INDEPENDENTLY-NORMALIZED SGD FOR GENERALIZED-SMOOTH NONCONVEX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies have shown that many nonconvex machine learning problems meet a so-called generalized-smooth condition that extends beyond traditional smooth nonconvex optimization. However, the existing algorithms designed for generalized-smooth nonconvex optimization encounter significant limitations in both their design and convergence analysis. In this work, we first study deterministic generalized-smooth nonconvex optimization and analyze the convergence of normalized gradient descent under the generalized Polyak-Łojasiewicz condition. Our results provide a comprehensive understanding of the interplay between gradient normalization and function geometry. Then, for stochastic generalized-smooth nonconvex optimization, we propose an independently-normalized stochastic gradient descent algorithm, which leverages independent sampling, gradient normalization and clipping to achieve an $\mathcal{O}(\epsilon^{-4})$ sample complexity under relaxed assumptions. Experiments demonstrate the fast convergence of our algorithm.

1 INTRODUCTION

In modern machine learning, the convergence of gradient-based optimization algorithms has been well studied in the standard smooth nonconvex setting. However, it has been shown recently that smoothness fails to characterize the global geometry of many nonconvex machine learning problems, including distributionally-robust optimization (DRO)(Levy et al., 2020; Jin et al., 2021), meta-learning (Nichol et al., 2018; Chayti & Jaggi, 2024) and language models (Liu et al., 2023; Zhang et al., 2019). Instead, these problems have been shown to satisfy a so-called *generalized-smooth* condition, in which the smoothness parameter can scale with the gradient norm in the optimization process (Zhang et al., 2019).

In the existing literature, various works have proposed different algorithms for solving generalized-smooth nonconvex optimization problems. Specifically, one line of work focuses on the classic stochastic gradient descent (SGD) algorithm (Li et al., 2024; Reisizadeh et al., 2023). However, the convergence of SGD either relies on adopting very large batch size or involves large constants, and the practical performance of SGD is often poor due to the ill-conditioned smoothness parameter when gradient is large. Another line of work focuses on clipped SGD, which adapts to the generalized-smooth geometry by leveraging gradient clipping and normalization (Zhang et al., 2019; 2020). However, to establish convergence guarantee, these studies rely on the strong assumption that the stochastic approximation error is bounded almost surely.

Motivated by the algorithmic and theoretical limitations discussed above, this work aims to explore the interplay between algorithm design and the geometry of generalized-smooth functions, and develop algorithms tailored for generalized-smooth nonconvex optimization. To achieve this overarching goal, we need to address several fundamental challenges. First, even in deterministic generalized-smooth nonconvex op-

timization, there is limited knowledge about how to adapt gradient-based algorithms to the geometry of generalized-smooth problems. Thus, we want to answer the following question.

- *Q1: In deterministic nonconvex optimization, how can we adapt algorithm hyperparameters to align with the Polyak-Łojasiewicz geometry of generalized-smooth problems? What are the convergence rates?*

Second, in stochastic generalized-smooth nonconvex optimization, the existing SGD-type algorithms are either impractical due to poor performance or relying on strong assumptions to establish convergence guarantee. Therefore, we aim to answer the following question.

- *Q2: Can we develop a novel algorithm tailored for stochastic generalized-smooth optimization that achieves fast convergence in practice while providing convergence guarantee under relaxed assumptions?*

In this work, we provide comprehensive answers to the above questions and develop new algorithms as well as convergence analysis in generalized-smooth nonconvex optimization. In light of the above discussions, we summarize our key contributions as following.

1.1 Our Contribution

We first consider deterministic generalized-smooth nonconvex optimization, and study the convergence of normalized gradient descent under the generalized Polyak-Łojasiewicz (PŁ) condition. Our result characterizes the algorithm convergence rate under a broad spectrum of function geometry characterized by the generalized-smooth and PŁ conditions, and provides deep insights into adapting algorithm hyper-parameters (such as learning rate and gradient normalization scale) to the underlying function geometry.

We then consider stochastic generalized-smooth nonconvex optimization, for which we propose a novel Independently-Normalized Stochastic Gradient Descent (I-NSGD) algorithm. Specifically, I-NSGD leverages normalized gradient updates with independent sampling and gradient clipping to reduce the bias and enhance stability. Consequently, we can establish convergence of I-NSGD with $\mathcal{O}(\epsilon^{-4})$ sample complexity under a relaxed assumption on the approximation error of stochastic gradient and constant-level batch size. This makes the algorithm well-suited for solving large-scale problems. We further study the convergence behavior of I-NSGD under the generalized PŁ condition.

We compare the numerical performance of our I-NSGD algorithm with other state-of-the-art stochastic algorithms in applications of nonconvex phase retrieval and nonconvex distributionally-robust optimization, both of which are generalized-smooth nonconvex problems. Our results demonstrate the efficiency of I-NSGD in solving generalized-smooth nonconvex problems and match our theoretical guidance.

1.2 Related Work

Generalized-Smoothness. The concept of generalized-smoothness was introduced by Zhang et al. (2019) with the (L_0, L_1)-smooth condition, which allows a function to either have an affine-bounded hessian norm or be locally L-smooth within a specific region. This idea was extended by Chen et al. (2023), who proposed the $\mathcal{L}^*_{asym}(\alpha)$ and $\mathcal{L}^*_{sym}(\alpha)$ conditions, controlling gradient changes globally with both a constant term and a gradient-dependent term associated with power α, thus applying more broadly. Later, Li et al. (2024) introduced ℓ-smoothness, which use a non-decreasing sub-quadratic polynomial to control gradient differences. Also, Mishkin et al. (2024) proposed directional smoothness, which preserves L-smoothness along specific directions.

Algorithms for Generalized-Smooth Optimization. Motivated by achieving comparable lower bounds presented in Arjevani et al. (2023) under standard assumptions, algorithms for solving generalized-smooth problems can be categorized into two main series. The first series focus on adaptive methods. In deterministic non-convex settings, Zhang et al. (2019; 2020) showed that Clipped GD can achieve a rate of $\mathcal{O}(\epsilon^{-2})$

under mild assumptions.Later, Chen et al. (2023) proposed β-GD, also achieving $\mathcal{O}(\epsilon^{-2})$ iteration complexity. Other methods, such as smoothed gradient clipping, gradient descent with Polyak step-sizes, triangles Method were explored by Gorbunov et al. (2024) under convex (L_0, L_1)-smooth conditions. In stochastic settings, when the approximation error of the stochastic gradient estimator is bounded, Zhang et al. (2019; 2020) proved clipped SGD achieves $\mathcal{O}(\epsilon^{-4})$ sample complexity. Inspired superior performance of Adagrad Duchi et al. (2011), Wang et al. (2023); Faw et al. (2023); Hong & Lin (2024) further studied AdaGrad under generalized smooth and relaxed variance assumption with different learning rate schemes. They all attains $\tilde{\mathcal{O}}(1/\sqrt{T})$ convergence rate under mild conditions.Xie et al. (2024a) studied trust-region methods convergence under generalized-smoothness. The second series focus on SGD methods with constant learning rate. Reisizadeh et al. (2023); Li et al. (2024) proved that SGD converges with sample complexity $\mathcal{O}(\epsilon^{-4})$ under generalized-smoothness. To ensure convergence, Reisizadeh et al. (2023) adopted a large batch size of $\mathcal{O}(\epsilon^{-2})$, while Li et al. (2024) relaxed this requirement but introduces additional variables of size $\mathcal{O}(\epsilon^{-1})$. Additionally, various acceleration methods have been explored under the generalized-smoothness condition. Zhang et al. (2020) proposed a general clipping framework with momentum updates; Jin et al. (2021) studied normalized SGD with momentum Cutkosky & Mehta (2020) under parameter-dependent achieves $\mathcal{O}(\epsilon^{-4})$ sample complexity; Hübler et al. (2024) studied normalized SGD with momentum Cutkosky & Mehta (2020) associated with parameter-agnostic learning rates, which establishes $\tilde{\mathcal{O}}(\epsilon^{-4})$ convergence rate and corresponding lower bound. By adjusting batch size, Chen et al. (2023); Reisizadeh et al. (2023) demonstrated that the SPIDER algorithm (Fang et al., 2018) can reach the optimal $\mathcal{O}(\epsilon^{-3})$ sample complexity. Furthermore, Zhang et al. (2024b); Wang et al. (2024a;b); Li et al. (2023) explored the convergence of RMSprop (Hinton et al., 2012) and Adam (Kingma, 2014) under generalized-smoothness. Jiang et al. (2024) studied variance-reduced sign-SGD convergence under generalized-smoothness.

Machine Learning Applications. Generalized smoothness has been studied under various machine learning framework. Levy et al. (2020); Jin et al. (2021) studied the dual formulation of regularized DRO problems, where the loss function objective satisfies generalized smoothness. Chayti & Jaggi (2024) identified their meta-learning objective's smoothness constant increases with the norm of the meta-gradient. Gong et al. (2024b); Hao et al. (2024); Gong et al. (2024a); Liu et al. (2022b) explored algorithms for bi-level optimization and federated learning within the context of generalized smoothness. Zhang et al. (2024a) developed algorithms for multi-task learning problem where the objective is generalized smooth. Xie et al. (2024b) studied online mirror descent when the objective is generalized smooth. Xian et al. (2024) studied min-max optimization algorithms' convergence behavior under generalized smooth condition.

2 Deterministic Generalized-Smooth Nonconvex Optimization

We first introduce generalized-smooth optimization problems. Then, we review the classic normalized gradient descent algorithm and study its complexity in the generalized-smooth and gradient-dominant setting.

We are interested in the following nonconvex optimization problem.

$$\min_{w \in \mathbf{R}^d} f(w), \tag{1}$$

where $f : \mathbf{R}^d \to \mathbf{R}$ denotes a nonconvex and differentiable function, and w corresponds to the model parameters. We assume that function f satisfies the following generalized-smooth condition introduced in (Jin et al., 2021; Chen et al., 2023)[1].

Assumption 1 (Generalized-smooth) *The objective function f satisfies the following conditions.*

1. *f is differentiable and bounded below, i.e., $f^* := \inf_{x \in \mathbf{R}^d} f(x) > -\infty$;*

[1]Jin et al. (2021) considered the special case $\alpha = 1$, and Chen et al. (2023) defined a symmetric version of equation 2.

2. *There exists constants $L_0, L_1 > 0$ and $\alpha \in [0,1]$ such that for any w, $w' \in \mathbf{R}^d$, it holds that*

$$\|\nabla f(w) - \nabla f(w')\| \leq \big(L_0 + L_1 \|\nabla f(w')\|^{\alpha}\big)\|w - w'\|. \tag{2}$$

The generalized-smooth condition in Assumption 1 is a generalization of the standard smooth condition, which corresponds to the special case of $L_1 = 0$. To elaborate, it allows the smoothness parameter to scale with the gradient norm polynomially, and therefore is able to model functions with highly irregular nonconvex geometry. In (Chen et al., 2023; Zhang et al., 2019), it has been shown that many complex machine learning problems belong to this function class with different parameter α, including distributionally-robust optimization ($\alpha = 1$), deep neural networks and phase retrieval ($\alpha = \frac{2}{3}$), etc. Following a standard proof, it is easy to show that generalized-smooth functions satisfy the following descent lemma.

Lemma 1 *Under Assumption 1, function f satisfies, for any $w, w' \in \mathbf{R}^d$,*

$$f(w) \leq f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{1}{2}(L_0 + L_1 \|\nabla f(w')\|^{\alpha})\|w - w'\|^2. \tag{3}$$

The main challenge of generalized-smooth optimization is to control the polynomial gradient norm term $\|\nabla f(w')\|^{\alpha}$ involved in the smoothness parameter. This key observation has motivated the existing studies to develop normalized gradient methods for solving generalized-smooth problems.

Chen et al. (2023) proposed a specialized normalized gradient descent (NGD) algorithm for generalized-smooth nonconvex optimization. The algorithm normalizes the gradient by its norm polynomially, i.e.,

$$\text{(NGD)} \quad w_{t+1} = w_t - \gamma \frac{\nabla f(w_t)}{\|\nabla f(w_t)\|^{\beta}}, \tag{4}$$

where $\gamma > 0$ denotes the learning rate and β is a scaling parameter that controls the normalization scale of the gradient norm. Intuitively, when the gradient norm is large, a smaller β would make the normalized gradient update more aggressive. Chen et al. (2023) has empirically demonstrated the effectiveness of equation 4 than gradient descent in deterministic settings and proved that, when choosing a proper γ and setting $\beta \in [\alpha, 1]$, NGD can find an ϵ-stationary point within $\mathcal{O}(\epsilon^{-2})$ number of iterations.

Moreover, From Liu et al. (2022a); Scaman et al. (2022), it has been observed that generalization of Polyak-Łojasiewicz (PŁ) condition, such as Polyak-Łojasiewicz*(PŁ)*, Kurdyka-Łojasiewicz*(KŁ*), and Separable-Łojasiewicz*(SŁ*) hold in the landscape under over-parametrized neural networks of several state-of-art losses, such as mean-squared loss and cross-entropy loss. Based on scalability of PŁ condition that can be extended to hold in deep learning, and the linear convergence rate studied in Karimi et al. (2016b). In this work, we study the convergence rate of NGD for solving generalized-smooth problems under the following generalized Polyak-Łojasiewicz (PŁ) condition (Karimi et al., 2016a).

Assumption 2 (Generalized Polyak-Łojasiewicz Condition) *There exists constants $\mu > 0$ and $\rho > 0$ such that $f(\cdot)$ satisfies, for all $w \in \mathbf{R}^d$,*

$$\frac{1}{2\mu}\|\nabla f(w)\|^{\rho} \geq f(w) - f^*. \tag{5}$$

Assumption 2 generalizes the standard PŁ condition (corresponds to $\rho = 2$) via flexible choice of the parameter ρ. In particular, some generalized-smooth functions satisfy the above generalized PŁ condition with different parameters. For example, the sigmoid-like function $f(w) = \frac{1}{2}w^2(\exp(w^2) - 1) + \frac{1}{2}w^2$ satisfies $\rho = 1, \mu = 0.1$, and the polynomial function $f(w) = w + \frac{1}{2}w^2 + w^4$ satisfies $\rho = 3, \mu = 0.1$.

We obtain the following convergence rate result of NGD, where we denote $\Delta_t := f(w_t) - f^*$.

Theorem 1 (Convergence of NGD) *Let Assumptions 1 and 2 hold. Choose learning rate* $\gamma = \frac{(2\mu\epsilon)^{\beta/\rho}}{8(L_0+L_1)+1}$ *where* ϵ *denotes the target accuracy, and set* $\alpha \leq \beta \leq 1$*. Then, the following statements hold.*

- *If* $0 < \rho < 2-\beta$*, then we have*

$$\Delta_t = \mathcal{O}\Big(\big(\frac{\rho}{(2-\beta-\rho)\gamma t}\big)^{\frac{\rho}{2-\rho-\beta}}\Big). \tag{6}$$

  Furthermore, in order to achieve $\Delta_t \leq \epsilon$*, the total number of iteration satisfies* $T = \mathcal{O}\big((\frac{1}{\epsilon})^{\frac{\beta}{\rho}}\big)$ *if* $2-2\beta < \rho < 2-\beta$*, and* $T = \mathcal{O}\big((\frac{1}{\epsilon})^{\frac{2-\rho-\beta}{\rho}}\big)$ *if* $0 < \rho \leq 2-2\beta$*.*

- *If* $\rho = 2-\beta$ *and choose* ϵ *such that* $\gamma < \frac{2}{\mu}$*, then we have*

$$\Delta_t = \mathcal{O}\Big(\big(1-\frac{\gamma\mu}{2}\big)^t\Big). \tag{7}$$

  In order to achieve $\Delta_t \leq \epsilon$*, the total number of iteration satisfies* $T = \mathcal{O}\big((\frac{1}{\epsilon})^{\frac{\beta}{\rho}}\log\frac{1}{\epsilon}\big)$*.*

- *If* $\rho > 2-\beta$*, then there exists* $T_0 \in \mathbf{N}$ *such that for all* $t \geq T_0$*, we have*

$$\Delta_t = \mathcal{O}\Big(\Big(\frac{\Delta_{T_0}}{\gamma^{\frac{\rho}{\rho+\beta-2}}}\Big)^{\frac{\rho}{2-\beta}^{t-T_0}}\Big). \tag{8}$$

  In order to achieve $\Delta_t \leq \epsilon$*, the total number of iterations after* T_0 *satisfies* $T = \Omega\big(\log((\frac{1}{\epsilon})^{\frac{\beta}{\rho+\beta-2}})\big)$*.*

Theorem 1 indicates that the convergence behavior of NGD depends on the parameter ρ in the generalized PŁ condition. When $\rho \leq 2-\beta$, the algorithm achieves slow sub-linear convergence rate. When $\rho > 2-\beta$, the algorithm achieves local linear convergence rate. These results match the intuition behind the generalized PŁ condition. Namely, a large ρ indicates that the gradient norm vanishes slowly when the function value gap approaches zero, corresponding to sharp geometry that leads to fast local convergence.

3 Stochastic Generalized-Smooth Nonconvex Optimization

In this section, we study the following stochastic generalized-smooth optimization problem, where f_ξ corresponds to the loss function associated with data sample ξ, and the expected loss function $F(\cdot)$ satisfies the generalized-smooth condition in Assumption 1.

$$\min_{w\in\mathbf{R}^d} F(w) := \mathbb{E}_{\xi\sim\mathbb{P}}\big[f_\xi(w)\big]. \tag{9}$$

3.1 Normalized SGD and Its Limitations

To solve the stochastic generalized-smooth problem in equation 9, one straightforward approach is to replace the full batch gradient in the NGD update in equation 4 with the stochastic gradient $\nabla f_\xi(w_t)$, resulting in the following normalized SGD (NSGD) algorithm.

$$\text{(NSGD)} \quad w_{t+1} = w_t - \gamma\frac{\nabla f_{\xi_t}(w_t)}{\|\nabla f_{\xi_t}(w_t)\|^\beta}. \tag{10}$$

NSGD-type algorithms have attracted a lot of attention recently for solving stochastic generalized-smooth problems (Zhang et al., 2019; 2020; Liu et al., 2022b). In particular, it has been proven in these works that NSGD with proper gradient clipping can achieve a sample complexity of $\mathcal{O}(\epsilon^{-4})$, which matches that of the standard SGD algorithm for solving stochastic smooth problems (Ghadimi & Lan, 2013). However, NSGD-type update has the following limitations.

1. *Biased gradient estimator:* The normalized stochastic gradient used in equation 10 is biased, i.e., $\mathbb{E}[\frac{\nabla f_{\xi_t}(w_t)}{\|\nabla f_{\xi_t}(w_t)\|^\beta}] \neq \frac{\nabla F(w_t)}{\|\nabla F(w_t)\|^\beta}$. This is due to the dependence between $\nabla f_{\xi_t}(w_t)$ and $\|\nabla f_{\xi_t}(w_t)\|^\beta$. In particular, the bias can be huge if the stochastic gradients are diverse, as illustrated in Figure 1.
2. *Strong assumption:* To control the estimation bias and establish theoretical convergence guarantee for NSGD-type algorithms in generalized-smooth nonconvex optimization, the existing studies need to adopt strong assumption. For example, Zhang et al. (2019; 2020) and Liu et al. (2022b) assume that the stochastic approximation error $\|\nabla f_\xi(w) - \nabla F(w)\|$ is bounded by a constant almost surely. In real applications, this constant can be a large numerical number if certain sample ξ is an outlier.

3.2 Independently-Normalized SGD

To overcome the aforementioned limitations, we propose the independently-normalized stochastic gradient (I-NSG) estimator

$$\text{(I-NSG estimator)} \quad \frac{\nabla f_\xi(w)}{\|\nabla f_{\xi'}(w)\|^\beta}, \tag{11}$$

where ξ and ξ' are samples draw *independently* from the underlying data distribution. Intuitively, the independence between ξ and ξ' decorrelates the denominator from the numerator, making it an unbiased stochastic gradient estimator (up to a scaling factor). Specifically, we formally have that

$$\mathbb{E}_{\xi,\xi'}\left[\frac{\nabla f_\xi(w)}{\|\nabla f_{\xi'}(w)\|^\beta}\right] = \mathbb{E}_{\xi'}\left[\frac{\mathbb{E}_\xi\big[\nabla f_\xi(w)\big]}{\|\nabla f_{\xi'}(w)\|^\beta}\right] \propto \nabla F(w). \tag{12}$$

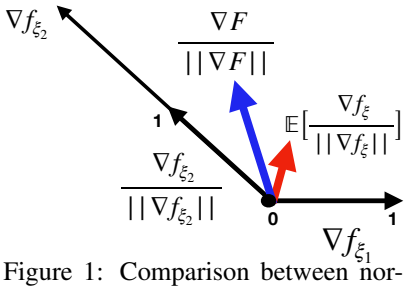

Figure 1: Comparison between normalized full gradient (blue) and expected normalized stochastic gradient (red). Here, ξ_1 and ξ_2 are sampled uniformly at random.

Moreover, as we show later under mild assumptions, the scaling factor $\mathbb{E}\big[\|\nabla f_{\xi'}(w)\|^{-\beta}\big]$ can be roughly bounded by the full gradient norm and hence resembling the full-batch NGD update. Based on this idea, we formally propose the following independently-normalized SGD (I-NSGD) algorithm, where $\nabla f_{\xi_B}(w_t)$ corresponds to the mini-batch stochastic gradient associated with a batch of samples B, and B' denotes another independent batch.

$$\text{(I-NSGD):} \;\; w_{t+1} = w_t - \gamma \frac{\nabla f_{\xi_B}(w_t)}{h_t^\beta}, \;\; \text{where} \;\; h_t = \max\Big\{1, (4L_1\gamma)^{\frac{1}{\beta}}\Big(2\big\|\nabla f_{\xi_{B'}}(w_t)\big\| + \delta\Big)\Big\}. \tag{13}$$

The above I-NSGD algorithm adopts a clipping strategy for the normalization term h_t. This is to impose a constant lower bound on h_t, which helps develop the theoretical convergence analysis and avoid numerical instability in practice. We note that I-NSGD requires querying two batches of samples in every iteration. However, as we show in the experiments later, the batch size $|B'|$ can be chosen far smaller than $|B|$.

3.3 Convergence Analysis of I-NSGD

We adopt the following standard assumptions on the stochastic gradient.

Assumption 3 (Unbiased stochastic gradient) *The stochastic gradient $\nabla f_\xi(w)$ is unbiased, i.e., $\mathbb{E}_{\xi\sim\mathbb{P}}\big[\nabla f_\xi(w)\big] = \nabla F(w)$ for all $w \in \mathbf{R}^d$.*

Assumption 4 (Approximation error) *There exists $\tau_1, \tau_2 > 0$ such that for any $w \in \mathbf{R}^d$, one has*

$$\big\|\nabla f_\xi(w) - \nabla F(w)\big\| \leq \tau_1\big\|\nabla F(w)\big\| + \tau_2 \;\; a.s. \;\; \forall \xi \sim \mathbb{P}. \tag{14}$$

We note that the above Assumption 4 is much weaker than the bounded approximation error assumption (i.e., $\tau_1 = 0$) adopted in (Zhang et al., 2019; 2020; Liu et al., 2022b). Specifically, it allows the approximation error to scale with the full gradient norm and only assumes bounded error at the stationary points. With these assumptions, we can lower bound the stochastic gradient norm with the full gradient norm as follows.

Lemma 2 *Let Assumptions 3 and 4 hold. Consider the mini-batch stochastic gradient ∇f_{ξ_B} with batch size $B = 16\tau_1^2$, then for all $w \in \mathbf{R}^d$ we have*

$$\left\|\nabla f_{\xi_B}(w)\right\| \geq \frac{1}{2}\left\|\nabla F(w)\right\| - \frac{\tau_2}{2\tau_1}. \tag{15}$$

Lemma 2 shows that with a constant-level batch size, the stochastic gradient norm can be lower bounded the full gradient norm up to a constant. This result is very useful in our convergence analysis to effectively bound the mini-batch stochastic gradient norm used in the normalized stochastic gradient update.

We obtain the following convergence result of I-NSGD.

Theorem 2 (Convergence of I-NSGD) *Let Assumptions 1, 3 and 4 hold. For the I-NSGD algorithm, choose learning rate $\gamma = \min\{\frac{1}{4L_0}, \frac{1}{4L_1}, \frac{1}{\sqrt{T}}, \frac{1}{8L_1(3\tau_2/\tau_1)^\beta}\}$, batch sizes $B = 2\tau_1^2$, $B' = 16\tau_1^2$ and $\delta = \frac{\tau_2}{\tau_1}$. Denote $\Lambda := F(w_0) - F^* + \frac{1}{2}(L_0 + L_1)(1 + \tau_2^2/\tau_1^2)^2$. Then, with probability at least $\frac{1}{2}$, I-NSGD produces a sequence satisfying $\min_{t\leq T}||\nabla F(w_t)|| \leq \epsilon$ if the total number of iteration T satisfies*

$$T \geq \Lambda \max\Big\{\frac{256\Lambda}{\epsilon^4}, \frac{640L_1}{\epsilon^{2-\beta}}, \frac{64(L_0 + L_1) + 128L_1(3\tau_2/\tau_1)^\beta}{\epsilon^2}\Big\}. \tag{16}$$

The choices of $B, B' = \mathcal{O}(\tau_1^2)$ are mainly to simplify the symbolic operation during the proof. By deploying normalizing during data pre-processing, the value of τ_1 can be approximately controlled as $\mathcal{O}(1)$ in practice. Thus, Theorem 2 indicates that I-NSGD achieves a sample complexity in the order of $\mathcal{O}(\epsilon^{-4})$ with constant-level batch sizes in generalized-smooth optimization. Compared to the existing studies on normalized/clipped SGD, this convergence result neither requires using extremely large batch sizes nor depending on the bounded error assumption. Through numerical experiments in Section 4 later, we show that it suffices to query a small number of independent samples for I-NSGD in practice.

Proof outline and novelty: The independent sampling strategy adopted by I-NSGD naturally decouples stochastic gradient from gradient norm normalization, making it easier to achieve the desired optimization progress in generalized-smooth optimization under relaxed conditions. By the descent lemma, we have that

$$\begin{aligned}\mathbb{E}_{\xi_B}\left[F(w_{t+1}) - F(w_t)\right] &\overset{(i)}{\leq} \frac{-\gamma\left\|\nabla F(w_t)\right\|^2}{h_t^\beta} + \frac{\gamma^2(L_0 + L_1\left\|\nabla F(w_t)\right\|^\alpha)}{2h_t^{2\beta}}\mathbb{E}_{\xi_B}\left[\left\|\nabla f_{\xi_B}(w_t)\right\|^2\right] \\ &\overset{(ii)}{\leq} \Big(\frac{\gamma}{h_t^\beta}\big(-1 + \gamma\frac{L_0 + L_1\left\|\nabla F(w_t)\right\|^\alpha}{h_t^\beta}\big)\Big)\left\|\nabla F(w_t)\right\|^2 + \frac{1}{2}\gamma^2\frac{L_0 + L_1\left\|\nabla F(w_t)\right\|^\alpha}{h_t^{2\beta}}\frac{\tau_2^2}{\tau_1^2},\end{aligned} \tag{17}$$

where the expectation (conditioned on w_t) in (i) is taken over ξ_B only, and note that h_t involves the independent mini-batch samples $\xi_{B'}$; (ii) leverages Assumption 4 to bound the second moment term $\mathbb{E}_{\xi_B}[\|\nabla f_{\xi_B}(w_t)\|^2]$ by $2\|\nabla F(w_t)\|^2 + \tau_2^2/\tau_1^2$. Then, for the first term in equation 17, we leverage the clipping structure of h_t to bound the coefficient $\gamma(L_0 + L_1||\nabla F(w)||^\alpha)/h_t^\beta$ by $\frac{1}{2}$. For the second term in equation 17, we again leverage the clipping structure of h_t and consider two complementary cases: when $||\nabla F(w_t)|| \leq \sqrt{1 + \tau_2^2/\tau_1^2}$, this term can be upper bounded by $\frac{1}{2}\gamma^2(L_0 + L_1)(1 + \tau_2^2/\tau_1^2)$; when $||\nabla F(w_t)|| > \sqrt{1 + \tau_2^2/\tau_1^2}$, this term can be upper bounded by $\frac{\gamma}{4h_t^\beta}||\nabla F(w_t)||^2$. Summing them up gives the desired bound. We refer to Lemma 6 in the appendix for more details. Substituting these bounds into equation 17 and rearranging the terms yields that

$$\frac{\gamma}{4h_t^\beta}\left\|\nabla F(w_t)\right\|^2 \leq \mathbb{E}_{\xi_B}\left[F(w_t) - F(w_{t+1})\right] + \frac{1}{2}(L_0 + L_1)\gamma^2(1 + \frac{\tau_2^2}{\tau_1^2})^2.$$

Furthermore, by leveraging the clipping structure of h_t^β and Assumption 4, the left hand side can be lower bounded as $\frac{\gamma\|\nabla F(w_t)\|^2}{h_t^\beta} \geq \min\{\gamma\|\nabla F(w_t)\|^2, \frac{\|\nabla F(w_t)\|^{2-\beta}}{20L_1}\}$. Finally, telescoping above inequalities over t and taking expectation leads to the desired bound in equation 16.

As a comparison, in the prior work on clipped SGD (Zhang et al., 2019; 2020), their stochastic gradient and normalization term h_t depend on the same mini-batch of samples, and therefore cannot be treated separately in the analysis. For example, their analysis proposed the following decomposition.

$$\mathbb{E}_{\xi_B}\frac{\|\nabla f_{\xi_B}(w_t)\|^2}{h_t^{2\beta}} = \mathbb{E}_{\xi_B}\frac{\|\nabla F(w_t)\|^2 + \|\nabla f_{\xi_B}(w_t) - \nabla F(w_t)\|^2 + 2\langle \nabla F(w_t), \nabla f_{\xi_B}(w_t) - \nabla F(w_t)\rangle}{h_t^{2\beta}}.$$

Hence their analysis need to assume a constant upper bound for the approximation error $\|\nabla f_{\xi_B}(w_t) - \nabla F(w_t)\|$ in order to obtain a comparable bound to ours.

We also analyze I-NSGD under the generalized PŁ condition by establishing a recursion similar to that proved in Theorem 1. Due to page limitation, we refer to Appendix G for more details.

4 EXPERIMENTS

We conduct numerical experiments to compare I-NSGD with other state-of-the-art stochastic algorithms, including the standard SGD (Ghadimi & Lan, 2013), normalized SGD, Clipped SGD (Zhang et al., 2019). The problems we consider are nonconvex phase retrieval and nonconvex distributionally-robust optimization.

4.1 NONCONVEX PHASE RETRIEVAL

The phase retrieval problem arises in optics, signal processing, and quantum mechanics (Drenth, 2007). The goal is to recover a signal from measurements where only the intensity is known, and the phase information is missing or difficult to measure. Specifically, denote the underlying object as $x \in \mathbf{R}^d$. Suppose we take m intensity measurements $y_r = |a_r^T x|^2 + n_r$ for $r = 1 \cdots m$, where a_r denotes the measurement vector and n_r is the additive noise. We aim to reconstruct x by solving the following regression problem.

$$\min_{z\in\mathbf{R}^d} f(z) = \frac{1}{2m}\sum_{r=1}^{m}\left(y_r - \left|a_r^T z\right|^2\right)^2. \tag{18}$$

Such nonconvex function is generalized-smooth with parameter $\alpha = \frac{2}{3}$ (Chen et al., 2023). In this experiment, we generate the initialization $z_0 \sim \mathcal{N}(1, 6)$ and the underlying signal $x \sim \mathcal{N}(0, 0.5)$ with dimension $d = 100$. We take $m = 3$k measurements with $a_r \sim \mathcal{N}(0, 0.5)$ and $n_r \sim \mathcal{N}(0, 4^2)$.

We implement all the stochastic algorithms with batch size $|B| = 64$, and we choose a small independent batch size $|B'| = 8$ for I-NSGD. We use fine-tuned learning rates for all algorithms, i.e., $\gamma = 5\text{e}{-5}$ for SGD, 0.25 for both normalized SGD and Clipped SGD, and 0.5 for I-NSGD. We set the maximal gradient clipping constant 45 and $\delta = 15$ for both Clipped SGD and I-NSGD. Moreover, we set the normalization parameter of I-NSGD as $\beta = \frac{2}{3}$, which matches the generalized-smoothness parameter α of phase retrieval.

Figure 2 (left) shows the comparison of objective value versus sample complexity. It can be seen that our I-NSGD consistently converges faster than other algorithms. This indicates that, the independently-normalized and clipped updates of I-NSGD are more adapted to the underlying generalized-smooth nonconvex geometry. In Figure 2 (middle), we test the performance of I-NSGD under different choices of the normalization parameter $\beta = 1, \frac{4}{5}, \frac{2}{3}, \frac{7}{10}, \frac{13}{20}$. It can be seen that I-NSGD converges the fastest as β matches the theoretically-suggested value $\frac{2}{3}$, demonstrating the importance of imposing a proper level of gradient normalization in generalized-smooth optimization. In Figure 2 (right), we further explore the effect of the batch size for I-NSGD's independent batch samples B'. Specifically, we test batch sizes $|B'| = 4, 8, 16, 32, 64$, while keeping all other hyper-parameters unchanged. The plot shows that I-NSGD can achieve both fast and stable convergence when choosing a very small batch size ($|B'| = 4$ or 8) for the independent batch samples.

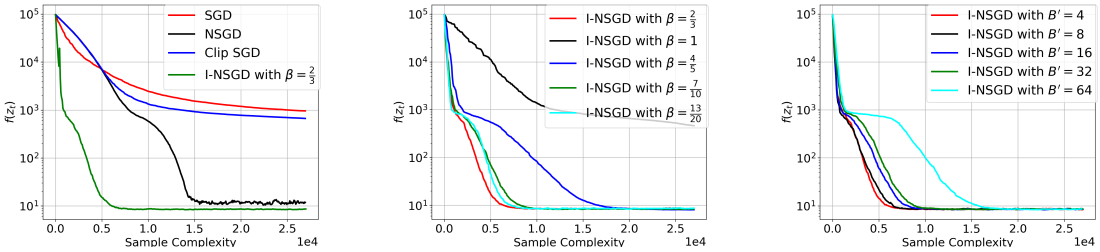

Figure 2: Left: Comparison of I-NSGD and stochastic algorithms. Middle: Performance of I-NSGD with different normalization parameters. Right: Performance of I-NSGD with different independent batch sizes.

4.2 Distributionally-Robust Optimization

Distributionally-robust optimization (DRO) is a popular approach to enhance robustness against data distribution shift. We consider the regularized DRO problem $\min_{w\in\mathcal{W}} f(w) = \sup_{\mathbb{Q}} \left\{\mathbb{E}_{\xi\sim\mathbb{Q}}[\ell_\xi(w)] - \lambda\Psi(\mathbb{P};\mathbb{Q})\right\}$, where $\mathbb{Q}$, $\mathbb{P}$ represents the underlying distribution and the nominal distribution respectively. λ denotes a regularization hyper-parameter and Ψ denotes a divergence metric. Under mild technical assumptions, Jin et al. (2021) showed that such a problem has the following equivalent dual formulation

$$\min_{w\in\mathcal{W}} L(w,\eta) = \lambda\mathbb{E}_{\xi\sim P}\Psi^*\left(\frac{\ell_\xi(w)-\eta}{\lambda}\right) + \eta, \tag{19}$$

where Ψ^* denotes the conjugate function of Ψ and η is a dual variable. In particular, such dual objective function is generalized-smooth with parameter $\alpha = 1$ (Jin et al., 2021; Chen et al., 2023). In this experiment, we use the life expectancy data (Arshi, 2017). We set $\lambda = 0.01$ and select $\Psi^*(t) = \frac{1}{4}(t+2)_+^2 - 1$, i.e., the conjugate of χ^2-divergence. We adopt the regularized loss $\ell_\xi(\mathbf{w}) = \frac{1}{2}(y_\xi - \mathbf{x}_\xi^\top\mathbf{w})^2 + 0.1\sum_{j=1}^{34}\ln(1+|w^{(j)}|)$.

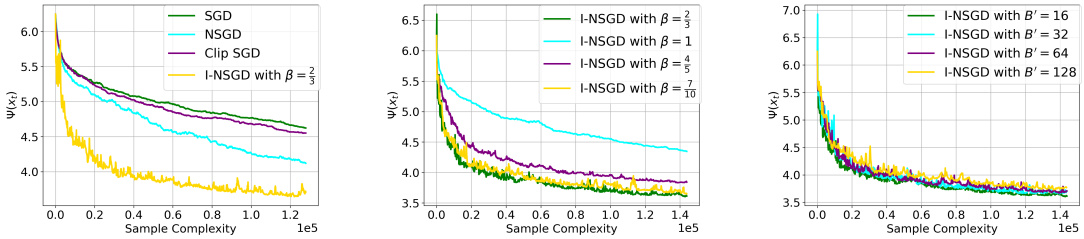

Figure 3: Left: Comparison of I-NSGD and stochastic algorithms. Middle: Performance of I-NSGD with different normalization parameters. Right: Performance of I-NSGD with different independent batch sizes.

We implement all the aforementioned stochastic algorithms with batch size $|B| = 128$, and we choose $|B'| = 16$ for I-NSGD. We use fine-tuned learning rates for all algorithms, i.e., $\gamma = 4\text{e}{-5}$ for SGD, $5\text{e}{-2}$ for normalized SGD, 0.18 for Clipped SGD and 0.28 for I-NSGD. We set the maximal gradient clipping constant 60 and $\delta = 45$ for both Clipped SGD and I-NSGD.

Figure 3 (left) shows the comparison of objective value versus sample complexity. It can be seen that $\frac{2}{3}$-I-NSGD consistently converges faster than other methods. This indicates independent normalization and clipped updates is also more adapted to function geometry of equation 19. In Figure 3 (middle), we test the performance of I-NSGD under different choices of the normalization parameter $\beta = 1, \frac{4}{5}, \frac{7}{10}, \frac{2}{3}$. It can be seen that $\beta = \frac{2}{3}$ outperforms all other choices in terms of both convergence speed and stability. In Figure 3 (right), we explore the effect of the batch size for I-NSGD's independent batch samples B'. We

test batch sizes $|B'| = 16, 32, 64, 128$ and keeping all other hyper-parameters unchanged. The plot shows the loss function as a function of sample complexity. We found the batch size of $|B'|$ has little effect the convergence speed. $|B'| = 16$ is sufficient to guarantee fast and stable convergence. This indicates I-NSGD doesn't require a large batch size to ensure convergence, and is more suitable for large-scale problems. To further demonstrate the effectiveness of I-NSGD, we compare our algorithm with additional baseline methods, normalized SGD with momentum Cutkosky & Mehta (2020) and SPIDER Fang et al. (2018). We also conduct ablation study to unify the normalization parameter β for all normalization method. Experiment results show the effectiveness of our proposed I-NSGD framework, which combines independent sampling with clipping updates, and normalization parameter β. We refer readers to check Section I in appendix for more details about experiments settings and corresponding results.

5 Conclusion

In this work, we study convergence of normalized gradient descent under generalized smooth and generalized PŁ condition. We propose independent normalized stochastic gradient descent for stochastic setting, achieving same sample complexity under relaxed assumptions. Our results extend the existing boundary of first-order nonconvex optimization and may inspire new developments in this direction. In the future, it is interesting to explore if the popular acceleration method such as stochastic momentum and variance reduction can be combined with independent sampling and normalization to improve the sample complexity.

References

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.

Kumar Ajay Arshi. Life expectancy data. Kaggle, 2017. URL `https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?resource=download`.

El Mahdi Chayti and Martin Jaggi. A new first-order meta-learning algorithm with convergence guarantees. *arXiv preprint arXiv:2409.03682*, 2024.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2260–2268. PMLR, 13–18 Jul 2020.

Jan Drenth. *Principles of protein X-ray crystallography*. Springer Science & Business Media, 2007.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In *Advances in neural information processing systems*, 2018.

Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive sgd. In *The Thirty Sixth Annual Conference on Learning Theory*, 2023.

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

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

Xiaochuan Gong, Jie Hao, and Mingrui Liu. A nearly optimal single loop algorithm for stochastic bilevel optimization under unbounded smoothness. In *Proceedings of the 41st International Conference on Machine Learning*, 2024b.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. *arXiv preprint arXiv:2409.14989*, 2024.

Jie Hao, Xiaochuan Gong, and Mingrui Liu. Bilevel optimization under unbounded smoothness: A new algorithm and convergence analysis. In *12th International Conference on Learning Representations*, 2024.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

Yusu Hong and Junhong Lin. Revisiting convergence of adagrad with relaxed assumptions. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 2024.

Wei Jiang, Sifan Yang, Wenhao Yang, and Lijun Zhang. Efficient sign-based optimization: Accelerating convergence via variance reduction. *arXiv preprint arXiv:2406.00489*, 2024.

Jikai Jin, Bohang Zhang, Haiyang Wang, and Liwei Wang. Non-convex distributionally robust optimization: Non-asymptotic analysis. In *Advances in Neural Information Processing Systems*, 2021.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pp. 795–811, 2016a.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pp. 795–811. Springer, 2016b.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems*, 2020.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. In *Advances in Neural Information Processing Systems*, 2023.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *Advances in Neural Information Processing Systems*, 2024.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022a.

Mingrui Liu, Zhenxun Zhuang, Yunwen Lei, and Chunyang Liao. A communication-efficient distributed gradient clipping algorithm for training deep neural networks. In *Advances in Neural Information Processing Systems*, 2022b.

Pengfei Liu, Xiang Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys (CSUR)*, 55(9):1–35, 2023.

Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert M. Gower. Directional smoothness and gradient methods: Convergence and adaptivity. *arXiv preprint arXiv:2403.04081*, 2024.

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

Amirhossein Reisizadeh, Haochuan Li, Subhro Das, and Ali Jadbabaie. Variance-reduced clipping for non-convex optimization. *arXiv preprint arXiv:2303.00883*, 2023.

Kevin Scaman, Cedric Malherbe, and Ludovic Dos Santos. Convergence rates of non-convex stochastic gradient descent under a generic lojasiewicz condition and local smoothness. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, 2023.

Bohan Wang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, and Wei Chen. On the convergence of adam under non-uniform smoothness: Separability from sgdm and beyond, 2024a.

Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, Tie-Yan Liu, Zhi-Quan Luo, and Wei Chen. Provable adaptivity of adam under non-uniform smoothness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024b.

Wenhan Xian, Ziyi Chen, and Heng Huang. Delving into the convergence of generalized smooth minimization optimization. In *12th International Conference on Learning Representations*, 2024.

Chenghan Xie, Chenxi Li, Chuwen Zhang, Qi Deng, Dongdong Ge, and Yinyu Ye. Trust region methods for nonconvex stochastic optimization beyond lipschitz smoothness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024a.

Yan-Feng Xie, Peng Zhao, and Zhi-Hua Zhou. Gradient-variation online learning under generalized smoothness. *arXiv preprint arXiv:2408.09074*, 2024b.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. In *Advances in Neural Information Processing Systems*, 2020.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2019.

Qi Zhang, Peiyao Xiao, Kaiyi Ji, and Shaofeng Zou. On the convergence of multi-objective optimization under generalized smoothness. *arXiv preprint arXiv:2405.19440*, 2024a.

Qi Zhang, Yi Zhou, and Shaofeng Zou. Convergence guarantees for rmsprop and adam in generalized-smooth non-convex optimization with affine noise variance. *arXiv preprint arXiv:2404.01436*, 2024b.