\nu_t + \frac{D_3\eta_x^2}{2}.\end{aligned}$$

Summing over above inequality with $t=0$ to $T-1$, we obtain

$$\sum_{t=0}^{T-1}\eta_x\left\|\nabla\mathcal{L}_\lambda^*(x_t)\right\|\nu_t \leq \mathcal{L}_\lambda^*(x_0) - \mathcal{L}_\lambda^*(x_T) + \frac{D_3\eta_x^2 T}{2}. \tag{36}$$

Let $\psi_t \coloneqq \mathbb{E}[\nu_t \mid \mathcal{F}_{t-1}]$ and denote $\{\Theta_t \coloneqq -\eta_x(\nu_t - \psi_t)\left\|\nabla\mathcal{L}_\lambda^*(x_t)\right\|\}_{t\in\mathbb{N}}$ as the martingale difference sequence with respect to $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$. Note that $|\nu_t| \leq 1$, then we have

$$\exp\left(\frac{\Theta_t^2}{4\eta_x^2\left\|\nabla\mathcal{L}_\lambda^*(x_t)\right\|^2}\right) = \exp\left(\frac{(\nu_t - \psi_t)^2}{4}\right) \leq \mathrm{e}.$$

Consequently, we apply Lemma 17 with $\Theta_t = -\eta_x(\nu_t - \psi_t)\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\|$, $\sigma_t = 2\eta_x\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\|$, and $\hat{\delta} = \delta/2$ to obtain that

$$\sum_{t=1}^{T} \eta_x(\psi_t - \nu_t)\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\| \leq 3\chi\eta_x^2 \sum_{t=1}^{T}\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\|^2 + \frac{1}{\chi}\ln\left(\frac{2}{\delta}\right)$$

holds with probability at least $1 - \delta/2$ for all $\chi > 0$. Rearranging above inequality, we have

$$\begin{aligned}
&\sum_{t=0}^{T-1} \eta_x(\psi_t - 3\chi\eta_x\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\|)\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\| \\
\leq &\sum_{t=1}^{T} \eta_x\nu_t\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\| + \frac{1}{\chi}\ln\left(\frac{2}{\delta}\right) \\
\leq &\mathcal{L}_\lambda^*(x_0) - \mathcal{L}_\lambda^*(x_T) + \frac{D_3\eta_x^2 T}{2} + \frac{1}{\chi}\ln\left(\frac{2}{\delta}\right)
\end{aligned} \tag{37}$$

holds with probability at least $1 - \delta/2$, where the last step is based on equation (36). In addition, the smoothness of $\mathcal{L}_\lambda^*$ shown in Lemma 22 implies

$$\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\| \leq \left\|\nabla \mathcal{L}_\lambda^*(x_0)\right\| + D_3 t\eta_x. \tag{38}$$

Then with probability at least $1 - \delta/2$, we have

$$\begin{aligned}
&\sum_{t=0}^{T-1} \eta_x\left(\psi_t - \frac{1}{2}\right)\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\| \\
\leq &\sum_{t=0}^{T-1} \eta_x\left(\psi_t - \frac{\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\|}{2(\left\|\nabla \mathcal{L}_\lambda^*(x_0)\right\| + \eta_x T D_3)}\right)\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\| \\
\leq &\mathcal{L}_\lambda^*(x_0) - \mathcal{L}_\lambda^*(x_T) + \frac{D_3\eta_x^2 T}{2} + 6(\eta_x\left\|\nabla \mathcal{L}_\lambda^*(x_0)\right\| + \eta_x^2 T D_3)\ln\left(\frac{2}{\delta}\right),
\end{aligned} \tag{39}$$

where the first inequality follows from (38) and the second inequality is due to equation (37) with $\chi = 1/(6(\eta_x\left\|\nabla \mathcal{L}_\lambda^*(x_0)\right\| + \eta_x^2 T D_3))$.

By applying Lemma 12 with $a = \nabla \mathcal{L}_\lambda^*(x_t)$ and $b = g_{x,t}$, we get

$$\psi_t\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\| \geq \left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\| - 2\mathbb{E}[\left\|g_{x,t} - \nabla \mathcal{L}_\lambda^*(x_t)\right\| \mid \mathcal{F}_{t-1}], \tag{40}$$

which implies

$$\begin{aligned}
&\sum_{t=0}^{T-1} \frac{\eta_x}{2}\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\| - 2\eta_x\mathbb{E}[\left\|g_{x,t} - \nabla \mathcal{L}_\lambda^*(x_t)\right\| \mid \mathcal{F}_{t-1}] \\
\leq &\sum_{t=0}^{T-1} \eta_x\left(\psi_t - \frac{1}{2}\right)\left\|\nabla \mathcal{L}_\lambda^*(x_t)\right\| \\
\leq &\mathcal{L}_\lambda^*(x_0) - \mathcal{L}_\lambda^*(x_T) + \frac{D_3\eta_x^2 T}{2} + 6(\eta_x\left\|\nabla \mathcal{L}_\lambda^*(x_0)\right\| + \eta_x^2 T D_3)\ln\left(\frac{2}{\delta}\right),
\end{aligned}$$

where the last step is based on (39). Rearranging above inequality, we have

$$\begin{aligned}
&\frac{1}{2}\sum_{t=0}^{T-1}\eta_x\left\|\nabla\mathcal{L}_\lambda^*(x_t)\right\| \\
\leq&\mathcal{L}_\lambda^*(x_0)-\mathcal{L}_\lambda^*(x_T)+2\eta_x\sum_{t=0}^{T-1}\mathbb{E}[\|g_{x,t}-\nabla\mathcal{L}_\lambda^*(x_t)\|\mid\mathcal{F}_{t-1}]+\frac{D_3\eta_x^2T}{2} \\
&+6(\eta_x\left\|\nabla\mathcal{L}_\lambda^*(x_0)\right\|+\eta_x^2TD_3)\ln\left(\frac{2}{\delta}\right) \\
\leq&\mathcal{L}_\lambda^*(x_0)-\mathcal{L}_\lambda^*(x_T)+2\eta_x\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|g_{x,t}-\nabla_x\tilde{\mathcal{L}}_\lambda(x_t,y_t,z_t)\right\|\mid\mathcal{F}_{t-1}\right] \\
&+2\eta_x\sum_{t=0}^{T-1}\left\|\nabla_x\tilde{\mathcal{L}}_\lambda(x_t,y_t,z_t)-\nabla\mathcal{L}_\lambda^*(x_t)\right\|+\frac{D_3\eta_x^2T}{2}+6(\eta_x\left\|\nabla\mathcal{L}_\lambda^*(x_0)\right\|+\eta_x^2TD_3)\ln\left(\frac{2}{\delta}\right).
\end{aligned}$$

where $\tilde{\mathcal{L}}_\lambda(x,y,z)=f(x,y)+\lambda\left(g(x,y)-g(x,z)\right)$. Here, we use triangle inequality in the last step. Dividing both sides by $\eta_xT/2$ leads to the desired result. ■

Under an appropriate choice of η_x, we can derive the following result.

Lemma 26 *Under Assumption 2.1 and 2.2, we run N²SBA (Algorithm 1) with $\lambda\geq 2L_f/\mu$ and $\eta_x=\sqrt{\Delta/D_3T}$. Then with probability at least $1-\delta/2$ for all $\delta\in(0,1)$, it holds*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\mathcal{L}_\lambda^*(x_t)\right\| \\
\leq&\frac{(7+50\ln(2/\delta))\sqrt{\Delta D_3}}{\sqrt{T}}+\frac{8\sigma_f+16\lambda\sigma_g}{M^{\frac{p-1}{p}}}+\frac{4}{T}\sum_{t=0}^{T-1}\left\|\nabla_x\tilde{\mathcal{L}}_\lambda(x_t,y_t,z_t)-\nabla\mathcal{L}_\lambda^*(x_t)\right\|.
\end{aligned}$$

Proof We follow the notation $\tilde{\mathcal{L}}_\lambda(x,y,z)=f(x,y)+\lambda\left(g(x,y)-g(x,z)\right)$ in the proof of Lemma 35. Note that x_t and y_t are $\mathcal{F}_{t-1}$ measurable and $\xi_{t,1}^t,\ldots,\xi_{t,M}^t$, $\zeta_{t,1}^t,\ldots,\zeta_{t,M}^t$ are independent of $\mathcal{F}_{t-1}$. Hence, we apply Lemma 15 to get

$$\begin{aligned}
&\mathbb{E}\left[\left\|g_{x,t}-\nabla_x\tilde{\mathcal{L}}_\lambda(x_t,y_t,z_t)\right\|\mid\mathcal{F}_{t-1}\right]=\mathbb{E}\left[\left\|g_{x,t}-\nabla_x\tilde{\mathcal{L}}_\lambda(x_t,y_t,z_t)\right\|\mid x_t,y_t,z_t\right] \\
\leq&\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^M\left\|\nabla_xF(x_t,y_t;\xi_{t,i})-\nabla_xf(x_t,y_t)\right\|\mid x_t,y_t,z_t\right] \\
&+\lambda\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^M\left\|\nabla_xG(x_t,y_t;\zeta_{t,i})-\nabla_xg(x_t,y_t)\right\|\mid x_t,y_t,z_t\right] \\
&+\lambda\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^M\left\|\nabla_xG(x_t,z_t;\zeta_{t,i})-\nabla_xg(x_t,z_t)\right\|\mid x_t,y_t,z_t\right] \\
\leq&\frac{2\sigma_f}{M^{\frac{p-1}{p}}}+\frac{4\lambda\sigma_g}{M^{\frac{p-1}{p}}},
\end{aligned}$$

where the first inequality follows the triangle inequality and the second inequality is due to Lemma 15 and Assumption 2.2. Plugging the above inequality into (35) yields

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\mathcal{L}_\lambda^*(x_t)\| \leq & \frac{2(\mathcal{L}_\lambda^*(x_0)-\mathcal{L}_\lambda^*(x_T))}{\eta_x T}+\frac{8\sigma_f+16\lambda\sigma_g}{M^{\frac{p-1}{p}}}+4\sum_{t=0}^{T-1}\frac{\left\|\nabla_x\tilde{\mathcal{L}}_\lambda(x_t,y_t,z_t)-\nabla\mathcal{L}_\lambda^*(x_t)\right\|}{T}\\ &+D_3\eta_x+12\left(\frac{\|\nabla\mathcal{L}_\lambda^*(x_0)\|}{T}+\eta_x D_3\right)\ln\left(\frac{2}{\delta}\right)\\ \leq & \frac{7\sqrt{\Delta D_3}}{\sqrt{T}}+\frac{8\sigma_f+16\lambda\sigma_g}{M^{\frac{p-1}{p}}}+\sum_{t=0}^{T-1}\frac{4\left\|\nabla_x\tilde{\mathcal{L}}_\lambda(x_t,y_t,z_t)-\nabla\mathcal{L}_\lambda^*(x_t)\right\|}{T}\\ &+12\left(\frac{\|\nabla\mathcal{L}_\lambda^*(x_0)\|}{T}+\frac{\sqrt{\Delta D_3}}{\sqrt{T}}\right)\ln\left(\frac{2}{\delta}\right)\\ \leq & \left(7+50\ln\left(\frac{2}{\delta}\right)\right)\frac{\sqrt{\Delta D_3}}{\sqrt{T}}+\frac{8\sigma_f+16\lambda\sigma_g}{M^{\frac{p-1}{p}}}\\ &+\sum_{t=0}^{T-1}\frac{4\left\|\nabla_x\tilde{\mathcal{L}}_\lambda(x_t,y_t,z_t)-\nabla\mathcal{L}_\lambda^*(x_t)\right\|}{T}\end{aligned} \tag{41}$$

holds with probability at least $1-\delta/2$, where the second inequality is based on the setting $\eta_x=\sqrt{\Delta/D_3T}$ and the fact $\mathcal{L}_\lambda^*(x_0)-\mathcal{L}_\lambda^*(x_T)\leq 3\Delta$ achieved by Lemma 20; the last step is based on the fact $\|\nabla\mathcal{L}_\lambda^*(x_0)\|\leq\sqrt{2\Delta D_3}$. ■

We now provide the high-probability result to bound $\|y_t-y_\lambda^*(x_t)\|$ and $\|z_t-y^*(x_t)\|$.

Lemma 27 *Following the setting of Theorem 4, it holds*

$$\|y_t-y_\lambda^*(x_t)\|=\mathcal{O}\left(\frac{\epsilon^2}{\ell^2\kappa^3}\right)\quad\text{and}\quad\|z_t-y^*(x_t)\|=\mathcal{O}\left(\frac{\epsilon^2}{\ell^2\kappa^3}\right)$$

with probability at least $1-\delta/2$ *for all* $t=0,\ldots,T-1$ *and* $\delta\in(0,1)$.

Proof We specify the parameter settings for N^{2}SBA as

$$\begin{aligned}&\eta_x=\sqrt{\frac{\Delta}{\ell\kappa^3T}},\quad \eta_{y,t}=\min\left\{\frac{1}{400(L_f+\lambda L_g)\ln(16TK/\delta)},\frac{2\ln(B_K)}{\lambda\mu K}\right\},\\ &\eta_{z,t}=\min\left\{\frac{1}{400\lambda L_g\ln(16TK/\delta)},\frac{\ln(B_K)}{\lambda\mu K}\right\},\quad K=\tilde{\mathcal{O}}\left(\frac{\ell^{\frac{p}{p-1}}\kappa^{\frac{4p}{p-1}}\sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{2p}{p-1}}}\right),\\ &\lambda=\max\left\{\frac{\kappa}{\sqrt{R}},\frac{\ell\kappa^2}{\Delta},\frac{\ell\kappa^3}{\epsilon}\right\},\quad \tau_{t,k}=\frac{\exp(-\eta_{z,t}\lambda\mu(1+k/2))R_z}{120\eta_{z,t}\ln(16TK/\delta)},\\ &\tau'_{t,k}=\frac{\exp(-\eta_{y,t}\lambda\mu(1/2+k/4))R_y}{120\eta_{y,t}\ln(16TK/\delta)},\end{aligned}$$

where we set

$$B_{y,t}=\max\left\{2,\frac{\lambda^2\mu^2K^{\frac{2(p-1)}{p}}R_y^2}{5400^{\frac{2}{p}}(\sigma_f+\lambda\sigma_g)^2(\ln(B_{y,t}))^2\left(\ln(16TK/\delta)\right)^{\frac{2(p-1)}{p}}}\right\},$$

$$B_{z,t} = \max\left\{2, \frac{\mu^2 K^{\frac{2(p-1)}{p}} R_z^2}{5400^{\frac{2}{p}}\sigma_g^2(\ln(B_{z,t}))^2\left(\ln\left(16TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}\right\},$$
$$R_y^2 \coloneqq 4R_0^2 + \frac{2\epsilon^4}{\ell^4\kappa^6} + \frac{32\epsilon^2}{\ell^2\kappa^4}, \quad R_z^2 \coloneqq R_0^2 + \frac{2\epsilon^4}{\ell^4\kappa^6} + \frac{2\epsilon^2}{\ell^2\kappa^4},$$
$$R_0 \geq \left\|\hat{y}_{0,0} - y^*(x_0)\right\|^2, \quad \Delta = \varphi(x_0) - \inf_{x\in\mathbb{R}^{d_x}}\varphi(x),$$

We first consider the term $\|y_t - y_\lambda^*(x_t)\|$. Recall that Algorithm 1 set $y_t = \hat{y}_{t,K}$ in line 10, where $\hat{y}_{t,K}$ can be regarded as the output of applying clipped stochastic gradient descent (lines 5–6) to minimize the function $\mathcal{L}_\lambda(x_t, y) = f(x_t, y) + \lambda(g(x_t, y) - g^*(x_t))$ with respect to y. The updates use the stochastic gradient estimate $\nabla_y F(x_t, y; \xi_t') + \lambda\nabla_y G(x_t, y; \zeta_t')$ for $\mathcal{L}_\lambda(x_t, y)$. Since the setting of λ and Lemma 19 implies the function $\mathcal{L}_\lambda(x_t, y)$ is $(\lambda\mu/2)$-strongly convex in y, we apply Lemma 32 with $h(y) = \mathcal{L}_\lambda(x_t, y)$, $\sigma_h = 2(\sigma_f + \lambda\sigma_g)$, $\ell_h = 2\lambda L_g$, $\mu_h = \lambda\mu/2$, and $\hat{\delta} = \delta/(4T)$ to obtain

$$\begin{aligned}\left\|\hat{y}_{t,K} - y_\lambda^*(x_t)\right\|^2 \leq& 2\hat{R}_{y,t}^2 \exp\left(-\frac{\mu K}{1600 L_g \ln\left(16TK/\delta\right)}\right)\\ &+ \frac{32\cdot 5400^{\frac{2}{p}}(\sigma_f + \lambda\sigma_g)^2(\ln(B_{y,t}))^2\left(\ln\left(16TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}{\lambda^2\mu^2 K^{\frac{2(p-1)}{p}}},\end{aligned} \tag{42}$$

for all given $\hat{R}_{y,t}^2 \geq \left\|\hat{y}_{t,0} - y_\lambda^*(x_t)\right\|^2$ with probability $1 - \delta/(4T)$ for all $t = 0, 1, \ldots, T-1$. Here, the choice of $\sigma_h = 2(\sigma_f + \lambda\sigma_g)$ that corresponds to the upper bound for the p-th central moment shown in Lemma 23.

We then use induction to show that $\left\|\hat{y}_{0,0} - y_\lambda^*(x_0)\right\|^2, \ldots, \left\|\hat{y}_{0,t} - y_\lambda^*(x_t)\right\|^2$ have a uniform upper bound for all $t = 0, \ldots, T-1$ with high-probability. In particular, we define the event E_t as

$$\left\|\hat{y}_{l,0} - y_\lambda^*(x_l)\right\|^2 \leq R_y^2 \coloneqq 4R_0^2 + \frac{2\epsilon^4}{\ell^4\kappa^6} + \frac{32\epsilon^2}{\ell^2\kappa^4} \tag{43}$$

holds for all $l = 0, 1, \ldots, t$, and target to show that $\mathbb{P}(E_t) \geq 1 - t\delta/(4T)$ holds for all $t = 0, \ldots, T-1$ by induction. For the induction base $t = 0$, we follow the derivation of equation (22) to achieve

$$\left\|\hat{y}_{0,0} - y_\lambda^*(x_0)\right\|^2 \leq 2\left\|\hat{y}_{0,0} - y^*(x_0)\right\|^2 + 2\left\|y^*(x_0) - y_\lambda^*(x_0)\right\|^2 \leq 2R_0^2 + 2R_0^2 = 4R_0^2,$$

i.e., $\mathbb{P}(E_0) = 1$. For the induction step, we suppose that $\mathbb{P}(E_t) \geq 1 - t\delta/(4T)$ holds for all $t = 0, 1, \ldots, T'-1$. In the case of $t = T'$, we follow the derivation of equation (23) to achieve

$$\left\|\hat{y}_{T',0} - y_\lambda^*(x_{T'})\right\|^2 \leq 2\left\|\hat{y}_{T'-1,K} - y_\lambda^*(x_{T'-1})\right\|^2 + \frac{32L_g^2\eta_x^2}{\mu^2}. \tag{44}$$

Following the derivation of equation (42), it holds

$$\begin{aligned}\left\|\hat{y}_{T'-1,K} - y_\lambda^*(x_{T'-1})\right\|^2 \leq& 2\hat{R}_{y,T'-1}^2 \exp\left(-\frac{\mu K}{1600 L_g \ln\left(16TK/\delta\right)}\right)\\ &+ \frac{32\cdot 5400^{\frac{2}{p}}(\sigma_f + \lambda\sigma_g)^2(\ln(B_{y,T'-1}))^2\left(\ln\left(16TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}{\lambda^2\mu^2 K^{\frac{2(p-1)}{p}}},\end{aligned} \tag{45}$$

with probability at least $1-\delta/(4T)$ for all given $\hat{R}_{y,T'-1}^2 \geq \left\|\hat{y}_{T'-1,0} - y_\lambda^*(x_{T'-1})\right\|^2$. In addition, the induction hypothesis $\mathbb{P}(E_{T'-1}) \geq 1-(T'-1)\delta/(4T)$ implies we can apply equation (45) by taking $\hat{R}_{y,T'-1}^2 = R_y^2$ and the union bound on events $E_{T'-1}$ and (45) to guarantee

$$\begin{aligned}\left\|\hat{y}_{T'-1,K} - y_\lambda^*(x_{T'-1})\right\|^2 \leq & 2R_y^2 \exp\left(-\frac{\mu K}{1600 L_g \ln\left(16TK/\delta\right)}\right) \\ & + \frac{32\cdot 5400^{\frac{2}{p}}(\sigma_f+\lambda\sigma_g)^2(\ln(B_{y,T'-1}))^2\left(\ln\left(16TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}{\lambda^2\mu^2 K^{\frac{2(p-1)}{p}}}\end{aligned}$$

holds with probability at least $1-T'\delta/(4T)$. Combining above inequality with equation (44), we guarantee

$$\begin{aligned}\left\|\hat{y}_{T',0} - y_\lambda^*(x_{T'})\right\|^2 \leq & 4R_y^2 \exp\left(-\frac{\mu K}{1600 L_g \ln\left(16TK/\delta\right)}\right) \\ & + \frac{64\cdot 5400^{\frac{2}{p}}(\sigma_f+\lambda\sigma_g)^2(\ln(B_{y,T'-1}))^2\left(\ln\left(64TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}{\lambda^2\mu^2 K^{\frac{2(p-1)}{p}}} + \frac{32L_g^2\eta_x^2}{\mu^2} \\ \leq & \frac{2\epsilon^4}{\ell^4\kappa^6} + \frac{32L_g^2\eta_x^2}{\mu^2} \leq \frac{2\epsilon^4}{\ell^4\kappa^6} + \frac{32\epsilon^2}{\ell^2\kappa^4} \leq R_y^2,\end{aligned}$$

holds with probability at least $1-T'\delta/(4T)$, i.e., $\mathbb{P}(E_{T'}) \geq 1-T'\delta/(4T)$, where the second inequality is based on the setting of K, the third inequality is based on the setting of η_x, and the last inequality is based on the definition of R_y^2. This completes the induction.

Therefore, we can apply Lemma 32 on $\|\hat{y}_{l,K} - y_\lambda^*(x_l)\|^2$ with $\hat{R}_{y,l} = R_y$ and the union bound on events E_t and equation (42) to guarantee

$$\begin{aligned}& \|y_l - y_\lambda^*(x_l)\|^2 = \|\hat{y}_{l,K} - y_\lambda^*(x_l)\|^2 \\ \leq & 2R_y^2 \exp\left(-\frac{\mu K}{1600 L_g \ln\left(16TK/\delta\right)}\right) \\ & + \frac{32\cdot 5400^{\frac{2}{p}}(\sigma_f+\lambda\sigma_g)^2(\ln(B_{y,l}))^2\left(\ln\left(16TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}{\lambda^2\mu^2 K^{\frac{2(p-1)}{p}}} = \mathcal{O}\left(\frac{\epsilon^4}{\ell^4\kappa^6}\right),\end{aligned}$$

holds for all $l = 0, \ldots, t$ with probability at least $1-(t+1)\delta/(4T)$, which implies

$$\|y_t - y_\lambda^*(x_t)\| = \mathcal{O}\left(\frac{\epsilon^2}{\ell^2\kappa^3}\right), \tag{46}$$

holds for all $t = 0, 1, \ldots, T-1$ with probability at least $1-\delta/4$.

Next, we consider the term $\|z_t - y^*(x_t)\|$. Recall that Algorithm 1 set $z_t = \hat{z}_{t,K}$ in line 11, where $\hat{z}_{t,K}$ can be regarded as the output of applying clipped stochastic gradient descent (lines 7–8) to minimize the function $\lambda g(x_t, y)$ with respect to y. The updates use the stochastic gradient estimate $\lambda\nabla_y G(x_t, y; \zeta_t)$ for $\lambda g(x_t, y)$. Since the function $\lambda g(x_t, y)$ is

$\lambda\mu$-strongly convex in y, we apply Lemma 32 with $h(y) = \lambda g(x_t, y)$, $\sigma_h = \lambda\sigma_g$, $\ell_h = \lambda L_g$, $\mu_h = \lambda\mu$, and $\hat{\delta} = \delta/(4T)$ to obtain

$$\begin{aligned}\|\hat{z}_{t,K} - y^*(x_t)\|^2 \leq & 2\hat{R}_{z,t}^2 \exp\left(-\frac{\mu K}{400 L_g \ln(16TK/\delta)}\right) \\ & + \frac{2 \cdot 5400^{\frac{2}{p}} \sigma_g^2 (\ln(B_{z,t}))^2 \left(\ln(16TK/\delta)\right)^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}}\end{aligned} \tag{47}$$

with probability at least $1 - \delta/(4T)$ for all given $\hat{R}_{z,t}^2 \geq \|\hat{z}_{t,0} - y^*(x_t)\|^2$.

We then use induction to show that $\|\hat{z}_{0,0} - y^*(x_0)\|^2, \ldots, \|\hat{z}_{0,t} - y^*(x_t)\|^2$ have a uniform upper bound for all $t = 0, \ldots, T-1$ with high-probability. In particular, we define the event E_t' as

$$\|\hat{z}_{l,0} - y^*(x_l)\|^2 \leq R_z^2 \coloneqq R_0^2 + \frac{2\epsilon^4}{\ell^4 \kappa^6} + \frac{2\epsilon^2}{\ell^2 \kappa^4} \tag{48}$$

holds for all $l = 0, 1, \ldots, t$, and target to show that $\mathbb{P}(E_t') \geq 1 - t\delta/(4T)$ holds for all $t = 0, \ldots, T-1$ by induction. For the induction base $t = 0$, we follow the derivation of equation (27) to achieve

$$\|\hat{z}_{0,0} - y^*(x_0)\|^2 = \|\hat{y}_{0,0} - y^*(x_0)\|^2 \leq R_0^2 \leq R_z^2,$$

i.e., $\mathbb{P}(E_0') = 1$. For the induction step, we suppose that $\mathbb{P}(E_t') \geq 1 - t\delta/(4T)$ holds for all $t = 0, \ldots, T-1$. In the case of $t = T'$, we follow the derivation of equation (28) to achieve

$$\left\|\hat{z}_{T',0} - y^*(x_{T'})\right\|^2 \leq 2\left\|\hat{z}_{T'-1,K} - y^*(x_{T'-1})\right\|^2 + \frac{2L_g^2\eta_x^2}{\mu^2}, \tag{49}$$

Following the derivation of equation (47), it holds

$$\begin{aligned}\left\|\hat{z}_{T'-1,K} - y^*(x_{T'-1})\right\|^2 \leq & 2\hat{R}_{z,T'-1}^2 \exp\left(-\frac{\mu K}{400 L_g \ln(16TK/\delta)}\right) \\ & + \frac{2 \cdot 5400^{\frac{2}{p}} \sigma_g^2 (\ln(B_{z,T'-1}))^2 \left(\ln(16TK/\delta)\right)^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}}\end{aligned} \tag{50}$$

with probability at least $1 - \delta/(4T)$ for all given $\hat{R}_{z,T'-1}^2 \geq \left\|\hat{z}_{T'-1,0} - y^*(x_{T'-1})\right\|^2$. In addition, the induction hypothesis $\mathbb{P}(E_{T'-1}') \geq 1 - (T'-1)\delta/(4T)$ implies we can apply equation (50) by taking $\hat{R}_{z,T'-1}^2 = R_z^2$ and the union bound on events $E_{T'-1}'$ and (50) to guarantee

$$\begin{aligned}\left\|\hat{z}_{T'-1,K} - y^*(x_{T'-1})\right\|^2 \leq & 2R_z^2 \exp\left(-\frac{\mu K}{400 L_g \ln(16TK/\delta)}\right) \\ & + \frac{2 \cdot 5400^{\frac{2}{p}} \sigma_g^2 (\ln(B_{z,T'-1}))^2 \left(\ln(16TK/\delta)\right)^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}}\end{aligned}$$

holds with probability at least $1-T'\delta/(4T)$. Combining above inequality with equation (49), we guarantee

$$\begin{aligned}\left\|\hat{z}_{T',0}-y^*(x_{T'})\right\|^2 \leq & 4R_z^2\exp\left(-\frac{\mu K}{400L_g\ln\left(16TK/\delta\right)}\right)\\ &+\frac{4\cdot 5400^{\frac{2}{p}}\sigma_g^2(\ln(B_{z,T'-1}))^2\left(\ln\left(16TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}{\mu^2K^{\frac{2(p-1)}{p}}}+\frac{2L_g^2\eta_x^2}{\mu^2}\\ \leq & \frac{2\epsilon^4}{\ell^4\kappa^6}+\frac{2L_g^2\eta_x^2}{\mu^2}\leq\frac{2\epsilon^4}{\ell^4\kappa^6}+\frac{2\epsilon^2}{\ell^2\kappa^4}\leq R_z^2,\end{aligned}$$

holds with probability at least $1-T'\delta/(4T)$, i.e., $\mathbb{P}(E'_{T'})\geq 1-T'\delta/(4T)$. Here, the second inequality is based on the setting of K, the third inequality is based on the setting of η_x, and the last inequality is based on the definition of R_z^2. This completes the induction.

Therefore, we can apply Lemma 32 on $\|\hat{z}_{l,K}-y^*(x_l)\|^2$ with $\hat{R}_{z,l}=R_z$ and the union bound on events E'_t and (47) to guarantee

$$\begin{aligned}&\|z_l-y^*(x_l)\|^2=\|\hat{z}_{l,K}-y^*(x_l)\|^2\\ \leq & 2R_z^2\exp\left(-\frac{\mu K}{400L_g\ln\left(16TK/\delta\right)}\right)\\ &+\frac{2\cdot 5400^{\frac{2}{p}}\sigma_g^2(\ln(B_{z,l}))^2\left(\ln\left(16TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}{\mu^2K^{\frac{2(p-1)}{p}}}=\mathcal{O}\left(\frac{\epsilon^4}{\ell^4\kappa^6}\right).\end{aligned}$$

holds for all $l=0,\ldots,t$ with probability at least $1-(t+1)\delta/(4T)$, which implies

$$\|z_t-y^*(x_t)\|=\mathcal{O}\left(\frac{\epsilon^2}{\ell^2\kappa^3}\right) \tag{51}$$

holds for all $t=0,\ldots,T-1$ with probability at least $1-\delta/4$.

Applying the union bound on events (46) and (51) leads to the desired result.

■

B.2.2 The Proof of Theorem 4

Proof We follow the parameter settings in the proof of Lemma 27 and additionally set

$$T=\tilde{\mathcal{O}}\left(\frac{\Delta\ell\kappa^3}{\epsilon^2}\right)\quad\text{and}\quad M=\mathcal{O}\left(\frac{\ell^{\frac{p}{p-1}}\sigma^{\frac{p}{p-1}}\kappa^{\frac{3p}{p-1}}}{\epsilon^{\frac{2p}{p-1}}}\right).$$

According to Lemma 26, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\mathcal{L}_\lambda^*(x_t)\| \\
\leq&\frac{(7+50\ln(2/\delta))\sqrt{\Delta D_3}}{\sqrt{T}}+\frac{8\sigma_f+16\lambda\sigma_g}{M^{\frac{p-1}{p}}}+\frac{4}{T}\sum_{t=0}^{T-1}\left\|\nabla_x\tilde{\mathcal{L}}_\lambda(x_t,y_t,z_t)-\nabla_x\mathcal{L}_\lambda^*(x_t)\right\| \qquad (52)\\
\leq&\frac{(7+50\ln(2/\delta))\sqrt{\Delta D_3}}{\sqrt{T}}+\frac{8\sigma_f+16\lambda\sigma_g}{M^{\frac{p-1}{p}}}+\frac{4\lambda L_g}{T}\sum_{t=0}^{T-1}(2\|y_t-y_\lambda^*(x_t)\|+\|z_t-y^*(x_t)\|)
\end{aligned}$$

with probability $1-\delta/2$, where $\tilde{\mathcal{L}}_\lambda(x,y,z)=f(x,y)+\lambda\left(g(x,y)-g(x,z)\right)$. Here, we achieve the last inequality in the derivation equation (52) as follows

$$\begin{aligned}
&\left\|\nabla_x\tilde{\mathcal{L}}_\lambda(x_t,y_t,z_t)-\nabla_x\mathcal{L}_\lambda^*(x_t)\right\| \\
\leq&\|\nabla_x f(x_t,y_t)-\nabla_x f(x_t,y_\lambda^*(x_t))\|+\lambda\|\nabla_x g(x_t,y_t)-\nabla_x f(x_t,y_\lambda^*(x_t))\| \\
&+\lambda\|\nabla_x g(x_t,z_t)-\nabla_x g(x_t,y_\lambda^*(x_t))\| \\
\leq&L_f\|y_t-y_\lambda^*(x_t)\|+\lambda L_g\|y_t-y_\lambda^*(x_t)\|+\lambda L_g\|z_t-y_\lambda^*(x_t)\| \\
\leq&2\lambda L_g\|y_t-y_\lambda^*(x_t)\|+\lambda L_g\|z_t-y_\lambda^*(x_t)\|,
\end{aligned}$$

where the first step is based on triangle inequality, the second step is based on the smoothness of f and g, and the last step is based on the setting $\lambda\geq 2L_f/\mu$.

Similar to the analysis in the proof of Theorem 3, we only need to consider the task of finding an $\mathcal{O}(\epsilon)$-stationary point of $\mathcal{L}_\lambda^*$. For the last line in equation (52), our settings of T and M guarantee

$$\frac{(7+50\ln(2/\delta))\sqrt{\Delta D_3}}{\sqrt{T}}=\mathcal{O}(\epsilon)\qquad\text{and}\qquad\frac{8\sigma_f+16\lambda\sigma_g}{M^{\frac{p-1}{p}}}=\mathcal{O}(\epsilon). \qquad (53)$$

According to Lemma 27, we have

$$\lambda L_g\|y_t-y_\lambda^*(x_t)\|=\mathcal{O}(\epsilon)\qquad\text{and}\qquad\lambda L_g\|z_t-y^*(x_t)\|=\mathcal{O}(\epsilon) \qquad (54)$$

holds for all $t=0,1,\ldots,T-1$ with probability at least $1-\delta/2$. Applying the union bound on (52), (53), and (54), it holds

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\mathcal{L}_\lambda^*(x_t)\|=\mathcal{O}(\epsilon)$$

with probability at least $1-\delta$. In addition, the total SFO complexity is

$$TM+TK=\tilde{\mathcal{O}}\left(\frac{\Delta\ell^{\frac{2p-1}{p-1}}\kappa^{\frac{7p-3}{p-1}}\sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{4p-2}{p-1}}}\right).$$

■

Appendix C. Proofs for Results of Stochastic Minimax Optimization

In this section, we establish the convergence guarantee of the N^{2}SGDA algorithm for solving the nonconvex–strongly-concave minimax optimization problem. The analysis in this section follows the counterpart for bilevel optimization, while we can omit the update on $\hat{z}_{t,k}$ to achieve the sharper complexity bounds for minimax problem.

C.1 The In-Expectation Convergence Guarantee

In this subsection, we present the proofs for the in-expectation convergence guarantee of our N^{2}SGDA shown in Section 4.

C.1.1 The Proof of Lemma 9

Proof According to Lemma 8, the function $\Phi(x)$ is $(\ell + \kappa\ell)$-smooth and we have

$$\Phi(x_{t+1}) - \Phi(x_t) - \langle x_{t+1} - x_t, \nabla\Phi(x_t)\rangle \leq (\kappa+1)\ell \left\|x_{t+1} - x_t\right\|^2. \tag{55}$$

Plugging update $x_{t+1} - x_t = -\eta_x \frac{g_{x,t}}{\|g_{x,t}\|}$ into the above inequality yields

$$\begin{aligned}\Phi(x_{t+1}) \leq& \Phi(x_t) - \eta_x \left\langle \frac{g_{x,t}}{\|g_{x,t}\|}, \nabla\Phi(x_t)\right\rangle + \eta_x^2(\kappa+1)\ell \\ \leq& \Phi(x_t) - \eta_x \left\|\nabla\Phi(x_t)\right\| + 2\eta_x \left\|g_{x,t} - \nabla\Phi(x_t)\right\| + \eta_x^2(\kappa+1)\ell,\end{aligned}$$

where the second inequality follows from Lemma 12. By the triangle inequality, one has

$$\left\|g_{x,t} - \nabla\Phi(x_t)\right\| \leq \left\|g_{x,t} - \nabla_x f(x_t, y_t)\right\| + \left\|\nabla_x f(x_t, y_t) - \nabla\Phi(x_t)\right\|.$$

Taking the expectation on both sides of the inequality, we have

$$\mathbb{E}[\|g_{x,t} - \nabla\Phi(x_t)\|] \leq \frac{2\sigma}{M^{\frac{p-1}{p}}} + \mathbb{E}[\|\nabla_x f(x_t, y_t) - \nabla\Phi(x_t)\|],$$

where the above inequality is due to Lemma 15. Then it follows that

$$\begin{aligned}\mathbb{E}[\Phi(x_{t+1})] \leq& \mathbb{E}[\Phi(x_t)] - \eta_x\mathbb{E}[\|\nabla\Phi(x_t)\|] + 2\eta_x\mathbb{E}[\|g_{x,t} - \nabla\Phi(x_t)\|] + \eta_x^2(\kappa+1)\ell \\ \leq& \mathbb{E}[\Phi(x_t)] - \eta_x\mathbb{E}[\|\nabla\Phi(x_t)\|] + \frac{4\sigma\eta_x}{M^{\frac{p-1}{p}}} \\ &+ 2\eta_x\mathbb{E}[\|\nabla_x f(x_t, y_t) - \nabla\Phi(x_t)\|] + \eta_x^2(\kappa+1)\ell.\end{aligned}$$

Summing the above inequality over $t = 0, \ldots, T-1$, we have

$$\begin{aligned}\mathbb{E}[\Phi(x_T)] \leq& \mathbb{E}[\Phi(x_0)] - \eta_x \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t)\|] + \frac{4\sigma\eta_x T}{M^{\frac{p-1}{p}}} \\ &+ 2\eta_x \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_x f(x_t, y_t) - \nabla\Phi(x_t)\|] + \eta_x^2(\kappa+1)\ell T \\ \leq& \mathbb{E}[\Phi(x_0)] - \eta_x \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(x_t)\|] + \frac{4\sigma\eta_x T}{M^{\frac{p-1}{p}}} \\ &+ 2\eta_x\ell \sum_{t=0}^{T-1} \mathbb{E}[\|y_t - y^*(x_t)\|] + \eta_x^2(\kappa+1)\ell T,\end{aligned}$$

where the second inequality follows from the ℓ-smoothness of the function f. Rearranging the above inequality and dividing both sides by $\eta_x T$, it follows that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(x_t)\|] \leq \frac{\mathbb{E}[\Phi(x_0)-\Phi(x_T)]}{\eta_x T} + \frac{4\sigma}{M^{\frac{p-1}{p}}} + \frac{2\ell\sum_{t=0}^{T-1}\mathbb{E}[\|y_t - y^*(x_t)\|]}{T} + \eta_x(\kappa+1)\ell.$$

■

C.1.2 The Proof of Theorem 10

We first provide the upper bound of $\|y_t - y^*(x_t)\|$.

Lemma 28 *Following the setting of Theorem 10, for all $t = 0, \ldots, T-1$, we have*

$$\mathbb{E}[\|y_t - y^*(x_t)\|] = \mathcal{O}\left(\frac{\epsilon}{\ell}\right).$$

Proof We first specify the parameter setting for N^{2}SGDA as

$$\eta_x = \frac{\epsilon}{\kappa\ell}, \quad \eta_{y,t} = \min\left\{\frac{1}{400\ell}, \frac{\ln(B_t)}{\mu K}\right\},$$
$$\tau_{t,k} = \frac{\exp(-\eta_{y,t}\mu(1+k/2))R_y}{120\eta_{y,t}}, \quad K = \tilde{\mathcal{O}}\left(\kappa + \left(\frac{\ell^2\sigma^2}{\mu^2\epsilon^2}\right)^{\frac{p}{2(p-1)}} + \frac{R_y\ell}{\epsilon}\right),$$

where we set

$$B_t = \max\left\{2, \frac{\mu^2 K^{\frac{2(p-1)}{p}} R_y^2}{5400^{\frac{2}{p}}\sigma^2(\ln(B_t))^2}\right\}, \quad \Delta_\Phi = \Phi(x_0) - \min_{x\in\mathbb{R}^{d_x}}\Phi(x), \quad R_y^2 \coloneqq \mathcal{O}\left(R_0^2 + \frac{\epsilon^2}{\ell^2}\right).$$

Recall that Algorithm 2 set $y_t = \hat{y}_{t,K}$ in line 8, where $\hat{y}_{t,K}$ can be regarded as the output of applying clipped stochastic gradient descent (lines 4–6) to minimize the function $-f(x_t, y)$ with respect to y. The updates use the stochastic gradient estimate $-\nabla_y F(x_t, y; \xi_t')$ for $-f(x_t, y)$. We apply Lemma 2 with $h(y) = -f(x_t, y)$, $\sigma_h = \sigma$, $\ell_h = \ell$, and $\mu_h = \mu$ to obtain

$$\mathbb{E}\left[\|\hat{y}_{t,K} - y^*(x_t)\|^2\right] \leq \mathcal{O}\left(\hat{R}_{y,t}^2 \exp\left(-\frac{K}{\kappa\ln(K\kappa)}\right) + \frac{\sigma^2(\ln(B_K))^2(\ln(K\kappa))^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}} + \frac{\hat{R}_{y,t}^2}{K^2}\right), \quad (56)$$

for all $\hat{R}_{y,t}^2 \geq \mathbb{E}[\|\hat{y}_{t,0} - y^*(x_t)\|^2]$.

We then use induction to show $\mathbb{E}[\|\hat{y}_{t,0} - y^*(x_t)\|^2]$ has the uniform upper bound, i.e.,

$$\mathbb{E}[\|\hat{y}_{t,0} - y^*(x_t)\|^2] \leq R_y^2 \coloneqq \mathcal{O}\left(\frac{\epsilon^2}{\ell^2} + R_0^2\right). \quad (57)$$

holds for all $t = 0, \ldots, T-1$. For the induction base $t = 0$, it follows from the setting of R_0 and R_y that

$$\mathbb{E}[\|\hat{y}_{0,0} - y^*(x_0)\|^2] \leq R_0^2 \leq R_y^2.$$

For the induction step, we suppose equation (21) holds for all $t = 0, \ldots, T' - 1$. In the case of $t = T'$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{y}_{T',0} - y^*(x_{T'})\right\|^2\right] \leq & 2\mathbb{E}\left[\left\|\hat{y}_{T'-1,K} - y^*(x_{T'-1})\right\|^2\right] + 2\mathbb{E}\left[\left\|y^*(x_{T'-1}) - y^*(x_{T'})\right\|^2\right] \\ \leq & 2\mathbb{E}\left[\left\|\hat{y}_{T'-1,K} - y^*(x_{T'-1})\right\|^2\right] + 2\kappa^2\left\|x_{T'-1} - x_{T'}\right\|^2 \\ = & 2\mathbb{E}\left[\left\|\hat{y}_{T'-1,K} - y^*(x_{T'-1})\right\|^2\right] + 2\kappa^2\eta_x^2.\end{aligned} \tag{58}$$

where the first inequality follows the Young's inequality, the second inequality is due to Lemma 8, and the last inequality is based on line 11 of Algorithm 2.

Following the derivation of equation (20), we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{y}_{T'-1,K} - y^*(x_{T'-1})\right\|^2\right] \\ =&\mathcal{O}\left(\hat{R}_{y,T'-1}^2 \exp\left(-\frac{K}{\kappa\ln(K\kappa)}\right) + \frac{\sigma^2(\ln(B_{T'-1}))^2(\ln(K\kappa))^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}} + \frac{\hat{R}_{y,T'-1}^2}{K^2}\right),\end{aligned} \tag{59}$$

for all $\hat{R}_{y,T'-1}^2 \geq \mathbb{E}[\left\|\hat{y}_{T'-1,0} - y^*(x_{T'-1})\right\|^2]$. In addition, the induction hypothesis (57) implies we can apply equation (59) by taking $\hat{R}_{y,T'-1}^2 = R_y^2$ to achieve

$$\mathbb{E}\left[\left\|\hat{y}_{T'-1,K} - y^*(x_{T'-1})\right\|^2\right] = \mathcal{O}\left(R_y^2 \exp\left(-\frac{K}{\kappa\ln(K\kappa)}\right) + \frac{\sigma^2(\ln(B_{T'-1}))^2(\ln(K\kappa))^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}} + \frac{R_y^2}{K^2}\right).$$

Combining above inequality with equation (58), we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{y}_{T',0} - y_\lambda^*(x_{T'})\right\|^2\right] \\ \leq&\mathcal{O}\left(R_y^2 \exp\left(-\frac{K}{\kappa\ln(K\kappa)}\right) + \frac{\sigma^2(\ln(B_{T'-1}))^2(\ln(K\kappa))^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}} + \frac{R_y^2}{K^2}\right) + 2\kappa^2\eta_x^2 \\ \leq&\mathcal{O}\left(\frac{\epsilon^2}{\ell^2}\right) \leq R_y^2,\end{aligned}$$

where the second inequality is based on the setting of K, the third inequality is based on the setting of η_x, and the last inequality is based on the definition of R_y^2. This completes the induction.

Therefore, we can apply Lemma 2 on $\|\hat{y}_{t,K} - y^*(x_t)\|^2$ with $\hat{R}_{y,t} = R_y$ to obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|y_t - y^*(x_t)\right\|^2\right] = \mathbb{E}\left[\left\|\hat{y}_{t,K} - y^*(x_t)\right\|^2\right] \\ \leq&\mathcal{O}\left(R_y^2 \exp\left(-\frac{K}{\kappa\ln(K\kappa)}\right) + \frac{\sigma^2(\ln(B_t))^2(\ln(K\kappa))^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}} + \frac{R_y^2}{K^2}\right) = \mathcal{O}\left(\frac{\epsilon^2}{\ell^2}\right).\end{aligned}$$

Consequently, we use Jensen's inequality to achieve

$$\mathbb{E}\left[\left\|y_t - y^*(x_t)\right\|\right] \leq \sqrt{\mathbb{E}\left[\left\|y_t - y^*(x_t)\right\|^2\right]} = \mathcal{O}\left(\frac{\epsilon}{\ell}\right).$$

■

We then provide the proof of Theorem 10.

Proof We follow the parameter settings in the proof of Lemma 28 and additionally set

$$T = \mathcal{O}\left(\frac{\Delta_\Phi \kappa \ell}{\epsilon^2}\right) \qquad \text{and} \qquad M = \left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}},$$

Recall that Lemma 9 implies that

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(x_t)\|] \leq & \frac{\mathbb{E}[\Phi(x_0)-\Phi(x_T)]}{\eta_x T} + \frac{4\sigma}{M^{\frac{p-1}{p}}} \\ & + \frac{2\ell\sum_{t=0}^{T-1}\mathbb{E}[\|y_t - y^*(x_t)\|]}{T} + \eta_x(\kappa+1)\ell.\end{aligned} \tag{60}$$

In the remainder of this proof, we will show that each term on the right-hand side of the equation (60) can be upper bounded by $\mathcal{O}(\epsilon)$ in expectation, which means that the output $\hat{x}$ is an $\mathcal{O}(\epsilon)$-stationary point of $\Phi(x)$.

Based on the settings of η_x, T and M, we have

$$\frac{\mathbb{E}[\Phi(x_0)-\Phi(x_T)]}{\eta_x T} = \mathcal{O}(\epsilon), \quad \frac{4\sigma}{M^{\frac{p-1}{p}}} = \mathcal{O}(\epsilon), \quad \text{and} \quad \eta_x(\kappa+1)\ell = \mathcal{O}(\epsilon). \tag{61}$$

According to Lemma 28, we have

$$\mathbb{E}[\|y_t - y^*(x_t)\|] = \mathcal{O}\left(\frac{\epsilon}{\ell}\right). \tag{62}$$

for all $t = 0, \ldots, T-1$. Combining equations (60), (61), and (62), we have

$$\mathbb{E}[\|\nabla\Phi(\hat{x})\|] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(x_t)\|] = \mathcal{O}(\epsilon),$$

where the first step is due to $\hat{x}$ is uniformly sampled from $\{x_t\}_{t=1}^T$. In addition, the total SFO complexity is

$$TM + TK = \tilde{\mathcal{O}}\left(\frac{\Delta_\Phi \kappa^{\frac{2p-1}{p-1}} \ell \sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}} + \frac{\Delta_\Phi \kappa^2 \ell}{\epsilon^2} + \frac{R_0 \Delta_\Phi \kappa \ell^2}{\epsilon^3}\right).$$

■

C.2 The High-Probability Convergence Guarantee

In this subsection, we present the proofs for the high-probability convergence guarantee of the N^{2}SGDA shown in Section 4. We denote the natural filtration of our method by $\mathcal{F}_t \coloneqq \sigma(\{g_{x,s}\}_{s\leq t}, \{g_{y,s,k}\}_{s\leq t+1, k\leq K})$.

C.2.1 Technical Lemmas

We first establish the following lemma to bound the gradient norm of Φ.

Lemma 29 *Let $\delta \in (0,1)$. Under Assumption 4.1, the iterates generated by N^{2}SGDA satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\Phi(x_t)\| \leq \frac{2\Delta_\Phi}{\eta_x T} + \sum_{t=1}^{T}\frac{4\mathbb{E}[\|g_{x,t}-\nabla_x f(x_t,y_t)\| \mid \mathcal{F}_{t-1}]}{T} + \sum_{t=1}^{T}\frac{4\|\nabla_x f(x_t,y_t)-\nabla\Phi(x_t)\|}{T} + 2\eta_x(\kappa+1)\ell + 12\left(\frac{\|\nabla\Phi(x_0)\|}{T} + 2\eta_x(\kappa+1)\ell\right)\ln\left(\frac{2}{\delta}\right)$$

with probability at least $1-\delta/2$.

Proof Let $\nu_t \coloneqq \frac{\langle\nabla\Phi(x_t), g_{x,t}\rangle}{\|\nabla\Phi(x_t)\|\|g_{x,t}\|}$. According to Lemma 9, we have

$$\Phi(x_{t+1}) \leq \Phi(x_t) - \eta_x\|\nabla\Phi(x_t)\|\nu_t + \eta_x^2(\kappa+1)\ell.$$

Summing over above inequality with $t=0$ to $T-1$, we obtain

$$\sum_{t=0}^{T-1}\eta_x\nu_t\|\nabla\Phi(x_t)\| \leq \Delta_\Phi + \eta_x^2(\kappa+1)\ell T. \tag{63}$$

Let $\psi_t \coloneqq \mathbb{E}[\nu_t \mid \mathcal{F}_{t-1}]$ and denote $\{\Theta_t \coloneqq -\eta_x(\nu_t-\psi_t)\|\nabla\Phi(x_t)\|\}_{t\in\mathbb{N}}$ as the martingale difference sequence with respect to $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$. Note that $|\nu_t| \leq 1$, then we have

$$\exp\left(\frac{\Theta_t^2}{4\eta_x^2\|\nabla\Phi(x_t)\|^2}\right) = \exp\left(\frac{(\nu_t-\psi_t)^2}{4}\right) \leq \mathrm{e},$$

Consequently, we apply Lemma 17 with $\Theta_t = -\eta_x(\nu_t-\psi_t)\|\nabla\Phi(x_t)\|$, $\sigma_t = 2\eta_x\|\nabla\Phi(x_t)\|$, and $\hat{\delta} = \delta/2$ to obtain that

$$\sum_{t=1}^{T}\eta_x(\psi_t-\nu_t)\|\nabla\Phi(x_t)\| \leq 3\chi\eta_x^2\sum_{t=1}^{T}\|\nabla\Phi(x_t)\|^2 + \frac{1}{\chi}\ln\left(\frac{2}{\delta}\right)$$

holds with probability at least $1-\delta/2$ for all $\chi > 0$. Rearranging above inequality, we have

$$\begin{aligned}&\sum_{t=0}^{T-1}\eta_x(\psi_t - 3\chi\eta_x\|\nabla\Phi(x_t)\|)\|\nabla\Phi(x_t)\| \\ \leq &\sum_{t=1}^{T}\eta_x\nu_t\|\nabla\Phi(x_t)\| + \frac{1}{\chi}\ln\left(\frac{2}{\delta}\right) \leq \Delta_\Phi + \eta_x^2(\kappa+1)\ell T + \frac{1}{\chi}\ln\left(\frac{2}{\delta}\right)\end{aligned} \tag{64}$$

holds with probability at least $1-\delta/2$, where the last step is based on equation (63). In addition, the smoothness of Φ shown in Lemma 8 implies

$$\|\nabla\Phi(x_t)\| \leq \|\nabla\Phi(x_0)\| + (\kappa+1)\ell t\eta_x. \tag{65}$$

Then with probability at least $1-\delta/2$, we have

$$\begin{aligned}
&\sum_{t=0}^{T-1} \eta_x \left(\psi_t - \frac{1}{2}\right) \|\nabla\Phi(x_t)\| \\
\leq &\sum_{t=0}^{T-1} \eta_x \left(\psi_t - \frac{\|\nabla\Phi(x_t)\|}{2(\|\nabla\Phi(x_0)\| + (\kappa+1)\ell\eta_x T)}\right) \|\nabla\Phi(x_t)\| \\
\leq &\Delta_\Phi + \eta_x^2(\kappa+1)\ell T + 6(\eta_x \|\nabla\Phi(x_0)\| + \eta_x^2 T(\kappa+1)\ell)\ln\left(\frac{2}{\delta}\right),
\end{aligned} \tag{66}$$

where the first inequality follows from (65) and the second inequality is due to equation (64) with $\chi \coloneqq 1/(6(\eta_x \|\nabla\Phi(x_0)\| + \eta_x^2 T(\kappa+1)\ell))$.

By applying Lemma 12 with $a = \nabla\Phi(x_t)$ and $b = g_{x,t}$, we get

$$\psi_t \|\nabla\Phi(x_t)\| \geq \|\nabla\Phi(x_t)\| - 2\mathbb{E}[\|g_{x,t} - \nabla\Phi(x_t)\| \mid \mathcal{F}_{t-1}]. \tag{67}$$

which implies

$$\begin{aligned}
&\sum_{t=0}^{T-1} \frac{\eta_x}{2} \|\nabla\Phi(x_t)\| - 2\eta_x\mathbb{E}[\|g_{x,t} - \nabla\Phi(x_t)\| \mid \mathcal{F}_{t-1}] \\
\leq &\sum_{t=0}^{T-1} \eta_x \left(\psi_t - \frac{1}{2}\right) \|\nabla\Phi(x_t)\| \\
\leq &\Delta_\Phi + \eta_x^2(\kappa+1)\ell T + 6(\eta_x \|\nabla\Phi(x_0)\| + \eta_x^2 T(\kappa+1)\ell)\ln\left(\frac{2}{\delta}\right),
\end{aligned}$$

where the last step is based on (66). Rearranging the above inequality, we have

$$\begin{aligned}
\frac{1}{2}\sum_{t=0}^{T-1} \eta_x \|\nabla\Phi(x_t)\| \leq &\Delta_\Phi + 2\eta_x \sum_{t=0}^{T-1} \mathbb{E}[\|g_{x,t} - \nabla\Phi(x_t)\| \mid \mathcal{F}_{t-1}] + \eta_x^2(\kappa+1)\ell T \\
&+ 6(\eta_x \|\nabla\Phi(x_0)\| + \eta_x^2 T(\kappa+1)\ell)\ln\left(\frac{2}{\delta}\right).
\end{aligned}$$

with probability at least $1-\delta/2$. Dividing both sides by $\eta_x T/2$ finishes the proof. ■

Under an appropriate choice of η_x, we can derive the following high-probability lemma.

Lemma 30 *Let $\delta \in (0,1)$. Under Assumption 4.1, we choose $\eta_x = \sqrt{\Delta_\Phi/(\varkappa+1)\ell T}$ for N^{2}SGDA. Then, with probability at least $1-\delta/2$, the iterates generated by the algorithm satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla\Phi(x_t)\| \leq \left(4 + 50\ln\left(\frac{2}{\delta}\right)\right)\frac{\sqrt{\Delta_\Phi(\kappa+1)\ell}}{\sqrt{T}} + \sum_{t=1}^{T}\frac{4\ell\|y_t - y^*(x_t)\|}{T} + \frac{8\sigma}{M^{\frac{p-1}{p}}}.$$

Proof Note that x_t and y_t are $\mathcal{F}_{t-1}$ measurable and $\xi_1^t, \ldots, \xi_M^t$ are independent of $\mathcal{F}_{t-1}$, then we have $\mathbb{E}[\|g_{x,t} - \nabla_x f(x_t, y_t)\| \mid \mathcal{F}_{t-1}] = \mathbb{E}[\|g_{x,t} - \nabla_x f(x_t, y_t)\| \mid x_t, y_t]$. Hence, applying Lemma 15 implies

$$\mathbb{E}[\|g_{x,t} - \nabla_x f(x_t, y_t)\| \mid \mathcal{F}_{t-1}] = \mathbb{E}[\|g_{x,t} - \nabla_x f(x_t, y_t)\| \mid x_t, y_t] \leq \frac{2\sigma}{M^{\frac{p-1}{p}}}.$$

Plugging above result into Lemma 29 yields

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\Phi(x_t)\| \leq & \frac{2\Delta_\Phi}{\eta_x T} + \frac{8\sigma}{M^{\frac{p-1}{p}}} + \sum_{t=1}^{T}\frac{4\|\nabla_x f(x_t, y_t) - \nabla\Phi(x_t)\|}{T} \\ & + 2\eta_x(\kappa+1)\ell + 12\left(\frac{\|\nabla\Phi(x_0)\|}{T} + 2\eta_x(\kappa+1)\ell\right)\ln\left(\frac{2}{\delta}\right).\end{aligned}$$

We choose $\eta_x = \sqrt{\Delta_\Phi/(\kappa+1)\ell T}$, then with probability at least $1-\delta/2$, it holds

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\Phi(x_t)\| \leq & \frac{4\sqrt{\Delta_\Phi(\kappa+1)\ell}}{\sqrt{T}} + \frac{8\sigma}{M^{\frac{p-1}{p}}} + \sum_{t=0}^{T-1}\frac{4\|\nabla_x f(x_t, y_t) - \nabla\Phi(x_t)\|}{T} \\ & + 12\left(\frac{\|\nabla\Phi(x_0)\|}{T} + \frac{2\sqrt{\Delta_\Phi(\kappa+1)\ell}}{\sqrt{T}}\right)\ln\left(\frac{2}{\delta}\right) \\ \leq & \left(4 + 50\ln\left(\frac{2}{\delta}\right)\right)\frac{\sqrt{\Delta_\Phi(\kappa+1)\ell}}{\sqrt{T}} + \sum_{t=0}^{T-1}\frac{4\ell\|y_t - y^*(x_t)\|}{T} + \frac{8\sigma}{M^{\frac{p-1}{p}}},\end{aligned}$$

where we use $\|\nabla\Phi(x_0)\| \leq \sqrt{2(\kappa+1)\ell\Delta_\Phi}$ to derive the last inequality. ■

C.2.2 The Proof of Theorem 11

We first provide lemmas to upper bound the term $\|y_t - y^*(x_t)\|$ with high probability.

Lemma 31 *Following the setting of Theorem 11, for all $t = 0, \ldots, T-1$, we have*

$$\|y_t - y^*(x_t)\| = \mathcal{O}\left(\frac{\epsilon}{\ell}\right), \tag{68}$$

with probability at least $1 - \delta/2$.

Proof We first specify the following hyperparameters for N2SGDA:

$$\begin{aligned}&\eta_x = \sqrt{\frac{\Delta_\Phi}{(\kappa+1)\ell T}}, \quad \eta_{y,t} = \min\left\{\frac{1}{400\ell\ln(8TK/\delta)}, \frac{\ln(B_t)}{\mu K}\right\}, \\ &\tau_{t,k} = \frac{\exp(-\eta_{y,t}\mu(1+k/2))R_y}{120\eta_{y,t}\ln(8TK/\delta)}, \quad K = \tilde{\mathcal{O}}\left(\kappa + \left(\frac{\ell^2\sigma^2}{\mu^2\epsilon^2}\right)^{\frac{p}{2(p-1)}}\right),\end{aligned}$$

where we set

$$B_t = \max\left\{2, \frac{\mu^2 K^{\frac{2(p-1)}{p}} R_y^2}{5400^{\frac{2}{p}}\sigma^2(\ln(B_t))^2\left(\ln\left(8TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}\right\},$$
$$\Delta_\Phi = \Phi(x_0) - \min_{x\in\mathbb{R}^{d_x}} \Phi(x), \quad R_y^2 = R_0^2 + \frac{10\epsilon^2}{\ell^2},$$

Recall that Algorithm 2 set $y_t = \hat{y}_{t,K}$ in line 8, where $\hat{y}_{t,K}$ can be regarded as the output of applying clipped stochastic gradient descent (lines 4–6) to minimize the function $-f(x_t, y)$ with respect to y. The updates use the stochastic gradient estimate $-\nabla_y F(x_t, y; \xi_t')$ for $-f(x_t, y)$. We apply Lemma 32 with $h(y) = -f(x_t, y)$, $\sigma_h = \sigma$, $\ell_h = \ell$, $\mu_h = \mu$, and $\hat{\delta} = \delta/(2T)$ to obtain

$$\begin{aligned}\left\|\hat{y}_{t,K} - y^*(x_t)\right\|^2 \leq & 2\hat{R}_{y,t}^2 \exp\left(-\frac{\mu K}{400\ell\ln\left(8TK/\delta\right)}\right) \\ & + \frac{2\cdot 5400^{\frac{2}{p}}\sigma^2(\ln(B_{y,t}))^2\left(\ln\left(8TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}},\end{aligned} \tag{69}$$

for all given $\hat{R}_{y,t}^2 \geq \left\|\hat{y}_{t,0} - y^*(x_t)\right\|^2$ with probability $1 - \delta/(2T)$ for all $t = 0, 1, \ldots, T-1$.

We then use induction to show that $\left\|\hat{y}_{0,0} - y^*(x_0)\right\|^2, \ldots, \left\|\hat{y}_{t,0} - y^*(x_t)\right\|^2$ for all $t = 0, \ldots, T-1$ have a uniform upper bound with high probability. In particular, we define the event E_t as

$$\left\|\hat{y}_{l,0} - y^*(x_l)\right\|^2 \leq R_y^2 \coloneqq R_0^2 + \frac{10\epsilon^2}{\ell^2} \tag{70}$$

holds for all $l = 0, 1, \ldots, t$ and target to show that $\mathbb{P}(E_t) \geq 1 - t\delta/(2T)$ for all $t = 0, \ldots, T-1$.

For the induction base $t = 0$, we follow the choice of R_0 to achieve

$$\left\|\hat{y}_{0,0} - y^*(x_0)\right\|^2 \leq R_0^2$$

that is, $\mathbb{P}(E_0) = 1$. For the induction step, we suppose that $\mathbb{P}(E_t) \geq 1 - t\delta/(2T)$ holds for all $t = 0, \ldots, T'-1$. In the case of $t = T'$, we follow the derivation of equation (58) to achieve

$$\left\|\hat{y}_{T',0} - y^*(x_{T'})\right\|^2 \leq 2\left\|\hat{y}_{T'-1,K} - y^*(x_{T'-1})\right\|^2 + 2\kappa^2\eta_x^2, \tag{71}$$

Following the derivation of equation (69), it holds

$$\begin{aligned}\left\|\hat{y}_{T'-1,K} - y^*(x_{T'-1})\right\|^2 \leq & 2\hat{R}_{y,T'-1}^2 \exp\left(-\frac{\mu K}{400\ell\ln\left(8TK/\delta\right)}\right) \\ & + \frac{2\cdot 5400^{\frac{2}{p}}\sigma^2(\ln(B_{y,T'-1}))^2\left(\ln\left(8TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}},\end{aligned} \tag{72}$$

with probability at least $1 - \delta/(2T)$ for all given $\hat{R}_{y,T'-1}^2 \geq \left\|\hat{y}_{T'-1,0} - y^*(x_{T'-1})\right\|^2$. In addition, the induction hypothesis $\mathbb{P}(E_{T-1}) \geq 1 - (T-1)\delta/(2T)$ implies we can apply

equation (72) by taking $\hat{R}^2_{y,T'-1} = R^2_y$ and union bound on $E_{T'-1}$ and (72) to guarantee

$$\begin{aligned}\left\|\hat{y}_{T'-1,K} - y^*(x_{T'-1})\right\|^2 \leq & 2R_y^2 \exp\left(-\frac{\mu K}{400\ell \ln\left(8TK/\delta\right)}\right) \\ & + \frac{2\cdot 5400^{\frac{2}{p}}\sigma^2(\ln(B_{y,T'-1}))^2\left(\ln\left(8TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}},\end{aligned}$$

holds with probability at least $1 - T'\delta/(2T)$. Combining above inequality with equation (71), we guarantee

$$\begin{aligned}\left\|\hat{y}_{T',0} - y^*(x_{T'})\right\|^2 \leq & 4R_y^2 \exp\left(-\frac{\mu K}{400\ell \ln\left(8TK/\delta\right)}\right) \\ & + \frac{4\cdot 5400^{\frac{2}{p}}\sigma^2(\ln(B_{y,T'-1}))^2\left(\ln\left(8TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}} + 2\kappa^2\eta_x^2 \\ \leq & \frac{8\epsilon^2}{\ell^2} + 2\kappa^2\eta_x^2 \leq \frac{10\epsilon^2}{\ell^2} \leq R_y^2.\end{aligned}$$

holds with probability at least $1 - T'\delta/(2T)$, i.e., $\mathbb{P}(E'_{T'}) \geq 1 - T'\delta/(2T)$, where the second inequality is based on the setting of K, the third inequality is based on the setting of η_x, and the last inequality is based on the definition of R_y^2. This completes the induction.

Therefore, we can apply Lemma 32 on $\|\hat{y}_{l,K} - y^*(x_l)\|^2$ with $\hat{R}_{y,l} = R_y$ and the union bound on E_t and (69) to guarantee

$$\begin{aligned}& \|y_l - y^*(x_l)\|^2 = \|\hat{y}_{l,K} - y^*(x_l)\|^2 \\ \leq & 2R_y^2 \exp\left(-\frac{\mu K}{400\ell \ln\left(8TK/\delta\right)}\right) \\ & + \frac{2\cdot 5400^{\frac{2}{p}}\sigma^2(\ln(B_{y,l}))^2\left(\ln\left(8TK/\delta\right)\right)^{\frac{2(p-1)}{p}}}{\mu^2 K^{\frac{2(p-1)}{p}}} = \mathcal{O}\left(\frac{\epsilon^2}{\ell^2}\right),\end{aligned}$$

holds for all $l = 0, \ldots, t$ with probability at least $1 - (t+1)\delta/(2T)$, which implies

$$\|y_t - y^*(x_t)\| = \mathcal{O}\left(\frac{\epsilon}{\ell}\right), \tag{73}$$

holds for all $t = 0, 1, \ldots, T-1$ with probability at least $1 - \delta/2$.

■

Then we provide the proof of Theorem 11.

Proof We follow the parameter settings in the proof of Lemma 31 and additionally set

$$T = \tilde{\mathcal{O}}\left(\frac{\Delta_\Phi \kappa \ell}{\epsilon^2}\right) \quad \text{and} \quad M = \left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}},$$

According to Lemma 30, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\Phi(x_t)\| \leq \frac{(4 + 50\ln(2/\delta))\sqrt{\Delta_\Phi(\kappa+1)\ell}}{\sqrt{T}} + \sum_{t=1}^{T}\frac{4\ell\|y_t - y^*(x_t)\|}{T} + \frac{8\sigma}{M^{\frac{p-1}{p}}}, \tag{74}$$

with probability $1-\delta/2$. Our choices of T and M guarantee

$$\frac{(4+50\ln(2/\delta))\sqrt{\Delta_\Phi(\kappa+1)\ell}}{\sqrt{T}}=\mathcal{O}(\epsilon) \quad \text{and} \quad \frac{8\sigma}{M^{\frac{p-1}{p}}}=\mathcal{O}(\epsilon), \tag{75}$$

According to Lemma 31, we have

$$\sum_{t=1}^{T}\frac{4\ell\left\|y_t-y^*(x_t)\right\|}{T}=\mathcal{O}(\epsilon) \tag{76}$$

with probability at least $1-\delta/2$. Combining equations (74), (75) and (76), it holds

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\Phi(x_t)\|=\mathcal{O}(\epsilon)$$

with probability at least $1-\delta$. In addition, the total SFO complexity is

$$TM+TK=\tilde{O}\left(\frac{\Delta_\Phi\kappa^{\frac{2p-1}{p-1}}\ell\sigma^{\frac{p}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}}+\frac{\Delta_\Phi\kappa^2\ell}{\epsilon^2}\right).$$

■

Appendix D. Theoretical Results for Strongly Convex Problem

In this section, we present both in-expectation and high-probability guarantees of clipped-SGD for minimizing strongly convex functions under heavy-tailed noise, which serves as our analysis for bilevel and minimax problems.

D.1 The High-Probability Convergence Analysis

The following result states the high-probability convergence guarantee of clipped SGD for a strongly convex function, as established in Lemma 3.1 in Sadiev et al. (2023).

Lemma 32 (Sadiev et al. (2023, Lemma 3.1)) *Assume the function $h(y)=\mathbb{E}[H(y;\nu)]$ is ℓ_h-smooth and μ_h-strongly convex with the stochastic index $\nu\sim\mathcal{D}_h$, and the stochastic gradient estimator $\nabla H(y;\nu)$ satisfies the conditions $\mathbb{E}[\|\nabla H(y;\nu)-\nabla h(y)\|^p]\le\sigma_h^p$ and $\mathbb{E}[\nabla H(y;\nu)]=\nabla h(y)$ for some $\sigma_h>0$. We run the clipped stochastic gradient descent*

$$\begin{cases}\nu_k\sim\mathcal{D}_h\\ \hat{y}_{k+1}=\hat{y}_k-\eta_y\text{clip}(\nabla H(\hat{y}_k;\nu_k),\tau_k)\end{cases} \tag{77}$$

for $k=0,1,\ldots,K$ with

$$\tau_k=\frac{\exp(-\eta_y\mu_h(1+k/2))\hat{R}}{120\eta_y\ln\left(4(K+1)/\hat{\delta}\right)},\quad \eta_y=\min\left\{\frac{1}{400\ell_h\ln\left(4(K+1)/\hat{\delta}\right)},\frac{\ln(B_K)}{\mu_h(K+1)}\right\},$$

$$B_K = \Theta\left(\max\left\{2, \frac{\mu_h^2 K^{\frac{2(p-1)}{p}} \hat{R}^2}{\sigma_h^2 \left(\ln\left(\max\left\{2, \frac{\mu_h^2 K^{\frac{2(p-1)}{p}} \hat{R}^2}{\sigma_h^2 (\ln(K/\hat{\delta}))^{\frac{2(p-1)}{p}}}\right\}\right)\right)^2 \left(\ln\left(4(K+1)/\hat{\delta}\right)\right)^{\frac{2(p-1)}{p}}}\right\}\right),$$

Then for all $\hat{R} \geq \|\hat{y}_0 - y^*\|$ *and* $\hat{\delta} \in (0,1)$ *satisfying* $\ln\left(4(K+1)/\hat{\delta}\right) \geq 1$*, we have*

$$\|\hat{y}_{K+1} - y^*\|^2 = \mathcal{O}\left(\max\left\{\hat{R}^2 \exp\left(-\frac{K\mu_h}{\ell_h \ln(K/\hat{\delta})}\right), \frac{\sigma_h^2 (\ln(B_K))^2 \left(\ln\left(K/\hat{\delta}\right)\right)^{\frac{2(p-1)}{p}}}{\mu_h^2 K^{\frac{2(p-1)}{p}}}\right\}\right).$$

with probability at least $1 - \hat{\delta}$*.*

For ease of the presentation, Lemma 32 shows that, by choosing parameters

$$\tau_k = \tilde{\mathcal{O}}\left(\frac{\exp(-\eta_y \mu_h k)\hat{R}}{\eta_y}\right),\ \eta_y = \tilde{\mathcal{O}}\left(\min\left\{\frac{1}{\ell_h}, \frac{1}{\mu_h K}\right\}\right),\ B_K = \tilde{\mathcal{O}}\left(\max\left\{2, \frac{\mu_h^2 K^{\frac{2(p-1)}{p}} \hat{R}^2}{\sigma_h^2}\right\}\right).$$

we have

$$\|\hat{y}_{K+1} - y^*\|^2 = \tilde{\mathcal{O}}\left(\max\left\{\hat{R}^2 \exp\left(-\frac{K\mu_h}{\ell_h}\right), \frac{\sigma_h^2}{\mu_h^2 K^{\frac{2(p-1)}{p}}}\right\}\right).$$

with probability at least $1-\hat{\delta}$ for all $\hat{R} \geq \|\hat{y}_0 - y^*\|$ and $\hat{\delta} \in (0,1)$ satisfying $\ln(4(K+1)/\hat{\delta}) \geq 1$.

D.2 The In-Expectation Convergence Guarantee (The Proof of Lemma 2)

We restate Lemma 2 below with the complete parameter specification.

Lemma 33 (Lemma 2 Restated) *Assume the function* $h(y) = \mathbb{E}[H(y;\nu)]$ *is* ℓ_h*-smooth and* μ_h*-strongly convex with the stochastic index* $\nu \sim \mathcal{D}_h$*, and the stochastic gradient estimator* $\nabla H(y;\nu)$ *satisfies the conditions* $\mathbb{E}[\nabla H(y;\nu)] = \nabla h(y)$ *and* $\mathbb{E}[\|\nabla H(y;\nu) - \nabla h(y)\|^p] \leq \sigma_h^p$ *for some* $\sigma_h > 0$*. We run the clipped stochastic gradient descent*

$$\begin{cases} \nu_k \sim \mathcal{D}_h \\ \hat{y}_{k+1} = \hat{y}_k - \eta_y \mathrm{clip}(\nabla H(\hat{y}_k; \nu_k), \tau_k) \end{cases} \tag{78}$$

for $k = 0, 1, \ldots, K$ *with*

$$\tau_k = \frac{\exp(-\eta_y \mu_h (1 + k/2))\hat{R}}{120\eta_y \ln\left(4(K+1)/\hat{\delta}\right)}, \quad \eta_y = \min\left\{\frac{1}{400\ell_h \ln\left(4(K+1)/\hat{\delta}\right)}, \frac{\ln(B_K)}{\mu_h(K+1)}\right\},$$

$$B_K = \Theta\left(\max\left\{2, \frac{\mu_h^2 K^{\frac{2(p-1)}{p}} \hat{R}^2}{\sigma_h^2 \left(\ln\left(\max\left\{2, \frac{\mu_h^2 K^{\frac{2(p-1)}{p}} \hat{R}^2}{\sigma_h^2 (\ln(K/\hat{\delta}))^{\frac{2(p-1)}{p}}}\right\}\right)\right)^2 \left(\ln\left(4(K+1)/\hat{\delta}\right)\right)^{\frac{2(p-1)}{p}}}\right\}\right),$$

where $\hat{\delta} = \mu_h^2 \ell_h^{-2} (K+1)^{-4}$. *Then for all* $\hat{R}^2 \geq \|\hat{y}_0 - y^*\|^2$, *we have*

$$\mathbb{E}[\|\hat{y}_K - y^*\|^2] = \tilde{\mathcal{O}}\left(\hat{R}^2 \exp\left(-\frac{K\mu_h}{\ell_h}\right) + \frac{\sigma_h^2}{\mu_h^2 K^{\frac{2(p-1)}{p}}} + \frac{\hat{R}^2}{K^2}\right).$$

Proof We apply the high-probability result established in Lemma 32 to derive the in-expectation result as follows. We first consider the upper bound of $\|\hat{y}_{K+1} - y^*\|^2$. The update rule of $\hat{y}_{k+1}$ implies that

$$\|\hat{y}_{k+1} - \hat{y}_k\| = \eta_y \|\mathrm{clip}(\nabla H(\hat{y}_k; \nu_k), \tau_k)\| \leq \eta_y \tau_k. \tag{79}$$

Combining equation (79) with the choice $\hat{R} \geq \|\hat{y}_0 - y^*\|$, we achieve

$$\|\hat{y}_{K+1} - y^*\| \leq \sum_{k=0}^{K} \|\hat{y}_{k+1} - \hat{y}_k\| + \|\hat{y}_0 - y^*\| \leq \eta_y \sum_{k=0}^{K} \tau_k + \hat{R}.$$

According to the parameter settings of η_y and τ_k, we have

$$\begin{aligned}
\|\hat{y}_{K+1} - y^*\| \leq & \sum_{k=0}^{K} \frac{\hat{R}\exp(-\eta_y \mu_h (1 + k/2))}{120 \ln(4(K+1)/\hat{\delta})} + \hat{R} \\
\leq & \frac{\hat{R}\exp(-\eta_y \mu_h)}{120 \ln(4(K+1)/\hat{\delta})} \cdot \frac{1}{1 - \exp(-\eta_y \mu_h/2)} + \hat{R} \\
\leq & \frac{\hat{R}}{30 \eta_y \mu_h \ln(4(K+1)/\hat{\delta})} + \hat{R}
\end{aligned}$$

where the second inequality is due to the summation of a geometric series, and the third inequality follows from $0 < \eta_y \mu_h \leq \ln(B_K)/(K+1) < 1$ and $1 - \exp(-x) \geq x/2$ for $x \in (0, 1)$. Following the choice of η_y, it holds

$$\begin{aligned}
\|\hat{y}_{K+1} - y^*\| \leq & \mathcal{O}\left(\frac{\ell_h \hat{R}}{\mu_h} + \frac{\hat{R} K}{\ln(B_K) \ln(K/\hat{\delta})} + \hat{R}\right) \\
= & \mathcal{O}\left(\frac{\ell_h \hat{R}}{\mu_h} + \frac{\hat{R} K}{\ln(B_K) \ln(K/\hat{\delta})}\right).
\end{aligned}$$

Using the fact $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\|\hat{y}_{K+1} - y^*\|^2 = \mathcal{O}\left(\frac{\ell_h^2 \hat{R}^2}{\mu_h^2} + \frac{\hat{R}^2 K^2}{(\ln(B_K))^2 (\ln(K/\hat{\delta}))^2}\right). \tag{80}$$

Lemma 32 indicates that with probability at least $1-\hat{\delta}$, one has

$$\|\hat{y}_{K+1}-y^*\|^2 = \mathcal{O}\left(\max\left\{\hat{R}^2\exp\left(-\frac{K\mu_h}{\ell_h\ln(K/\hat{\delta})}\right), \frac{\sigma_h^2(\ln(B_K))^2\big(\ln\big(K/\hat{\delta}\big)\big)^{\frac{2(p-1)}{p}}}{\mu_h^2 K^{\frac{2(p-1)}{p}}}\right\}\right). \quad (81)$$

Combining equations (80) with (81) by choosing $\hat{\delta}=\mu_h^2\ell_h^{-2}(K+1)^{-4}$, we can obtain the following in-expectation result

$$\begin{aligned}
&\mathbb{E}[\|\hat{y}_{K+1}-y^*\|^2]\\
\leq&(1-\hat{\delta})\mathcal{O}\left(\max\left\{\hat{R}^2\exp\left(-\frac{K\mu_h}{\ell_h\ln(K/\hat{\delta})}\right), \frac{\sigma_h^2(\ln(B_K))^2\big(\ln\big(K/\hat{\delta}\big)\big)^{\frac{2(p-1)}{p}}}{\mu_h^2 K^{\frac{2(p-1)}{p}}}\right\}\right)\\
&+\hat{\delta}\mathcal{O}\left(\frac{\ell_h^2\hat{R}^2}{\mu_h^2}+\frac{\hat{R}^2K^2}{(\ln(B_K))^2(\ln(K/\hat{\delta}))^2}\right)\\
\leq&\mathcal{O}\left(\max\left\{\hat{R}^2\exp\left(-\frac{K\mu_h}{\ell_h\ln(K\ell_h/\mu_h)}\right), \frac{\sigma_h^2(\ln(B_K))^2\big(\ln\big(K\ell_h/\mu_h\big)\big)^{\frac{2(p-1)}{p}}}{\mu_h^2 K^{\frac{2(p-1)}{p}}}\right\}\right)\\
&+\mathcal{O}\left(\frac{\hat{R}^2}{K^4}+\frac{\hat{R}^2\mu_h^2}{\ell_h^2K^2}\right)\\
\leq&\mathcal{O}\left(\hat{R}^2\exp\left(-\frac{K\mu_h}{\ell_h\ln(K\ell_h/\mu_h)}\right)+\frac{\sigma_h^2(\ln(B_K))^2\big(\ln\big(K\ell_h/\mu_h\big)\big)^{\frac{2(p-1)}{p}}}{\mu_h^2 K^{\frac{2(p-1)}{p}}}+\frac{\hat{R}^2}{K^2}\right)\\
=&\tilde{\mathcal{O}}\left(\hat{R}^2\exp\left(-\frac{K\mu_h}{\ell_h}\right)+\frac{\sigma_h^2}{\mu_h^2K^{\frac{2(p-1)}{p}}}+\frac{\hat{R}^2}{K^2}\right).
\end{aligned}$$

■

Appendix E. The Improved SFO Bound for F^{2}BSA

In this section, we improve the upper bound by applying stochastic gradient descent to solve the stochastic strongly convex problem provided by Chen et al. (2025, Theorem D.2). By replacing Theorem D.2 of Chen et al. (2025) with the following theorem, we directly obtain an improved SFO complexity of $\tilde{\mathcal{O}}(\kappa^{11}\sigma^2\epsilon^{-6})$ for the stochastic bilevel optimization under the bounded variance.

Theorem 34 *Suppose $h(x)\colon\mathbb{R}^d\to\mathbb{R}$ is β-smooth and α-strongly convex. Consider the following update of stochastic gradient descent*

$$x_{t+1}=x_t-\frac{1}{\beta}\nabla h(x_t;\mathcal{B}_t),$$

for $t = 0, \ldots, T-1$, where the gradient estimator satisfies

$$\mathbb{E}[\nabla h(x_t; \mathcal{B}_t)] = \nabla h(x_t) \qquad \text{and} \qquad \mathbb{E}[\|\nabla h(x_t; \mathcal{B}_t) - \nabla h(x_t)\|^2] \leq \frac{\sigma^2}{B}.$$

Then it holds that

$$\mathbb{E}[\|x_T - x^*\|^2] \leq \left(1 - \frac{\alpha}{\beta}\right)^T \mathbb{E}[\|x_0 - x^*\|^2] + \frac{\sigma^2}{\alpha\beta B}.$$

where $x^* = \arg\min_{x \in \mathbb{R}^d} h(x)$.

Proof According to the update rule, we have

$$\|x_{t+1} - x^*\|^2 = \|x_t - x^*\|^2 - \frac{2}{\beta}\langle x_t - x^*, \nabla h(x_t; \mathcal{B}_t)\rangle + \frac{1}{\beta^2}\|\nabla h(x_t; \mathcal{B}_t)\|^2.$$

Taking the expectation on both sides of above inequality, we have

$$\begin{aligned}
\mathbb{E}[\|x_{t+1} - x^*\|^2] =& \mathbb{E}[\|x_t - x^*\|^2] - \frac{2}{\beta}\mathbb{E}[\langle x_t - x^*, \nabla h(x_t)\rangle] + \frac{1}{\beta^2}\mathbb{E}[\|\nabla h(x_t; \mathcal{B}_t)\|^2] \\
\leq& \mathbb{E}[\|x_t - x^*\|^2] - \frac{2}{\beta}\mathbb{E}[\langle x_t - x^*, \nabla h(x_t)\rangle] + \frac{\sigma^2}{\beta^2 B} \\
\leq& \mathbb{E}[\|x_t - x^*\|^2] - \frac{\alpha}{\beta}\mathbb{E}[\|x_t - x^*\|^2] + \frac{2}{\beta}\mathbb{E}[h(x^*) - h(x_t)] + \frac{\sigma^2}{\beta^2 B} \\
\leq& \left(1 - \frac{\alpha}{\beta}\right)\mathbb{E}[\|x_t - x^*\|^2] + \frac{\sigma^2}{\beta^2 B},
\end{aligned}$$

where the first inequality is due to the bounded variance assumption, the second inequality follows from the strong convexity assumption. The last inequality comes from the optimality of x^*. Unrolling the above inequality for T iterations, we have

$$\mathbb{E}[\|x_T - x^*\|^2] \leq \left(1 - \frac{\alpha}{\beta}\right)^T \mathbb{E}[\|x_0 - x^*\|^2] + \frac{\sigma^2}{\alpha\beta B}.$$

■

Appendix F. Lower Bound for Stochastic Bilevel Optimization

In this section, we present the proof of the lower bound for the stochastic bilevel optimization problem shown in Section 3.3. The construction of our lower bound is based on the zero-chain property of the hard instance. In particular, we first introduce the notion of "progress function" (Arjevani et al., 2023, Section 3.1) as

$$\text{prog}_\alpha(x) \coloneqq \max\{i \geq 0 \mid |x_{(i)}| > \alpha\}$$

for some $\alpha \in [0, 1)$, where $x_{(i)}$ is the i-th coordinate of the vector x for $i \geq 1$ and we let $x_{(0)} \equiv 1$ in convention. We then introduce the probabilistic zero-chain property as follows.

Definition 35 (Arjevani et al. (2023, Definition 2)) *We say a stochastic gradient function $h(x;\xi)$ is probability-q zero-chain if it holds*

$$\mathbb{P}\big(\forall x : \mathrm{prog}_0(h(x;\xi)) \le \mathrm{prog}_{\frac{1}{4}}(x) \text{ and } h(x;\xi) = h(x_{\le \mathrm{prog}_{\frac{1}{4}}(x)};\xi)\big) \ge 1-q,$$
$$\mathbb{P}\big(\forall x : \mathrm{prog}_0(h(x;\xi)) \le 1+\mathrm{prog}_{\frac{1}{4}}(x) \text{ and } h(x;\xi) = h(x_{\le 1+\mathrm{prog}_{\frac{1}{4}}(x)};\xi)\big) = 1.$$

where the j-th coordinate of the vector $x_{\le i}$ is $(x_{\le i})_{(j)} = x_j \mathbb{1}\{i \le j\}$.

We then introduce the following stochastic nonconvex function together with its stochastic gradient estimator.

Definition 36 (Arjevani et al. (2023, Section 3.2)) *Define the function $\bar{f}^{\text{nc-s}} : \mathbb{R}^T \to \mathbb{R}$ and its stochastic gradient estimator $\hat{\nabla}\bar{F}^{\text{nc-s}}(x;\xi)$ as*

$$\bar{f}^{\text{nc-s}}(x) = -\Psi(1)\Phi(x_{(1)}) + \sum_{i=2}^{T}\big(\Psi(-x_{(i-1)})\Phi(-x_{(i)}) - \Psi(x_{(i-1)})\Phi(x_{(i)})\big)$$

and

$$(\hat{\nabla}\bar{F}^{\text{nc-s}}(x;\xi))_{(j)} = (\nabla\bar{f}^{\text{nc-s}}(x))_{(j)}\left(1 + \mathbb{1}\left\{j > \mathrm{prog}_{\frac{1}{4}}(x)\right\}\cdot\left(\frac{\xi}{q}-1\right)\right),$$

respectively, where $x = [x_{(1)},\dots,x_{(T)}]^\top \in \mathbb{R}^T$ and $\xi \sim \text{Bernoulli}(q)$ with some $q \in [0,1]$. Here, the component functions $\Psi : \mathbb{R}\to\mathbb{R}$ and $\Phi : \mathbb{R}\to\mathbb{R}$ are defined as

$$\Psi(t) = \begin{cases} 0, & t \le \frac{1}{2}, \\ \exp\left(1 - \frac{1}{(2t-1)^2}\right) & t > \frac{1}{2}, \end{cases} \qquad \text{and} \qquad \Phi(t) = \sqrt{\mathrm{e}}\int_{-\infty}^{t}\exp\left(-\frac{\tau^2}{2}\right)\mathrm{d}\tau,$$

respectively.

The stochastic first-order oracle $\hat{\nabla}\bar{F}^{\text{nc-s}}(x;\xi)$ has the following property.

Lemma 37 (Arjevani et al. (2023, Lemma 3)) *Following Definition 36, the stochastic first-order oracle $\hat{\nabla}\bar{F}^{\text{nc-s}}(x;\xi)$ is an unbiased estimator of $\nabla\bar{f}^{\text{nc-s}}(x)$ and it is a probability-q zero chain.*

Consequently, we define the following hard instance for our lower bound construction.

Definition 38 (Cutkosky et al. (2023, Section F.2)) *For all $U \in \mathrm{Orth}(d,T) \coloneqq \{U \in \mathbb{R}^{d\times T} \mid U^\top U = I_T\}$, we define the functions*

$$\bar{f}_U^{\text{nc-rs}}(x) = \bar{f}^{\text{nc-s}}(U^\top\rho(x)) + \frac{1}{10}\langle x, \rho(x)\rangle$$

and

$$\hat{\nabla}\bar{F}_U^{\text{nc-rs}}(x;\xi) = J(x)^\top U\hat{\nabla}\bar{F}^{\text{nc-s}}(U^\top\rho(x);\xi) + \frac{1}{10}(\rho(x) + J(x)^\top x),$$

where $\xi \sim \text{Bernoulli}(q)$ with some $q \in [0,1]$, $\rho(x) = x/\sqrt{1+\|x\|^2/R^2}$ with $R = 230\sqrt{T}$, and $J(x)$ is the Jacobian of ρ.

We present several useful properties for $\bar{F}^{\text{nc-rs}}$ and $\hat{\nabla}\bar{F}^{\text{nc-rs}}$) as follows

Lemma 39 (Cutkosky et al. (2023, Lemmas 25 and 26)) *Following definitions of $\bar{f}_U^{\text{nc-rs}}$ and $\hat{\nabla}\bar{F}_U^{\text{nc-rs}}$, we have:*

- *It holds $\bar{f}_U^{\text{nc-rs}}(0) - \inf_{x\in\mathbb{R}^T} \bar{f}_U^{\text{nc-rs}}(x) \le 12T$;*
- *It holds $\left\|\hat{\nabla}\bar{f}_U^{\text{nc-rs}}(x)\right\| \le 92\sqrt{T}$ for all $x \in \mathbb{R}^T$;*
- *The gradient of $\bar{f}_U^{\text{nc-rs}}(x)$ is 156-Lipschitz continuous;*
- *Let $\{x_t\}_{t\in\mathbb{N}}$ be the sequence generated by a randomized algorithm $A \in \mathcal{A}^{\text{sfo}}$. For any deterministic bijective mapping $\mathcal{M} : \mathbb{R}^T \to \mathbb{R}^T$, with probability at least $1-\delta$, it holds $\left\|\nabla\bar{f}_U^{\text{nc-rs}}(\mathcal{M}(x_t))\right\| \ge 1/2$ for all $t \le (T - \log(2/\delta))/(2q)$.*

In addition, we can show the following property of the stochastic gradient $\hat{\nabla}\bar{F}_U^{\text{nc-rs}}(x;\xi)$.

Lemma 40 *For given $p \in (1,2]$ and $U \in \text{Orth}(d,T)$, it holds*

$$\mathbb{E}[\hat{\nabla}\bar{F}_U^{\text{nc-rs}}(x;\xi)] = \nabla\bar{f}_U^{\text{nc-rs}}(x) \quad \textit{and} \quad \mathbb{E}\left[\left\|\hat{\nabla}\bar{F}_U^{\text{nc-rs}}(x;\xi) - \nabla\bar{f}_U^{\text{nc-rs}}(x)\right\|^p\right] \le \frac{2\cdot 23^p}{q^{p-1}}$$

for all $x \in \mathbb{R}^T$.

Proof We first show that $\hat{\nabla}\bar{F}^{\text{nc-s}}(x;\xi)$ is unbiased estimator of $\nabla\bar{f}^{\text{nc-s}}(x)$. By the definition of $\hat{\nabla}\bar{F}^{\text{nc-s}}(x;\xi)$ and Lemma 37, it holds

$$\mathbb{E}[(\hat{\nabla}\bar{F}^{\text{nc-s}}(x;\xi))_{(j)}] = (\nabla\bar{f}^{\text{nc-s}}(x))_{(j)} \quad \text{for all} \quad j \le \text{prog}_{\frac{1}{4}}(x);$$

and

$$\mathbb{E}[(\hat{\nabla}\bar{F}^{\text{nc-s}}(x;\xi))_{(j)}] = (\nabla\bar{f}^{\text{nc-s}}(x))_{(j)}\left(1 + \frac{\mathbb{E}[\xi]}{q} - 1\right) = (\nabla\bar{f}^{\text{nc-s}}(x))_{(j)} \quad \text{for} \quad j = \text{prog}_{\frac{1}{4}}(x) + 1.$$

Therefore, we know that $\hat{\nabla}\bar{F}^{\text{nc-s}}(x;\xi)$ is an unbiased estimator of $\nabla\bar{f}^{\text{nc-s}}(x)$.

We then bound the difference of $\hat{\nabla}\bar{F}^{\text{nc-s}}(x;\xi)$ and $\nabla\bar{f}^{\text{nc-s}}(x)$ as follows

$$\begin{aligned}
&\mathbb{E}\left[\left\|\hat{\nabla}\bar{F}^{\text{nc-s}}(x;\xi) - \nabla\bar{f}^{\text{nc-s}}(x)\right\|^p\right] \\
\le& \left\|\nabla\bar{f}^{\text{nc-s}}(x)\right\|_\infty^p \mathbb{E}\left[\left|\frac{\xi}{q} - 1\right|^p\right] \\
=& \left\|\nabla\bar{f}^{\text{nc-s}}(x)\right\|_\infty^p \left(1 - q + \frac{(1-q)^p}{q^{p-1}}\right) \\
\le& \frac{2\left\|\nabla\bar{f}^{\text{nc-s}}(x)\right\|_\infty^p}{q^{p-1}} \\
\le& \frac{2\cdot 23^p}{q^{p-1}},
\end{aligned} \tag{82}$$

where the first step is based on definitions of $\hat{\nabla}\bar{F}^{\text{nc-s}}(x;\xi)$ and $\nabla\bar{f}^{\text{nc-s}}(x)$, the second step is based on $\xi \sim \text{Bernoulli}(q)$, and the last step is based on the result $\left\|\nabla\bar{f}^{\text{nc-s}}(x)\right\|_\infty \leq 23$ from Lemma 2 of Arjevani et al. (2023).

Now, we focus on the properties $\hat{\nabla}\bar{f}^{\text{nc-rs}}$. According to the definition of $\hat{\nabla}\bar{f}^{\text{nc-rs}}$ and the unbiasedness of $\hat{\nabla}\bar{F}^{\text{nc-s}}$, it holds

$$\begin{aligned}
&\mathbb{E}[\hat{\nabla}\bar{F}_U^{\text{nc-rs}}(x;\xi)]\\
=&J(x)^\top U\mathbb{E}[\hat{\nabla}\bar{F}^{\text{nc-s}}(U^\top\rho(x);\xi)] + \frac{1}{10}(\rho(x)+J(x)^\top x)\\
=&J(x)^\top U\nabla\bar{f}^{\text{nc-s}}(U^\top\rho(x)) + \frac{1}{10}(\rho(x)+J(x)^\top x)\\
=&\nabla\bar{f}_U^{\text{nc-rs}}(x).
\end{aligned}$$

This implies that $\hat{\nabla}\bar{F}_U^{\text{nc-rs}}(x;\xi)$ is an unbiased estimator of $\nabla\bar{f}_U^{\text{nc-rs}}(x)$.

Recall that we have defined $J(x)$ as the Jacobian of the function $\rho(x) = x/\sqrt{1+\|x\|^2/R^2}$. We let $s = 1+\|x\|^2/R^2$, then it holds

$$J(x)_{i,i} = \frac{\partial\rho_i}{\partial x_i} = \frac{1}{\sqrt{s}} - \frac{x_i^2/R^2}{s^{3/2}} \qquad \text{and} \qquad J(x)_{i,j} = \frac{\partial\rho_i}{\partial x_j} = -\frac{x_i x_j/R^2}{s^{3/2}}.$$

which indicates

$$J(x) = \frac{1}{\sqrt{s}}I - \frac{1}{R^2 s^{3/2}}xx^\top = \frac{I-\rho(x)\rho(x)^\top/R^2}{\sqrt{1+\|x\|^2/R^2}}.$$

According to Lemma 13 of Arjevani et al. (2023), we know that

$$\|J(x)\| \leq 1. \tag{83}$$

Finally, bound the p-th central moment of $\hat{\nabla}\bar{F}_U^{\text{nc-rs}}(x;\xi)$ as

$$\begin{aligned}
&\mathbb{E}\left[\left\|\hat{\nabla}\bar{F}_U^{\text{nc-rs}}(x;\xi) - \nabla\bar{F}_U^{\text{nc-rs}}(x)\right\|^p\right]\\
=&\mathbb{E}\left[\left\|J(x)^\top U\hat{\nabla}\bar{F}^{\text{nc-s}}(U^\top\rho(x);\xi) - J(x)^\top U\nabla\bar{f}^{\text{nc-s}}(U^\top\rho(x))\right\|^p\right]\\
\leq&\mathbb{E}\left[\|J(x)\|^p\|U\|^p\left\|\hat{\nabla}\bar{F}^{\text{nc-s}}(U^\top\rho(x);\xi) - \nabla\bar{f}^{\text{nc-s}}(U^\top\rho(x))\right\|^p\right]\\
\leq&\frac{2\cdot 23^p}{q^{p-1}},
\end{aligned}$$

where the first inequality is due to the submultiplicativity of the matrix norm, the second inequality is based on the fact $\|U\| \leq 1$ and equations (82) and (83). ■

We now present our hard instance for the stochastic bilevel optimization problem.

F.1 Proof of Theorem 7

Proof Following Definition 38, we consider the bilevel optimization problem

$$\begin{aligned} &\min_{x\in\mathbb{R}^{d_x}} \varphi(x) \coloneqq f(x, y^*(x)), \\ &\text{s.t.} \quad y^*(x) \coloneqq \arg\min_{y\in\mathbb{R}^{d_y}} g(x, y), \end{aligned} \tag{84}$$

where

$$\begin{aligned} &f(x,y) \coloneqq \frac{L_f\beta^2}{156}\bar{f}_U^{\text{nc-rs}}\left(\frac{y}{\beta}\right), \quad g(x,y) \coloneqq \frac{\mu}{2}\|y\|^2 - L_g\langle x, y\rangle, \\ &\beta = \frac{624\epsilon}{L_f\kappa_g}, \qquad \kappa_g = \frac{L_g}{\mu}, \qquad \text{and} \qquad U \in \text{Orth}(d, T). \end{aligned} \tag{85}$$

Furthermore, we introduce the the stochastic first-order oracles for the upper- and lower-level functions $f : \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}$ and $g : \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}$ as

$$\hat{\nabla}F_U(x, y; \xi) = \left(0_T, \frac{L_f\beta}{156}\hat{\nabla}\bar{F}_U^{\text{nc-rs}}\left(\frac{y}{\beta}; \xi\right)\right) \qquad \text{and} \qquad \hat{\nabla}G(x, y) = \nabla g(x, y), \tag{86}$$

respectively . By appropriate parameter settings, we verify that the bilevel problem (84) belongs to $\mathcal{F}^{\text{nc-scq}}(C_f, L_f, L_g, \mu, \Delta)$ and the stochastic gradient estimator $\hat{\nabla}F_U(x, y; \xi)$ is unbiased and its p-th central moment is bounded by σ^p as follows.

- We take

$$T = \left\lfloor \frac{156\Delta}{12L_f\beta^2} \right\rfloor, \tag{87}$$

  then the definition of $f(x, y)$ in equation (85) implies that

$$\|\nabla f(x, y)\| = \frac{L_f\beta}{156}\left\|\nabla\bar{f}_U^{\text{nc-rs}}\left(\frac{y}{\beta}\right)\right\| \le L_f\beta\sqrt{T} \le \sqrt{13L_f\Delta} \le C_f,$$

  where the first inequality is based on Lemma 39, the second equality is based on the setting of T, and the last step is base on the condition $C_f \ge \sqrt{13L_f\Delta}$ in the statement of this theorem. Hence, we know that the function $f(x, y)$ is C_f-Lipschitz, also C_f-Lipschitz in y.

- The definition of $f(x, y)$ in equation (85) implies that

$$\left\|\nabla^2 f(x, y)\right\| = \frac{L_f}{156}\left\|\nabla^2\bar{f}_U^{\text{nc-rs}}\left(\frac{y}{\beta}\right)\right\| \le L_f,$$

  where the last inequality is due to the third statement of Lemma 39. Hence, the function $f(x, y)$ is L_f-gradient Lipschitz.

- The definition of $g(x, y)$ in equation (85) directly implies it is L_g-gradient Lipschitz and μ-strongly convex in y.

- Note that the definition

$$g(x,y) \coloneqq \frac{\mu}{2}\|y\|^2 - L_g\langle x,y\rangle,$$

implies that $y^*(x) = L_g x/\mu$. Hence, the definition

$$f(x,y) \coloneqq \frac{L_f\beta^2}{156}\bar{f}_U^{\text{nc-rs}}\left(\frac{y}{\beta}\right)$$

implies

$$\varphi(x) = f(x,y^*(x)) = \frac{L_f\beta^2}{156}\bar{f}_U^{\text{nc-rs}}\left(\frac{\kappa_g x}{\beta}\right), \tag{88}$$

where $\kappa_g \coloneqq L_g/\mu$. This implies

$$\begin{aligned}&\varphi(0) - \inf_{x\in\mathbb{R}^T}\varphi(x) \overset{(88)}{=} \frac{L_f\beta^2}{156}\left(\bar{f}_U^{\text{nc-rs}}(0) - \inf_{x\in\mathbb{R}^T}\bar{f}_U^{\text{nc-rs}}\left(\frac{\kappa_g x}{\beta}\right)\right)\\ \le&\frac{L_f\beta^2}{156}\cdot 12T \le \Delta,\end{aligned}$$

where the last line is due to first statement in Lemma 39 and the setting of T in (87).

- We take

$$q = 2^{\frac{1}{p-1}}\left(\frac{23L_f\beta}{156\sigma}\right)^{\frac{p}{p-1}},$$

then the p-th central moment of $\hat{\nabla}F$ is bounded as

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{\nabla}F_U(x,y;\xi) - \nabla f(x,y)\right\|^p\right]\\ =&\left(\frac{L_f\beta}{156}\right)^p\mathbb{E}\left[\left\|\hat{\nabla}\bar{F}_U^{\text{nc-rs}}\left(\frac{y}{\beta};\xi\right) - \nabla\bar{f}_U^{\text{nc-rs}}\left(\frac{y}{\beta}\right)\right\|^p\right]\\ \le&\left(\frac{L_f\beta}{156}\right)^p\cdot\frac{2\cdot 23^p}{q^{p-1}} = \sigma^p,\end{aligned}$$

where the last step is based on Lemma 40 and the setting of q.

Following Definition 6, we consider the randomized algorithm A consists of the mappings $\{A_t\}_{t\in\mathbb{N}}$ recursively defined as

$$(x_t,y_t) = A_t(r,\hat{\nabla}\bar{F}_U^{\text{nc-rs}}(y_0;\xi),\hat{\nabla}G(x_0,y_0),\ldots,\hat{\nabla}\bar{F}_U^{\text{nc-rs}}(y_{t-1};\xi),\hat{\nabla}G(x_{t-1},y_{t-1})),$$

where r is the random seed. Note that $\hat{\nabla}G(x,y) = \nabla g(x,y)$ is a deterministic mapping of (x,y), then there exists mappings $\{A_t^1\}_{t\in\mathbb{N}}$ satisfying the recursion

$$(x_t,y_t) = A_t^1(r,\hat{\nabla}\bar{F}_U^{\text{nc-rs}}(y_0;\xi),x_0,y_0,\ldots,\hat{\nabla}\bar{F}_U^{\text{nc-rs}}(y_{t-1};\xi),x_{t-1},y_{t-1}).$$

This implies there exist mappings $\{A_t^2\}_{t\in\mathbb{N}}$ satisfying

$$(x_t, y_t) = A_t^2(r, \hat{\nabla}\bar{F}_U^{\text{nc-rs}}(y_0;\xi), \ldots, \hat{\nabla}\bar{F}_U^{\text{nc-rs}}(y_{t-1};\xi)).$$

Recall that equation (88) implies

$$\nabla\varphi(x) = \frac{L_f\beta\kappa_g}{156}\nabla\bar{f}_U^{\text{nc-rs}}\left(\frac{\kappa_g x}{\beta}\right). \tag{89}$$

We take

$$\beta = \frac{624\epsilon}{L_f\kappa_g},$$

then using Lemma 39 with $\delta = 1/2$ implies

$$\mathbb{E}[\|\nabla\varphi(x_t)\|] \overset{(89)}{=} \frac{L_f\beta\kappa_g}{156}\mathbb{E}\left[\left\|\nabla\bar{f}_U^{\text{nc-rs}}\left(\frac{\kappa_g x_t}{\beta}\right)\right\|\right] \geq \frac{L_f\beta\kappa_g}{624} = \epsilon \tag{90}$$

for all $t \leq (T - \log(2/\delta))/(2q)$. Therefore, achieving an ϵ-stationary point of the function $\varphi(x)$ requires at least the SFO complexity of

$$\Omega\left(\frac{T-\log(2/\delta)}{2q}\right) = \Omega\left(\frac{T}{q}\right) = \Omega\left(\frac{\Delta}{L_f\beta^2}\cdot\frac{\sigma^{\frac{p}{p-1}}}{L_f^{\frac{p}{p-1}}\beta^{\frac{p}{p-1}}}\right) = \Omega\left(\frac{\Delta L_f\sigma^{\frac{p}{p-1}}\kappa_g^{\frac{3p-2}{p-1}}}{\epsilon^{\frac{3p-2}{p-1}}}\right).$$

■

Appendix G. Discussion on the Concurrent Work

In a concurrent work, Zhang et al. (2025b) studied decentralized stochastic bilevel optimization under heavy-tailed noise. We clarify the difference between our setting and that of related work as follows.

- The concurrent work proposed a normalized stochastic gradient descent algorithm based on momentum-based variance reduction (Cutkosky and Orabona, 2019), which requires all stochastic components $F(x, y; \xi)$ and $G(x, y; \zeta)$ have the Lipschitz continuous gradient (Zhang et al., 2025b, Assumption 3.1). In contrast, our algorithm design and analysis only require the upper-level function $f(x, y)$ and the lower-level function $g(x, y)$ have the Lipschitz continuous gradient, which is weaker than that of Zhang et al. (2025b).
- The concurrent work assumes both the lower-level function $g(x, y)$ and the penalized function $\delta f(x, y) + g(x, y)$ satisfies Polyak–Lojasiewicz (PL) condition with respect to y (Zhang et al., 2025b, Assumption 3.1), where $\delta > 0$ is the penalized parameter which is similar to the role of $1/\lambda$ in our analysis. Although the PL condition on $g(x, y)$ is weaker than the strong convexity used in our work, the PL condition on $\delta f(x, y) + g(x, y)$ is difficult to verify since the choice of their δ depends on the iteration number of the algorithm (Zhang et al., 2025b, Theorem 4.5).

References

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *International Conference on Learning Representations*, 2024.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Lesi Chen and Jingzhao Zhang. On the condition number dependency in bilevel optimization. *arXiv preprint arXiv:2511.22331*, 2025.

Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *Conference on Learning Theory*, pages 947–980, 2024.

Lesi Chen, Yaohua Ma, and Jingzhao Zhang. Near-optimal nonconvex-strongly-convex bilevel optimization with fully first-order oracles. *Journal of Machine Learning Research*, 26(109):1–56, 2025.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *Advances in Neural Information Processing Systems*, volume 34, pages 25294–25307, 2021.

Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2466–2488, 2022.

Xuxing Chen, Minhui Huang, Shiqian Ma, and Krishna Balasubramanian. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *International Conference on Machine Learning*, pages 4641–4671, 2023.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *International Conference on Machine Learning*, pages 2260–2268, 2020.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, volume 32, pages 15236–15245, 2019.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pages 6643–6670, 2023.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge University Press, 2019.

Kacha Dzhaparidze and J.H. Van Zanten. On Bernstein-type inequalities for Martingales. *Stochastic Processes and Their Applications*, 93(1):109–117, 2001.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1082–1092, 2020.

Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning: Methods, systems, challenges*, pages 3–33. Springer International Publishing Cham, 2019.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135, 2017.

Adrien Fradin, Abdurakhmon Sadiev, Laurent Condat, and Peter Richtárik. Tight lower bounds and optimal algorithms for stochastic nonconvex optimization with heavy-tailed noise. In *International Conference on Artificial Intelligence and Statistics*, pages 2548–2556, 2026.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577, 2018.

David A. Freedman. On tail probabilities for Martingales. *the Annals of Probability*, pages 100–118, 1975.

Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, Zico Kolter, Zachary Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *International Conference on Machine Learning*, pages 3610–3619, 2021.

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

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Xiaochuan Gong, Jie Hao, and Mingrui Liu. An accelerated algorithm for stochastic bilevel optimization under unbounded smoothness. *Advances in Neural Information Processing Systems*, 37:78201–78243, 2024a.

Xiaochuan Gong, Jie Hao, and Mingrui Liu. A nearly optimal single loop algorithm for stochastic bilevel optimization under unbounded smoothness. In *International Conference on Machine Learning*, pages 15854–15892, 2024b.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

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

Eduard Gorbunov, Marina Danilova, David Dobre, Pavel Dvurechenskii, Alexander Gasnikov, and Gauthier Gidel. Clipped stochastic methods for variational inequalities with heavy-tailed noise. In *Advances in Neural Information Processing Systems*, volume 35, pages 31319–31332, 2022.

Eduard Gorbunov, Abdurakhmon Sadiev, Marina Danilova, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. In *International Conference on Machine Learning*, pages 15951–16070, 2023.

Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. High-probability complexity bounds for non-smooth stochastic convex optimization with heavy-tailed noise. *Journal of Optimization Theory and Applications*, pages 1–60, 2024.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758, 2020.

Jie Hao, Xiaochuan Gong, and Mingrui Liu. Bilevel optimization under unbounded smoothness: A new algorithm and convergence analysis. In *International Conference on Learning Representations*, 2024.

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.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Feihu Huang. Optimal Hessian/Jacobian-free nonconvex-PL bilevel optimization. In *International Conference on Machine Learning*, pages 19598–1962, 2024.

Minhui Huang, Xuxing Chen, Kaiyi Ji, Shiqian Ma, and Lifeng Lai. Efficiently escaping saddle points in bilevel optimization. *Journal of Machine Learning Research*, 26(1):1–61, 2025.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *International Conference on Artificial Intelligence and Statistics*, pages 2413–2421, 2024.

Kaiyi Ji and Yingbin Liang. Lower bounds and accelerated algorithms for bilevel optimization. *Journal of Machine Learning Research*, 24(22):1–56, 2023.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892, 2021.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Theoretical convergence of multi-step model-agnostic meta-learning. *Journal of Machine Learning Research*, 23(29):1–41, 2022.

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889, 2020.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in Neural Information Processing Systems*, 34:30271–30283, 2021.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D. Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages 18083–18113, 2023.

Yassine Laguel, Necdet Serhat Aybat, and Mert Gürbüzbalaban. High probability and risk-averse guarantees for stochastic saddle point problems. *arXiv preprint arXiv:2304.00444*, 2023.

Yassine Laguel, Yasa Syed, Necdet Serhat Aybat, and Mert Gürbüzbalaban. High-probability complexity guarantees for nonconvex minimax problems. In *Advances in Neural Information Processing System*, volume 38, pages 140937–140969, 2024.

Paul Lévy. *Calcul des probabilités*. Gauthier-Villars, 1925.

Haochuan Li, Yi Tian, Jingzhao Zhang, and Ali Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. In *Advances in Neural Information Processing Systems*, volume 34, pages 1792–1804, 2021.

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive SGD with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093, 2020.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. BOME! bilevel optimization made easy: A simple first-order approach. In *Advances in Neural Information Processing Systems*, volume 35, pages 17248–17262, 2022.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *International Conference on Learning Representations*, 2019.

Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. In *International Conference on Machine Learning*, pages 21884–21914, 2023.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, 2025.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Benoit Mandelbrot. The Pareto-Levy law and the distribution of income. *International Economic Review*, 1(2):79–106, 1960.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. High probability convergence of clipped-SGD under heavy-tailed noise. *arXiv preprint arXiv:2302.05437*, 2023.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D. Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, volume 32, pages 14934–14942, 2019.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning*, pages 737–746, 2016.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI Blog 1.8*, 2019.

Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing systems*, volume 32, pages 113–124, 2019.

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

Soroosh Shafieezadeh Abadeh, Peyman M. Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, volume 28, pages 1576–1584, 2015.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, pages 30992–31015, 2023.

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

Xingyou Song, Wenbo Gao, Yuxiang Yang, Krzysztof Choromanski, Aldo Pacchiano, and Yunhao Tang. ES-MAML: Simple Hessian-free meta learning. In *International Conference on Learning Representations*, 2020.

Daouda Sow, Kaiyi Ji, and Yingbin Liang. On the convergence theory for Hessian-free bilevel algorithms. In *Advances in Neural Information Processing Systems*, volume 35, pages 4136–4149, 2022.

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

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*, volume 135. MIT press Cambridge, 1998.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An instruction-following LLaMa model, 2023.

Davoud Ataee Tarzanagh, Mingchen Li, Christos Thrampoulidis, and Samet Oymak. FedNest: Federated bilevel, minimax, and compositional optimization. In *International Conference on Machine Learning*, pages 21146–21179, 2022.

Nuozhou Wang, Junyu Zhang, and Shuzhong Zhang. Efficient first order method for saddle point problems with higher order smoothness. *SIAM Journal on Optimization*, 34(4): 3342–3370, 2024.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(7), 2009.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 34, pages 13670–13682, 2021.

Shuoguang Yang, Xuezhou Zhang, and Mengdi Wang. Decentralized gossip-based stochastic bilevel optimization over communication networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 238–252, 2022.

Wei Yao, Haian Yin, Shangzhi Zeng, and Jin Zhang. Overcoming lower-level constraints in bilevel optimization: A novel approach with regularized gap functions. In *International Conference on Learning Representations*, 2025.

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.

Miao Zhang, Steven W. Su, Shirui Pan, Xiaojun Chang, Ehsan M. Abbasnejad, and Reza Haffari. iDARTS: Differentiable architecture search with stochastic implicit gradients. In *International Conference on Machine Learning*, pages 12557–12566, 2021.

Siqi Zhang, Xing Huang, and Feihu Huang. Generalized smooth bilevel optimization with nonconvex lower-level. In *International Conference on Machine Learning*, 2025a.

Xinwen Zhang, Yihan Zhang, and Hongchang Gao. Nonconvex decentralized stochastic bilevel optimization under heavy-tailed noises. *arXiv preprint arXiv:2509.15543*, 2025b.

Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.