Adaptive Gradient Normalization and Independent Sampling for (Stochastic) Generalized-Smooth Optimization

Anonymous authors
Paper under double-blind review

Abstract

Recent studies have shown that many nonconvex machine learning problems meet generalized-smooth condition that extends beyond traditional smooth nonconvex optimization. However, the existing algorithms cannot fully adapt to generalized-smooth nonconvex geometry and encounter significant technical limitations on convergence analysis. In this work, we first justify the advantage of using adaptive gradient normalization. We analyze the overall effects of adaptive normalization and function geometry on convergence rate. Our results provide a comprehensive understanding of the interplay between adaptive gradient normalization and function geometry. For stochastic generalized-smooth nonconvex optimization, we propose **I**ndependent-**A**daptively **N**ormalized **S**tochastic **G**radient **D**escent algorithm, which leverages adaptive gradient normalization, independent sampling, and gradient clipping to achieve an $\mathcal{O}(\epsilon^{-4})$ sample complexity under relaxed assumptions. Experiments on large-scale nonconvex generalized-smooth problems 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 L-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, various works have demonstrated that *deterministic* first-order algorithms such as gradient descent, normalized gradient descent and clipped gradient descent can achieve $\mathcal{O}(\epsilon^{-2})$ iteration complexity under mild assumptions (Li et al., 2024; Zhang et al., 2019; Chen et al., 2023; Gorbunov et al., 2024; Vankov et al., 2024b; Reisizadeh et al., 2023). These complexity results match the lower bound obtained by classical first-order methods. In particular, Chen et al. (2023) empirically demonstrated that proper usage of adaptive gradient normalization can substantially accelerate convergence in practice. However, the formal theoretical justification and understanding of adaptive gradient normalization is still lacking for first-order algorithms in generalized-smooth optimization.

On the other hand, some other works studied first-order *stochastic* algorithms in generalized-smooth nonconvex optimization (Li et al., 2024; Zhang et al., 2019; 2020). Specifically, one line of work focused on the classic stochastic gradient descent (SGD) algorithm (Li et al., 2024). However, in the generalized-smooth setting, the convergence analysis of SGD either relies on adopting very large batch size or involves large constants (Arjevani et al., 2023). Moreover, the empirical performance of SGD is often unstable due to the ill-conditioned smoothness parameter when the gradient is large (Chen et al., 2023). Another line of work focused on clipped SGD, which leverages gradient normalization and clipping to handle the generalized-smooth geometry (Zhang et al., 2019; 2020; Reisizadeh et al., 2023). Although clipped SGD has demonstrated su-

perior performance in solving large-scale problems, its existing theoretical analysis has several limitations. First, in order to establish convergence guarantee, the existing studies highly rely on the strong assumption that the stochastic approximation error is bounded almost surely. Second, the existing designs of clipped SGD adopt the standard stochastic gradient normalization scheme, which is not fully adapted to the function geometry characterized by the generalized-smooth condition.

Having observed the algorithmic and theoretical limitations discussed above, we aim to advance the algorithm design and analysis for generalized-smooth optimization through investigating the following two fundamental and complementary questions.

- *Q1: In deterministic generalized-smooth optimization, how does adaptive gradient normalization affect the convergence rate of first-order algorithm, e.g., under Polyak-Łojasiewicz-type conditions ?*
- *Q2: In stochastic generalized-smooth optimization, can we develop a novel algorithm with convergence guarantee under relaxed noise assumptions?*

In this work, we provide comprehensive answers to both questions by developing new algorithms and convergence analysis in generalized-smooth nonconvex optimization. We summarize our contributions as follows.

1.1 Our Contributions

To understand the advantage of using adaptive gradient normalization, we first study the convergence rate of adaptive normalized gradient descent (ANGD) in deterministic generalized-smooth optimization under the generalized Polyak-Łojasiewicz (PŁ) condition over a board spectrum of gradient normalization parameters. Our results reveal the interplay among learning rate, gradient normalization parameter and function geometry parameter, and characterize their impact on the type of convergence rate. In particular, our results reveal the advantage of using adaptive gradient normalization and provide theoretical guidance on choosing proper gradient normalization parameter to improve convergence rate.

We further propose a novel Independent-Adaptively Normalized Stochastic Gradient Descent (IAN-SGD) algorithm tailored for stochastic generalized-smooth nonconvex optimization. Specifically, IAN-SGD leverages normalized gradient updates with independent sampling and gradient clipping to reduce bias and enhance algorithm stability. Consequently, we are able to establish convergence of IAN-SGD 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 compare the numerical performance of our IAN-SGD algorithm with other state-of-the-art stochastic algorithms in applications of nonconvex phase retrieval, nonconvex distributionally-robust optimization and training deep neural networks, all of which are generalized-smooth nonconvex problems. Our results demonstrate the efficiency of IAN-SGD in solving generalized-smooth nonconvex problems.

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 definition 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. 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 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 $\mathcal{O}(\epsilon^{-1})$. The second series focus on adaptive methods. In nonconvex deterministic 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 achieving $\mathcal{O}(\epsilon^{-2})$ iteration complexity. Vankov et al. (2024b) studies clip and normalized gradient descent under mild-conditions, where they retrieve best-known convergence rate under each separate cases, including strong convex, convex settings etc. Gorbunov et al. (2024) varies the learning rate to study smoothed gradient clipping, gradient descent with Polyak step-sizes, triangles Method under convex (L_0, L_1)-smooth conditions, where they also achieve standard convergence rate under convex case. 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. Wang et al. (2023); Faw et al. (2023); Hong & Lin (2024) studied AdaGrad Duchi et al. (2011b) 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. Several works also studied stochastic acceleration methods 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. There is a concurrent work (Vankov et al., 2024a) using independent sampling with clipped SGD framework to solve variation inequality problem (SVI). Based on this idea, they also propose stochastic Korpelevich method for clipped SGD. Under generalized-smooth condition, they proved almost-sure convergence in terms of distance to solution set tailored for solving stochastic SVI problems.

3 Generalized-Smooth nonconvex Optimization

We first introduce generalized-smooth optimization problems. Consider the following 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.

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

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

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

The generalized-smooth condition in Assumption 1 is a generalization of the standard smooth condition, which corresponds to the special case of $L_1 = 0$. It allows the smoothness parameter to scale with the gradient norm polynomially, and therefore Assumption 1 can model functions with highly irregular nonconvex geometry. Moreover, following the 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 $w, w' \in \mathbf{R}^d$,*

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

We note that there are several variants of generalized-smooth conditions proposed by the previous works (Zhang et al., 2019; Jin et al., 2021; Chen et al., 2023). Below, we briefly discuss the relationship between the generalized-smooth condition in Assumption 1 and these existing notions.

Remark 1 ((L_0, L_1)-generalized-smooth condition) *The (L_0, L_1)-generalized-smooth condition proposed in (Zhang et al., 2019) is a special case of equation 3 with $\alpha = 1$. The corresponding descent lemma is given as*

$$f(w) \leq f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{1}{2}(4L_0 + 5L_1 ||\nabla f(w')||)||w - w'||^2,$$

which is the same as Lemma 1 with $\alpha = 1$ up to differences in the constant coefficients.

Remark 2 (Symmetric generalized-smooth condition) *Chen et al. (2023) introduced a symmetric version of generalized-smooth condition, by replacing $||\nabla f(w')||^{\alpha}$ in equation 2 with $\max_{w_\theta} ||f(w_\theta)||^{\alpha}$, where $w_\theta = \theta w' + (1-\theta)w$. We notice that the asymmetric generalized-smooth condition adopted in our Assumption 1 slightly generalizes such a symmetric generalized-smooth condition, since it directly implies the following stronger symmetric generalized-smooth condition.*

$$\big\|\nabla f(w) - \nabla f(w')\big\| \leq \big(L_0 + L_1 \big(\frac{\big\|\nabla f(w)\big\|^{\alpha} + \big\|\nabla f(w')\big\|^{\alpha}}{2}\big)\big)\big\|w - w'\big\|. \tag{4}$$

The following propositions show that various nonconvex machine learning problems such as phase retrieval and distributionally robust optimization (DRO) satisfy our generalized-smooth condition.

Proposition 1 *The nonconvex phase retrieval objective function (see equation 19 in the appendix) satisfies equation 4 with $\alpha = \frac{2}{3}$.*

Proposition 2 *The distributionally robust optimization (DRO) objective function (see equation 20 in experiment section) satisfies equation 4 with $\alpha = 1$.*

Thus, throughout this work, we consider the generalized-smooth condition in Assumption 1, which generalizes the existing notions and simplifies arithmetic of the proof.

In the following sections, we first consider deterministic generalized-smooth optimization and study the impact of adaptive gradient normalization on the convergence rate of gradient methods. Then, we consider stochastic generalized-smooth optimization and propose a novel independent sampling scheme for improving the convergence guarantee of stochastic gradient methods.

4 Adaptive Gradient Normalization for Deterministic Generalized-Smooth Optimization

In deterministic generalized-smooth optimization, many previous works have empirically demonstrated the faster convergence of normalized gradient descent-type algorithms (e.g., clipped-GD) over the standard gradient descent algorithm in various machine learning applications (Jin et al., 2021; Zhang et al., 2019).

On the other hand, theoretically, these algorithms were only shown to achieve the same iteration complexity $\mathcal{O}(\epsilon^{-2})$ as the gradient descent algorithm in generalized-smooth optimization. In this section, to further advance the theoretical understanding and explain the inconsistency between theory and practice, we explore the advantage of adapting gradient normalization to the special Polyak-Łojasiewicz-type (PŁ) geometry in generalized-smooth optimization. We aim to show that gradient normalization, when properly adapted to the underlying PŁ geometry, can help accelerate the convergence rate in generalized-smooth optimization.

Specifically, we consider the class of generalized-smooth problems that satisfy the following generalized PŁ geometry.

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

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

The above generalized PŁ condition is inspired by the Kurdyka Łojasiewicz (KŁ)-exponent condition proposed in (Li & Pong, 2018). When $\rho > 1$, equation 5 reduces to the KŁ-exponent condition. When $\rho = 2$, equation 5 reduces to the standard PŁ condition. Moreover, some recent works have shown that PŁ-type geometries widely exist in the loss landscape of over-parametrized deep neural networks (Liu et al., 2022a; Scaman et al., 2022), and we hope that our analysis based on assumption 5 will allow researchers to rethink the relationship between optimization algorithms and loss function geometry.

Here, we consider the adaptively normalized gradient descent (AN-GD) algorithm proposed by Chen et al. (2023) for generalized-smooth nonconvex optimization. The algorithm normalizes the gradient update as follows

$$\text{(AN-GD)} \quad w_{t+1} = w_t - \gamma \frac{\nabla f(w_t)}{\|\nabla f(w_t)\|^{\beta}}, \tag{6}$$

where $\gamma > 0$ denotes the learning rate and β is a normalization scaling parameter that allows us to adapt the normalization scale of the gradient norm to the underlying function geometry. Intuitively, when the gradient norm is large, a smaller β would make the normalized gradient update more aggressive; when the gradient norm is small, normalization can slow down gradient vanishing and improve numerical stability.

Chen et al. (2023) studied AN-GD in generalized-smooth nonconvex optimization, showing that it achieves the standard $\mathcal{O}(\epsilon^{-2})$ iteration complexity lower bound. In the following theorem, we obtain the convergence rate of under the generalized PŁ condition.

Theorem 1 (Convergence Rate of AN-GD) *Let Assumptions 1 and 2 hold. Denote $\Delta_t := f(\omega_t) - f^*$ as the function value gap. Choose learning rate $\gamma = \frac{(2\mu\epsilon)^{\beta/\rho}}{8(L_0+L_1)+1}$ where ϵ denotes the target accuracy, and choose $\beta \in [\alpha, 1]$. Then, the following statements hold.*

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

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

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

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

$$\Delta_t = \mathcal{O}\Big(\big(1 - \frac{\gamma\mu}{2}\big)^t\Big). \tag{8}$$

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

- *If* $1 \geq \beta > 2-\rho$*, then there exists* $T_0 \in \mathbf{N}$ *such that for all* $t \geq T_0$*, we have*

$$\Delta_t = \mathcal{O}\Big(\Big(\frac{\Delta_{T_0}}{\gamma^{\frac{\rho}{\rho+\beta-2}}}\Big)^{\frac{\rho}{2-\beta}^{t-T_0}}\Big). \tag{9}$$

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 characterizes the convergence rates of AN-GD under different choices of ρ and β. In particular, when $\alpha = \beta = 0$ and $\rho = 2$, Theorem 1 reduces to the linear convergence rate achieved by gradient descent under the standard PŁ condition. Theorem 1 also guides the choice of gradient normalization parameter β under different geometric conditions. For example, if there exists $\beta \in [\alpha, 1]$ such that $\rho + \beta > 2$, AN-GD can boost its convergence rate from polynomial to linear convergence. When $\rho + \beta \leq 2$, the iteration complexities derived from equation 7 and equation 8 depend on the specific values of ρ and β. For example, when $\rho = 1$ and consider two different choices of gradient normalization parameters $\beta_1 = \frac{2}{3}, \beta_2 = 1$, AN-GD achieves the iteration complexities $\mathcal{O}((\frac{1}{\epsilon})^{\frac{2}{3}})$ and $\mathcal{O}(\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))$, respectively. This result matches the empirical observation made in (Chen et al., 2023) that choosing a smaller $\beta \in [\alpha, 1]$ can improve the convergence rate.

4.1 Comparison between AN-GD and GD.

The following Proposition 3 further obtains the convergence rate of gradient descent (GD) under the same Assumptions 1 and 2.

Proposition 3 (Convergence of GD) *Let Assumptions 1 and 2 hold. Assume there exists a positive constant* G *such that* $||\nabla f(x_t)|| \leq G, \forall t \in T$*, When* $\rho = 1$ *and setting* $\alpha = \beta = 1$*, by selecting* $\gamma \leq \min\{\frac{1}{L_0}, \frac{1}{2L_1 G}\}$*, gradient descent converges to an* ϵ*-stationary point within* $T = \Omega(\frac{G}{\epsilon})$ *iterations.*

From Theorem 1, under the same setting of $\alpha = \beta = 1$, $\rho = 1$, AN-GD converges to an ϵ-stationary point after $\tilde{\mathcal{O}}(\frac{1}{\epsilon})$ iterations. As a comparison, although GD can converge after $\mathcal{O}(\frac{G}{\epsilon})$ iterations, this convergence is established under additional assumption that the gradient norm is upper bounded by a constant G. Such constant normally takes a large numerical value in generalized-smooth optimization and is hard to estimate in general. Consequently, it restricts the learning rate to be very small. Similar observation is also made in (Li et al., 2024), where the convergence guarantee is established based on some constants related to the upper bound of gradient norm.

5 Independently-and-Adaptively Normalized SGD for Stochastic Generalized-Smooth Optimization

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

$$\min_{w \in \mathbf{R}^d} F(w) = \mathbb{E}_{\xi \sim \mathbb{P}}\big[f_\xi(w)\big]. \tag{10}$$

Having discussed the superior theoretical performance of AN-GD in the previous section, we aim to leverage adaptive gradient normalization to further develop an adaptively normalized algorithm tailored for stochastic generalized-smooth optimization.

5.1 Normalized SGD and Its Limitations

To solve the stochastic generalized-smooth problem in equation 10, one straightforward approach is to replace the full batch gradient in the AN-GD update rule, equation 6 with the stochastic gradient $\nabla f_\xi(w_t)$, resulting in the following adaptively normalized SGD (AN-SGD) algorithm.

$$\text{(AN-SGD)} \quad w_{t+1} = w_t - \gamma \frac{\nabla f_{\xi_t}(w_t)}{\|\nabla f_{\xi_t}(w_t)\|^\beta}. \tag{11}$$

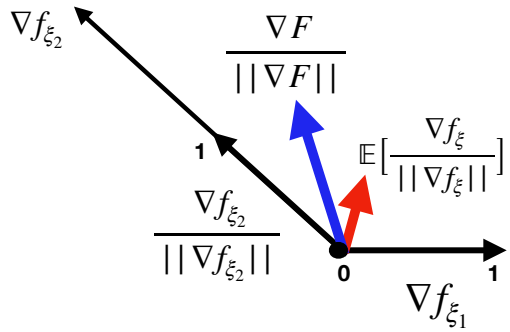

Figure 1: Comparison between normalized full gradient (blue) and expected normalized stochastic gradient (red). Here, ξ_1 and ξ_2 are sampled uniformly at random.

AN-SGD-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, previous works have shown that, when choosing $\beta = 1$ and using gradient clipping or momentum acceleration, AN-SGD's variations (e.g., NSGD, NSGD with momentum Cutkosky & Mehta (2020), clipped SGD Zhang et al. (2019; 2020)) can achieve a sample complexity of $\mathcal{O}(\epsilon^{-4})$. This result matches that of the standard SGD for solving classic stochastic smooth problems (Ghadimi & Lan, 2013). However, AN-SGD has several limitations as summarized below.

- *Biased gradient estimator:* The normalized stochastic gradient used in equation 11 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.
- *Strong assumption:* To control the estimation bias and establish theoretical convergence guarantee for ANSGD-type algorithms in generalized-smooth nonconvex optimization, the existing studies need to adopt strong assumptions. 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 practical applications, this constant can be a large numerical number if certain sample ξ happen to be an outlier.

5.2 Independently-and-Adaptively Normalized SGD

To overcome the aforementioned limitations, we propose the following independently-and-adaptively normalized stochastic gradient (IAN-SG) estimator

$$\text{(IAN-SG estimator)} \quad \frac{\nabla f_\xi(w)}{\|\nabla f_{\xi'}(w)\|^\beta}, \tag{12}$$

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

$$\mathbb{E}_{\xi,\xi'}\left[\frac{\nabla f_\xi(w)}{\|\nabla f_{\xi'}(w)\|^\beta}\right] = \mathbb{E}_{\xi'}\left[\frac{\mathbb{E}_\xi\big[\nabla f_\xi(w)\big]}{\|\nabla f_{\xi'}(w)\|^\beta}\right] \propto \nabla F(w). \tag{13}$$

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 AN-GD update. Based on this idea, we formally propose the following independently-and-adaptively normalized SGD (IAN-SGD) 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.

$$\text{(IAN-SGD):} \quad w_{t+1} = w_t - \gamma \frac{\nabla f_{\xi_B}(w_t)}{h_t^\beta},$$
$$\text{where} \quad h_t = \max\Big\{1, (4L_1\gamma)^{\frac{1}{\beta}}\Big(2\big\|\nabla f_{\xi_{B'}}(w_t)\big\| + \delta\Big)\Big\}. \tag{14}$$

The above IAN-SGD 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 IAN-SGD requires estimating the value of δ and querying two batches of samples in every iteration. However, as we show in the ablation study presented in the appendix, the convergence of IAN-SGD is robust with regard to the choice of δ, and the batch size $|B'|$ can be chosen far smaller than $|B|$.

5.3 Convergence Analysis of IAN-SGD

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*

$$\big\|\nabla f_\xi(w) - \nabla F(w)\| \le \tau_1 \big\|\nabla F(w)\big\| + \tau_2 \quad a.s. \quad \forall \xi \sim \mathbb{P}. \tag{15}$$

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*

$$\big\|\nabla f_{\xi_B}(w)\big\| \ge \frac{1}{2}\big\|\nabla F(w)\big\| - \frac{\tau_2}{2\tau_1}. \tag{16}$$

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 IAN-SGD.

Theorem 2 (Convergence of IAN-SGD) *Let Assumptions 1, 3 and 4 hold. For the IAN-SGD 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 $\frac{1}{2}$, IAN-SGD produces a sequence satisfying $\min_{t\le T} ||\nabla F(w_t)|| \le \epsilon$ if the total number of iteration T satisfies*

$$T \ge \Lambda \max\Big\{\frac{256\Lambda}{\epsilon^4}, \frac{640L_1}{\epsilon^{2-\beta}}, \frac{64(L_0 + L_1) + 128L_1(3\tau_2/\tau_1)^\beta}{\epsilon^2}\Big\}. \tag{17}$$

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 IAN-SGD 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 approximation error assumption. Through numerical experiments in Section 6 and ablation study A.1.1 later, we show that it suffices to query a small number of independent samples for IAN-SGD in practice.

5.3.1 Proof outline and novelty

The independent sampling strategy adopted by IAN-SGD 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

$$\begin{aligned}
&\mathbb{E}_{\xi_B}\big[F(w_{t+1}) - F(w_t)\big] \\
&\overset{(i)}{\leq} \frac{-\gamma\big\|\nabla F(w_t)\big\|^2}{h_t^\beta} + \frac{\gamma^2(L_0 + L_1\big\|\nabla F(w_t)\big\|^\alpha)}{2h_t^{2\beta}}\mathbb{E}_{\xi_B}\big[\big\|\nabla f_{\xi_B}(w_t)\big\|^2\big] \\
&\overset{(ii)}{\leq} \Big(\frac{\gamma}{h_t^\beta}\big(-1 + \gamma\frac{L_0 + L_1\big\|\nabla F(w_t)\big\|^\alpha}{h_t^\beta}\big)\Big)\big\|\nabla F(w_t)\big\|^2 + \frac{1}{2}\gamma^2\frac{L_0 + L_1\big\|\nabla F(w_t)\big\|^\alpha}{h_t^{2\beta}}\frac{\tau_2^2}{\tau_1^2},
\end{aligned} \tag{18}$$

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 18, 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 18, 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 18 and rearranging the terms yields that

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

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 the above inequalities over t and taking expectation leads to the desired bound in equation 17.

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.

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

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.

6 Experiments

We conduct numerical experiments to compare IAN-SGD with other state-of-the-art stochastic algorithms, including the standard SGD (Ghadimi & Lan, 2013), normalized SGD (NSGD), clipped SGD (Zhang et al., 2019)(Clip SGD), SPIDER (Fang et al., 2018), normalized SGD with momentum (Cutkosky & Mehta, 2020)(NSGDm) etc. The problems we consider are nonconvex phase retrieval, nonconvex distributionally-robust optimization and training deep neural networks.

6.1 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. Specially, 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.

$$\min_{z\in\mathbf{R}^d} f(z) = \frac{1}{2m}\sum_{r=1}^{m}\big(y_r - \big|a_r^T z\big|^2\big)^2. \tag{19}$$

Such nonconvex function satisfies generalized-smooth with parameter $\alpha = \frac{2}{3}$. 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 in original form described in previous literatures. We set batch size $|B| = 64$, and for IAN-SGD, we choose a small independent batch size $|B'| = 4$. We use fine-tuned learning rate for all algorithms, i.e., $\gamma = 5e-5$ for SGD, $\gamma = 0.2$ for NSGD and NSGD with momentum, $\gamma = 0.6$ for clipped SGD, and $\gamma = 0.25$ for SPIDER and IAN-SGD. We set the maximal gradient clipping constant as 20, $\delta = 1e-3$ for both clipped SGD and IAN-SGD. And we set normalization parameter $\beta = \frac{2}{3}$. Figure 2 (left) shows the comparison of objective value versus sample complexity. It can be observed that IAN-SGD consistently converges faster than other baseline algorithms.

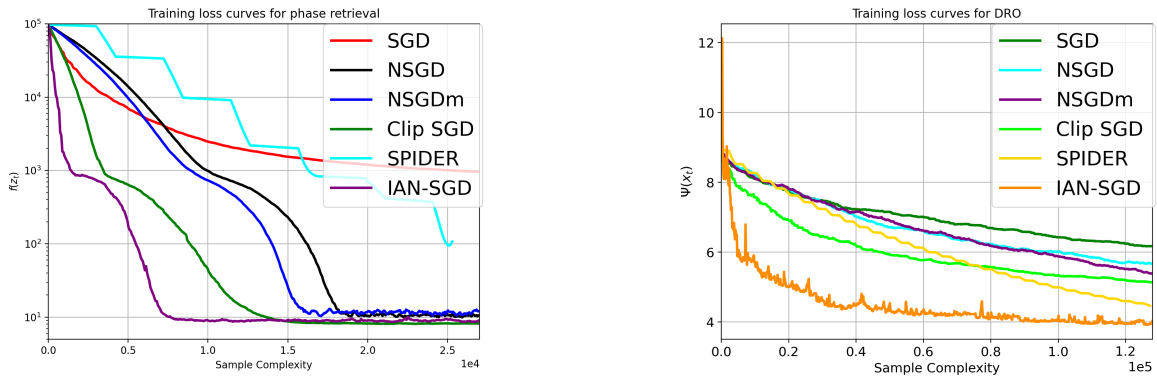

Figure 2: Experimental Results on Phase Retrieval and DRO

6.2 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

$$\min_{w\in\mathcal{W}} L(w,\eta) = \lambda\mathbb{E}_{\xi\sim P}\Psi^*\Big(\frac{\ell_\xi(w)-\eta}{\lambda}\Big) + \eta, \tag{20}$$

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 $\ell_\xi(\mathbf{w}) = \frac{1}{2}(y_\xi - \mathbf{x}_\xi^\top\mathbf{w})^2 + 0.1\sum_{j=1}^{34}\ln(1+|w^{(j)}|)$.

For moving average parameter used for acceleration method, we set it as 0.1 and 0.25 for NSGD with momentum and SPIDER respectively. For stochastic algorithms without usage of multiple mini-batches, i.e., SGD, NSGD, NSGD with momentum and clipped SGD, we set their batch sizes as $|B| = 128$. For SPIDER, we set $|B| = 128$ and $|B'| = 2313$, where the algorithm will conduct a full-gradient computation after every 15 iterations. For IAN-SGD, we set the batch size for two batch samples as $|B| = 128$ and $|B'| = 8$. We used fine-tuned learning rate for all algorithms, i.e., $\gamma = 4e-5$ for SGD, $\gamma = 5e-3$ for NSGD, NSGD with momentum and SPIDER, $\gamma = 0.11$ for clipped SGD and IAN-SGD. We set the $\delta = 1e-1$, maximal gradient clipping constant as 30, 25 for clipped SGD and IAN-SGD respectively. And we set normalization parameter $\beta = \frac{2}{3}$. Figure 2 (Right) shows the comparison of objective value versus sample

complexity. It can be observed that objective value optimized by IAN-SGD consistently converges faster than other baselines algorithms.

6.3 Deep Neural Networks

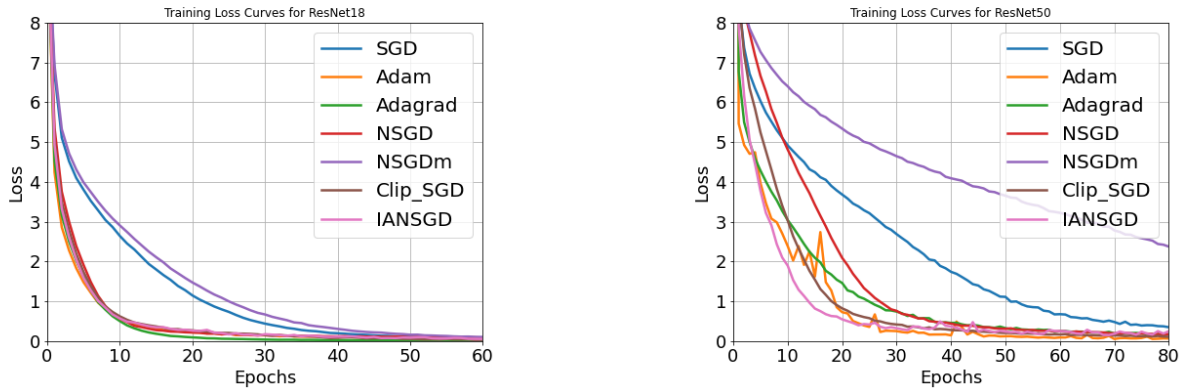

Figure 3: Experimental Result on training ResNet18, ResNet50.

According to Zhang et al. (2019), generalized-smooth has been observed to hold in deep neural networks. To further demonstrate the effectiveness of IAN-SGD algorithm, we conduct experiments for training deep neural networks. Specially, we train ResNet18, ResNet50 (He et al., 2016) on CIFAR10 Dataset (Krizhevsky, 2009) from scratch. We resize images as 32×32 and normalize images with standard derivation equals to 0.5 on each dimension. At the beginning of each algorithm, we fix random seed and initialize model parameters using Kaiming initialization. We compare our algorithm with baseline methods, including SGD (Robbins & Monro, 1951), Adam (Kingma, 2014), Adagrad (Duchi et al., 2011a), NSGD, NSGD with momentum (Cutkosky & Mehta, 2020) and clipped SGD (Zhang et al., 2019).

For SGD, Adam and Adagrad, we utilize pytorch built-in optimizer to implement training pipeline. We implement training pipeline for NSGD, NSGD with momentum, clipped SGD and IAN-SGD. The normalization constant is computed through all model parameters at each iteration. The detailed algorithm settings are as following. For batch size, all algorithms use $B = 128$, and $B' = 32$ for IAN-SGD. For moving average parameter, we use 0.9, 0.99 for Adam, and 0.25 for normalized SGD with momentum. For clipping threshold used in clipped SGD and IAN-SGD, we set them as 2 and $\delta = 1e - 1$. The normalization power used for IAN-SGD is $\beta = \frac{2}{3}$. We use fine-tuned learning rate for all algorithms, i.e., $\gamma = 1e - 3$ for SGD, Adam and Adagrad, $\gamma = 1e - 1$ for NSGD and NSGD with momentum, $\gamma = 2e - 1$ for clipped SGD and IAN-SGD. We trained ResNet18, ResNet50 on CIFAR10 dataset for 30 epochs and plot the training loss in Figure 3. Figure 3 (left) shows the training loss of ResNet18, Figure3 (right) shows the training loss of ResNet50. As we can see from these figures, the (pink) loss curve optimized by IAN-SGD indicates fast convergence rate comparable with several baselines, including SGD, NSGD NSGDm, clipped SGD, which demonstrate the effectiveness of IAN-SGD framework.

7 Conclusions

In this work, we provide theoretical insights on how normalization interplays with function geometry, and their overall effects on convergence. We then 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(61):2121–2159, 2011a.

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

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

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.

Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.

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.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

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.

Daniil Vankov, Angelia Nedich, and Lalitha Sankar. Generalized smooth stochastic variational inequalities: Almost sure convergence and convergence rates, 2024a.

Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian U. Stich. Optimizing (l_0, l_1)-smooth functions by gradient methods, 2024b.

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.

Appendix Table of Contents

1 Introduction **1**

2 Related Work **2**

3 Generalized-Smooth nonconvex Optimization **3**

4 Adaptive Gradient Normalization for Deterministic Generalized-Smooth Optimization **4**

5 Independently-and-Adaptively Normalized SGD for Stochastic Generalized-Smooth Optimization **6**

6 Experiments **9**

7 Conclusions **11**

A Ablation Study **16**

B Proof of Descent Lemma 1 **19**

C Proof of Propositions 1, 2 **19**

D Proof of Descent Lemma under Generalized PŁ condition **22**

E Proof of Theorem 1 **23**

F Proof of Proposition 3 **26**

G Proof of Theorem 2 **27**

H Proof of Lemma 2 **30**

I Lemma 6 and proof **32**

A Ablation Study

In order to have a comprehensive understanding of the performance of IAN-SGD in practical problems, in this section, we conduct ablation study on Phase Retrieval and DRO regarding on important components of IAN-SGD separately, i.e., adaptive normalization, batch size of independent samples and numerical stabilizer δ.

A.1 Effects of β

To justify the advantage of using adaptive normalization in practice, we conduct the following two experiments.

In first experiment, we unify the normalization parameter of all the normalized methods, i.e., NSGD, NSGD with momentum, clipped SGD, SPIDER and IAN-SGD to $\beta = \frac{2}{3}$. To guarantee convergence, we adjust the learning rate accordingly, i.e., $\gamma = 0.03$ for NSGD and NSGD with momentum, $\gamma = 0.05$ for SPIDER, $\gamma = 0.17$ for both clipped SGD and IAN-SGD. To make a fair comparison, we keep other parameters unchanged. Figure 4 (left) shows the comparison of objective value versus sample complexity for Phase Retrieval problem. It can be observed that, by adjusting $\beta = \frac{2}{3}$, the objective value optimized by all algorithms decreases much faster compared with Figure 2, this indicates adaptive normalization can accelerate convergence. Moreover, compared with other normalization methods, even though IAN-SGD requires additional sampling at each iteration, the training loss still decreases faster than NSGD SGD with momentum and SPIDER.

Similarly, for DRO, we unify the normalization parameter of all the normalized methods, i.e., NSGD, NSGD with momentum, clipped SGD, SPIDER and IAN-SGD to have the same $\beta = \frac{2}{3}$. To guarantee algorithm convergence, we adjust learning rate correspondingly, for NSGD, NSGDm and SPIDER, we keep learning rate unchanged, for clipped SGD, and IAN-SGD, we set $\gamma = 0.08$. To make a fair comparison, we keep other parameters unchanged. Figure 4 (Right) shows the comparison of objective value versus sample complexity. It can be observed that, by setting $\beta = \frac{2}{3}$, the objective value optimized by all normalization methods decreases faster than Figure 2. This verifies the effectiveness of adaptive normalization. Even though IAN-SGD requires additional sampling, it converges faster than other normalized methods.

In summary, results in Figure 4 indicates independent sampling and adaptive normalization doesn't increase the sample complexity to find a stationary point, which justifies IAN-SGD framework's effectiveness when dealing with nonconvex geometry characterized by generalized-smooth condition.

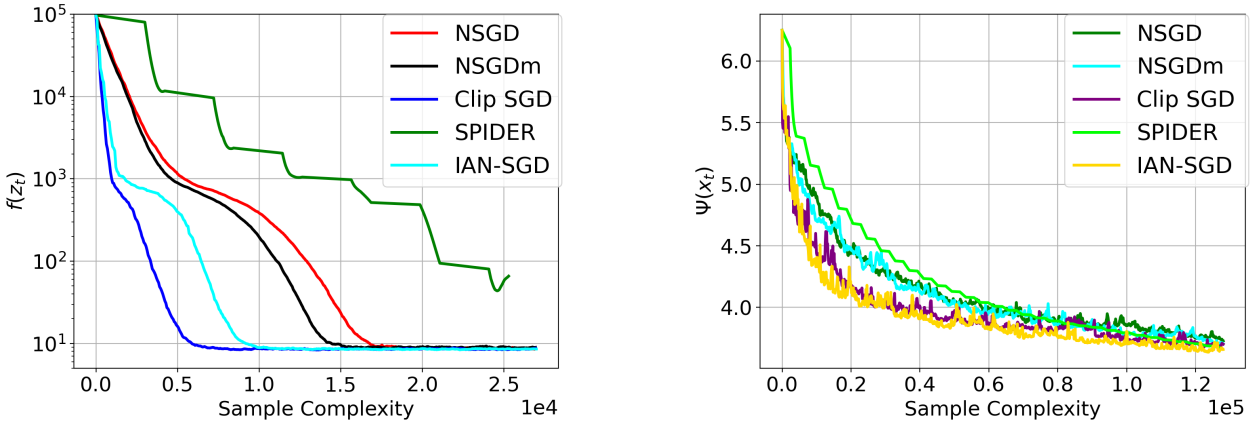

Figure 4: advantage of using adaptive normalization on normalized first-order algorithms

In second experiment, we vary β for IAN-SGD and keep other parameters unchanged. Figure 5 shows the convergence result with different β for Phase Retrieval and DRO. It can be observed that decreasing β

in general accelerate convergence. However, small β can make convergence unstable. In DRO experiment showed in Figure 5 (Right), we observed when $\beta = \frac{3}{5}$, the objective value curves vibrates a lot. If β is further reduced less than $\frac{3}{5}$, the algorithm fails to converge. This phenomena coincides with implications of theorem 1, where β must satisfy $\beta \in [\alpha, 1]$ to guarantee convergence.

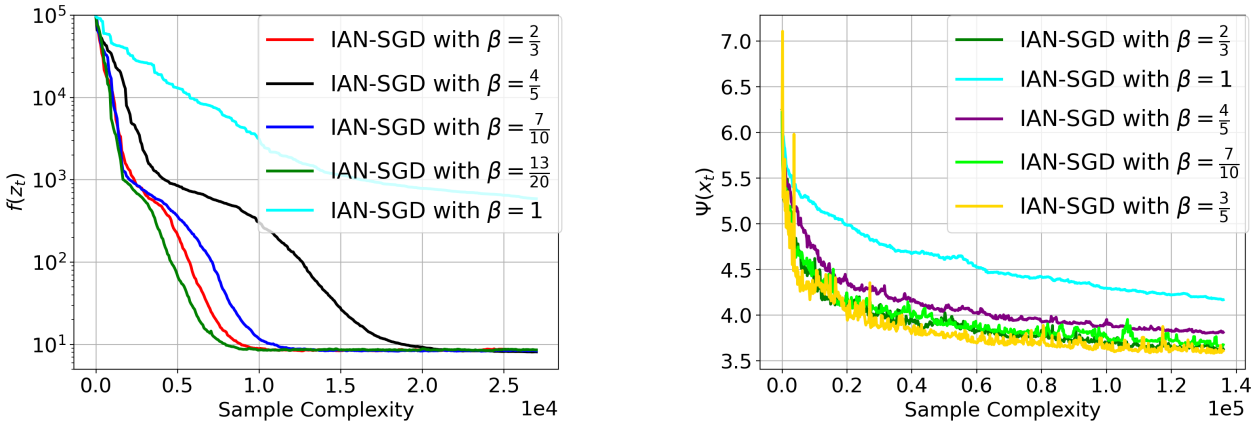

Figure 5: Effects of adaptive normalization on convergence of IAN-SGD

A.1.1 Effects of Batch size

In this section, we justify the effects of batch size for independent samples used in IAN-SGD. For Phase retrieval, we keep $|B| = 64$ and other parameters same as section A.1, and we vary independent batch sizes to be $|B'| = 4, 8, 16, 32, 64$. Similarly, for DRO, we keep $|B| = 128$ and other parameters same as section A.1, and we vary batch size of independent samples $|B'| = 16, 32, 64, 128$.

The following figure 6 shows the convergence of IAN-SGD under different batch size choices for Phase Retrieval (Left) and DRO (right). It can be observed for Phase retrieval and DRO problem, small batch

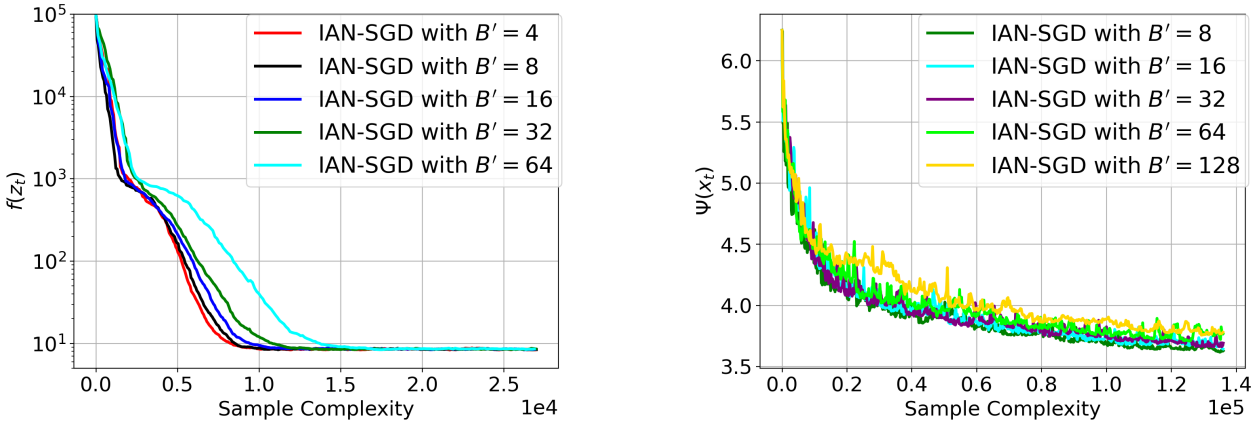

Figure 6: Effects of independent samples' batch size on convergence

size $|B'| = 4, 8$ are enough to guarantee algorithm convergence, and increasing independent samples' batch size $|B'|$ doesn't increase sample complexity too much for finding a stationary point, which verifies the effectiveness of independent sampling strategy.

A.1.2 Effects of δ

In Theorem 2, we set $\delta = \frac{\tau_2}{\tau_1}$ based on assumption 4, which assumes an affine bound for approximation error $||\nabla f_\xi(w) - \nabla F(w)||$. This assumption relaxes strong assumptions used in (Zhang et al., 2019; Liu et al., 2022b; Zhang et al., 2020), where they assume approximation error $||\nabla f(w) - \nabla F(w)||$ is upper bounded by a constant. The major weakness is when certain samples leads to gradient outlier, such assumption leads to loose upper bound. Thus, to verify the effectiveness of assumption 4, we expect convergence of IAN-SGD under wide range of δ, especially for smaller δ.

To justify the δ effects on convergence, we keep other parameters same as section A.1 and only vary δ. For Phase Retrieval, we vary $\delta = \{1e^{-8}, 1e^{-3}, 1e^{-1}, 1, 10\}$ and figure 7 (Left) shows the corresponding convergence result. For Distributionally Robust Optimization, we vary $\delta = \{1e^{-8}, 1e^{-3}, 1e^{-1}, 1, 10\}$ and figure 7 (Right) shows the convergence result. We observe that IAN-SGD convergence is robust to the choice of δ. But When $\delta = 1e^{-3}, 1e^{-1}$ demonstrate better convergence than others for Phase retrieval and Distributionally Robust Optimization respectively. This result indicates small δ is enough to guarantee convergence, which verifies the effectiveness of assumption 4 and IAN-SGD when dealing with stochastic nonconvex generalized-smooth geometry.

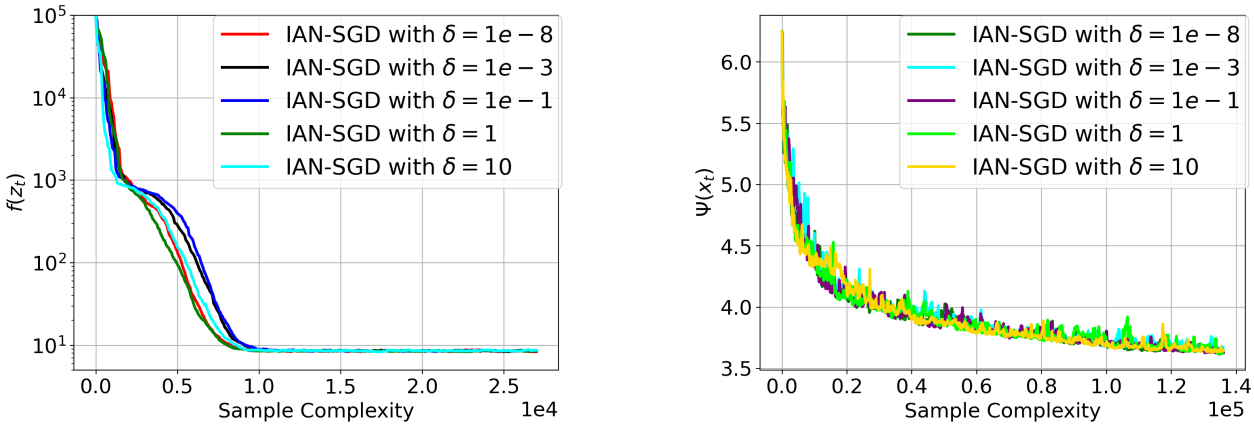

Figure 7: Effects of δ on convergence

B Proof of Descent Lemma 1

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

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

Proof 1 *Use fundamental theorem of calculus, we have*

$$\begin{aligned} & f(w') - f(w) - \langle \nabla f(w), w' - w \rangle \\ & = \int_0^1 \langle \nabla f(w_\theta), w' - w \rangle \mathrm{d}\theta - \int_0^1 \langle \nabla f(w), w' - w \rangle \mathrm{d}\theta, \end{aligned}$$

where $w_\theta = \theta w' + (1-\theta) w$. Since the integration integrates over w_θ, integrating second term doesn't affect the result. Now replacing above term by $\mathcal{L}^*_{asym}(\alpha)$ condition, we have

$$\begin{aligned} & f(w') - f(w) - \langle \nabla f(w), w' - w \rangle \\ & = \int_0^1 \langle \nabla f(w_\theta), w' - w \rangle \mathrm{d}\theta - \int_0^1 \langle \nabla f(w), w' - w \rangle \mathrm{d}\theta \\ & = \int_0^1 \langle \nabla f(w_\theta) - \nabla f(w), w' - w \rangle \mathrm{d}\theta \\ & \leq \int_0^1 \|\nabla f(w_\theta) - \nabla f(w)\| \|w' - w\| \mathrm{d}\theta \\ & \leq \int_0^1 \theta (L_0 + L_1 \|\nabla f(w_t)\|^{\alpha}) \|w' - w\|^2 \mathrm{d}\theta \\ & = \frac{1}{2}(L_0 + L_1 \|\nabla f(w_t)\|^{\alpha}) \|w' - w\|^2, \end{aligned} \tag{21}$$

where the first inequality is due to Cauchy-schwarz inequality, the second inequality is due to Assumption 1 regarding on $\mathcal{L}^*_{asym}(\alpha)$ generalized-smooth. Reorganize above inequality gives us the desired result.

C Proof of Propositions 1, 2

C.1 Proof for nonconvex Phase Retrieval

Proposition 1 *The nonconvex phase retrieval objective function (see equation 19 in the appendix) satisfies equation 4 with $\alpha = \frac{2}{3}$.*

The proof of generalized-smooth property for nonconvex Phase retrieval problem is similar as proof in Chen et al. (2023) with minor changes. We present the proof details here for completeness.

Proof 2 *The objective function of phase retrieval problem can be rewritten in the stochastic form $f(z) = \mathbb{E}_\xi f_\xi(z)$ where ξ is obtained from $\{1, 2...., m\}$ uniformly at random and*

$$f_\xi(z) = \frac{1}{2}\left(y_\xi - \left|a_\xi^\top z\right|^2\right)^2. \tag{22}$$

To prove that f satisfies induced symmetric generalized-smooth equation 4, it suffices to prove that for every ξ, f_ξ satisfies inequality equation 4 with $\alpha = \frac{2}{3}$.

For arbitrary $z \in \mathbf{R}^d$, the gradient of $f_\xi(z)$ has the following lower bound

$$\begin{aligned}
\|\nabla f_\xi(z)\|^{\frac{2}{3}} &= \frac{1}{2^{\frac{2}{3}}}\|\big(|a_\xi^\top z|^2 - y_\xi\big)\big(a_\xi a_\xi^\top\big)z\|^{\frac{2}{3}} \\
&\geq \frac{1}{2}\Big||a_\xi^\top z|^3 - y_\xi|a_\xi^\top z|\Big|^{\frac{2}{3}}\|a_\xi\|^{\frac{2}{3}} \\
&\overset{(i)}{\geq} \frac{1}{2}\big(|a_\xi^\top z|^2 - |y_\xi\|a_\xi^\top z|^{\frac{2}{3}}\big)|\,\|a_\xi\|^{\frac{2}{3}} \\
&\overset{(ii)}{\geq} \frac{1}{3}\big(|a_\xi^\top z|^2 - |y_\xi|^{\frac{3}{2}}\big)|\,\|a_\xi\|^{\frac{2}{3}},
\end{aligned} \tag{23}$$

where (i) uses inequality $|a-b|^{\frac{2}{3}} \geq |a|^{\frac{2}{3}} - |b|^{\frac{2}{3}}$ for any $a, b \in \mathbf{R}$; (iii) uses young's inequality, where we have $|y_\xi|^{\frac{2}{3}}a^{\frac{2}{3}} \leq \frac{1}{3}a^2 + \frac{2}{3}|y_\xi|^{\frac{3}{2}}$ for any $a > 0$.

Then, for any $z, z' \in \mathbf{R}^d$, we have

$$\begin{aligned}
&\|\nabla f_\xi(z') - \nabla f_\xi(z)\| \\
&= \frac{1}{2}\|\big(|a_\xi^\top z'|^2 - y_\xi\big)\big(a_\xi a_\xi^\top\big)z' - \big(|a_\xi^\top z|^2 - y_\xi\big)\big(a_\xi a_\xi^\top\big)z\| \\
&= \frac{1}{4}\|2\big(|a_\xi^\top z'|^2 - y_\xi\big)\big(a_\xi a_\xi^\top\big)z' - 2\big(|a_\xi^\top z|^2 - y_\xi\big)\big(a_\xi a_\xi^\top\big)z \\
&\quad + \big(|a_\xi^\top z'|^2 - y\big)\big(a_\xi a_\xi^\top\big)z - \big(|a_\xi^\top z|^2 - y\big)(a_\xi a_\xi^\top)z' \\
&\quad - \big(|a_\xi^\top z'|^2 - y\big)\big(a_\xi a_\xi^\top\big)z + \big(|a_\xi^\top z|^2 - y\big)(a_\xi a_\xi^\top)z'\| \\
&= \frac{1}{4}\|\big(|a_\xi^\top z'|^2 - y_\xi\big)\big(a_\xi a_\xi^\top\big)z' - \big(|a_\xi^\top z'|^2 - y\big)\big(a_\xi a_\xi^\top\big)z \\
&\quad + \big(|a_\xi^\top z|^2 - y\big)(a_\xi a_\xi^\top)z' - \big(|a_\xi^\top z|^2 - y_\xi\big)\big(a_\xi a_\xi^\top\big)z \\
&\quad + \big(|a_\xi^\top z'|^2 - y_\xi\big)\big(a_\xi a_\xi^\top\big)z' + \big(|a_\xi^\top z'|^2 - y\big)\big(a_\xi a_\xi^\top\big)z \\
&\quad - \big(|a_\xi^\top z|^2 - y\big)(a_\xi a_\xi^\top)z' - \big(|a_\xi^\top z|^2 - y_\xi\big)\big(a_\xi a_\xi^\top\big)z\| \\
&= \frac{1}{4}\|\big(|a_\xi^\top z'|^2 + |a_\xi^\top z|^2 - 2y_\xi\big)\big(a_\xi a_\xi^\top\big)\big(z' - z\big) \\
&\quad + \big(|a_\xi^\top z'|^2 - |a_\xi^\top z|^2\big)\big(a_\xi a_\xi^\top\big)\big(z' + z\big)\| \\
&\overset{(i)}{\leq} \frac{1}{4}\|a_\xi\|^2\big(|a_\xi^\top z'|^2 + |a_\xi^\top z|^2 + 2|y_\xi|\big)\|z' - z\| \\
&\quad + \frac{1}{4}\|a_\xi\|^2\big(|a_\xi^\top z'| + |a_\xi^\top z|\big)^2\|z' - z\| \\
&\overset{(ii)}{\leq} \frac{1}{4}\|z' - z\|\|a_\xi\|^2\big(3|a_\xi^\top z'|^2 + 3|a_\xi^\top z|^2 + 2|y_\xi|\big) \\
&\leq \frac{1}{4}\|z' - z\|\|a_\xi\|^{\frac{4}{3}}\|a_\xi\|^{\frac{2}{3}} \\
&\quad \cdot \big(3|a_\xi^\top z'|^2 + 3|a_\xi^\top z|^2 - 3|y_\xi| - 3|y_\xi| + 8|y_\xi|\big) \\
&\overset{(iii)}{\leq} \|z' - z\|\big(\frac{9}{4}a_{\max}^{\frac{4}{3}}\|\nabla f_\xi(z')\|^{\frac{2}{3}} \\
&\quad + \frac{9}{4}a_{\max}^{\frac{4}{3}}|\nabla f_\xi(z)|^{\frac{2}{3}} + 2y_{\max}a_{\max}^2\big)
\end{aligned} \tag{24}$$

where (i) uses triangular inequality, $||a_\xi a_\xi^\top|| = ||a_\xi||^2$, $|y_\xi| \leq 1$ and inequality equation 25; (ii) uses $(|a_\xi^\top z'| + |a_\xi^\top z|)^2 \leq 2|a_\xi^\top z'|^2 + 2|a_\xi^\top z|^2$; (iii) uses equation 23 and denotes$y_{\max} = \max_{1\leq r\leq m}|y_r|$ and $a_{\max} = \max_{1\leq r\leq m}\|a_r\|$. Thus, in summary, nonconvex Phase retrieval equation 19 satisfies induced

generalized-smooth equation 4 with $\alpha = \frac{2}{3}$, $L_0 = 2y_{\max}a_{\max}^2$, *and* $L_1 = \frac{9}{2}a_{\max}^{\frac{4}{3}}$.

$$\begin{aligned}\left|\left|a_\xi^\top z'\right|^2 - \left|a_\xi^\top z\right|^2\right| &= \left(\left|a_\xi^\top z'\right| + \left|a_\xi^\top z\right|\right)\left(\left|a_\xi^\top z'\right| - \left|a_\xi^\top z\right|\right) \\ &\leq \left(\left|a_\xi^\top z'\right| + \left|a_\xi^\top z\right|\right)\left\|a_\xi^\top (z' - z)\right\| \\ &\leq \left\|a_\xi^\top\right\|\left(\left|a_\xi^\top z'\right| + \left|a_\xi^\top z\right|\right)\left\|z' - z\right\|. \end{aligned} \tag{25}$$

C.2 Proof for DRO

Proposition 2 *The distributionally robust optimization (DRO) objective function (see equation 20 in experiment section) satisfies equation 4 with* $\alpha = 1$.

The proof of generalized-smooth property for Distributionally Robust Optimization is exactly same as Jin et al. (2021); Chen et al. (2023). We refer readers to check Appendix D.2 in Chen et al. (2023) for details. In short, DRO problem satisfies asymmetric generalized-smooth 2 in assumption 1, thus it also satisfies induced inequality 4 with $\alpha = 1$

D Proof of Descent Lemma under Generalized PŁ condition

Lemma 3 *For any $x \geq 0$, $C \in [0,1]$, $\Delta > 0$, and $0 \leq w \leq w'$ such that $\Delta \geq w' - w$, we have the following inequality hold*

$$Cx^{w} \leq x^{w'} + C^{\frac{w'}{\Delta}}. \tag{26}$$

The proof details for this lemma can be found at Chen et al. (2023), Lemma E.2 at Appendix.

Lemma 4 (Descent Lemma under Generalized PL condition) *Let Assumption 1 and 2 hold. Apply AN-GD, choose $\beta \in [\alpha, 1]$ or $\beta \in (\alpha, 1]$, when $\alpha \in (0,1]$ or $\alpha = 0$ respectively. Set the target accuracy ϵ satisfy $0 \leq \epsilon \leq \min\{1, 1/2\mu\}$. Define the step size $\gamma = \frac{(2\mu\epsilon)^{\beta/\rho}}{8(L_0+L_1)+1}$. Denote $\Delta_t = f(w_t) - f^*$, then we have descent lemma*

$$\Delta_{t+1} \leq \Delta_t - \frac{\gamma(2\mu)^{\frac{2-\beta}{\rho}}}{4}\Delta_t^{\frac{2-\beta}{\rho}}. \tag{27}$$

Proof 3 *Start from descent lemma 1, we have*

$$\begin{aligned}
&f\big(w_{t+1}\big) - f\big(w_t\big) \\
&\overset{(i)}{\leq} \nabla f\big(w_t\big)^{\top}\big(w_{t+1} - w_t\big) + \frac{1}{2}\big(L_0 + L_1\big\|\nabla f\big(w_t\big)\big\|^{\alpha}\big)\big\|w_{t+1} - w_t\big\|^2 \\
&\overset{(ii)}{=} -\gamma\big\|\nabla f\left(w_t\right)\big\|^{2-\beta}\frac{\gamma}{4}\big(2L_0\gamma \cdot \big\|\nabla f\big(w_t\big)\big\|^{2-2\beta} + 2L_1\gamma \cdot \big\|\nabla f\left(w_t\right)\big\|^{2+\alpha-2\beta}\big) \\
&\overset{(iii)}{\leq} -\gamma\big\|\nabla f\big(w_t\big)\big\|^{2-\beta} + \frac{\gamma}{4}\big(2\big\|\nabla f\big(w_t\big)\big\|^{2-\beta} + \big(2L_0\gamma\big)^{\frac{2}{\beta}-1} + \big(2L_1\gamma\big)^{\frac{2}{\beta}-1}\big) \\
&\overset{(iv)}{\leq} -\frac{\gamma}{2}\big\|\nabla f\left(w_t\right)\big\|^{2-\beta} + \gamma^{\frac{2}{\beta}}\big(2L_0 + 2L_1\big)^{\frac{2}{\beta}-1} \\
&\overset{(v)}{\leq} -\frac{\gamma}{2}\big\|\nabla f\left(w_t\right)\big\|^{2-\beta} + \frac{(2\mu\epsilon)^{\frac{2}{\rho}}}{(8(L_0+L_1)+1)^{\frac{2}{\beta}}}(\frac{1}{4})^{\frac{2}{\beta}-1}(8(L_0+L_1)+1)^{\frac{2}{\beta}-1} \\
&\overset{(vi)}{\leq} -\frac{\gamma}{2}\big\|\nabla f\left(w_t\right)\big\|^{2-\beta} + \frac{1}{4}\frac{(2\mu\epsilon)^{\frac{\beta}{\rho}}(2\mu\epsilon)^{\frac{2-\beta}{\rho}}}{8(L_0+L_1)+1} \\
&= -\frac{\gamma}{2}\big\|\nabla f\left(w_t\right)\big\|^{2-\beta} + \frac{\gamma}{4}(2\mu\epsilon)^{\frac{2-\beta}{\rho}},
\end{aligned} \tag{28}$$

where (i) follows from lemma 1; (ii) follows from update rule of AN-GD, namely replacing $w_{t+1} - w_t$ by $\frac{\nabla f(w_t)}{||\nabla f(w_t)||^{\beta}}$, (iii) follows from aggregates constant term by 6 and utilize technical lemma 3 by letting $\omega' = 2-\beta$, $\Delta = \beta$ and applying it to $2L_0\gamma||\nabla f(w_t)||^{2-2\beta}$, $2L_1\gamma||\nabla f(w_t)||^{2+\alpha-2\beta}$ twice gives the desired result; (iv) follows from $a^{\tau} + b^{\tau} \leq (a+b)^{\tau}$ holds for $\tau = 2/\beta - 1 > 1$ and $a, b \geq 0$, (v) follows from the step size rule $\gamma = (2\mu\epsilon)^{\beta/\rho}/(8(L_0+L_1)+1)$, (vi) following from the fact $0 < \beta \leq 1$, thus $\frac{1}{4}^{(2/\beta)-1} < \frac{1}{4}$. For function satisfying generalized PŁ-condition proposed in definition 5, we have

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

This is equivalent as

$$\big\|\nabla f(w)\big\|^{2-\beta} \geq (2\mu)^{\frac{2-\beta}{\rho}}\big(f(w) - f^*\big)^{\frac{2-\beta}{\rho}}. \tag{29}$$

Substitute equation 29 into equation 30, we have

$$f(w_{t+1}) - f(w_t) \leq -\frac{\gamma}{2}(2\mu)^{\frac{2-\beta}{\rho}}\big(f(w_t) - f^*\big)^{\frac{2-\beta}{\rho}} + \frac{\gamma}{4}(2\mu\epsilon)^{\frac{2-\beta}{\rho}}.$$

Subtract f^* on both sides, it is equivalent as

$$f(w_{t+1}) - f^* \leq f(w_t) - f^* - \frac{\gamma}{2}(2\mu)^{\frac{2-\beta}{\rho}}(f(w) - f^*)^{\frac{2-\beta}{\rho}} + \frac{\gamma}{4}(2\mu\epsilon)^{\frac{2-\beta}{\rho}}.$$

Now, denote $\Delta_t = f(w_t) - f^*$, we have the equivalent representation

$$\Delta_{t+1} \leq \Delta_t - \frac{\gamma(2\mu)^{\frac{2-\beta}{\rho}}}{2}\Delta_t^{\frac{2-\beta}{\rho}} + \frac{\gamma}{4}(2\mu\epsilon)^{\frac{2-\beta}{\rho}}. \tag{30}$$

By Choosing the stopping criterion as

$$T = \inf\left\{t | \Delta_t = f(w_t) - f^* \leq \epsilon\right\}, \text{ where } 0 < \epsilon \leq \min\{1, \frac{1}{2\mu}\}.$$

We conclude before algorithm terminates, $\Delta_t > \epsilon$ for all $t \leq T$, thus $-\frac{\gamma(2\mu)^{\frac{2-\beta}{\rho}}}{2}\Delta_t^{\frac{2-\beta}{\rho}}$ dominates $\frac{\gamma}{4}(2\mu\epsilon)^{\frac{2-\beta}{\rho}}$. Moreover, by definition of β, we have $\frac{2-\beta}{\rho} > 0$ and thus

$$\Delta_t^{\frac{2-\beta}{\rho}} > \epsilon^{\frac{2-\beta}{\rho}},$$

which is equivalent to claim

$$\frac{\gamma}{4}(2\mu)^{\frac{2-\beta}{\rho}}\Delta_t^{\frac{2-\beta}{\rho}} > \frac{\gamma}{4}(2\mu\epsilon)^{\frac{2-\beta}{\rho}}.$$

Thus, equation 30 reduces to relaxed descent inequality

$$\Delta_{t+1} \leq \Delta_t - \frac{\gamma(2\mu)^{\frac{2-\beta}{\rho}}}{4}\Delta_t^{\frac{2-\beta}{\rho}}. \tag{31}$$

E Proof of Theorem 1

Theorem 1 (Convergence Rate of AN-GD) *Let Assumptions 1 and 2 hold. Denote $\Delta_t := f(\omega_t) - f^*$ as the function value gap. Choose learning rate $\gamma = \frac{(2\mu\epsilon)^{\beta/\rho}}{8(L_0+L_1)+1}$ where ϵ denotes the target accuracy, and choose $\beta \in [\alpha, 1]$. Then, the following statements hold.*

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

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

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

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

$$\Delta_t = \mathcal{O}\Big(\big(1 - \frac{\gamma\mu}{2}\big)^t\Big). \tag{8}$$

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

- *If $1 \geq \beta > 2 - \rho$, then there exists $T_0 \in \mathbf{N}$ such that for all $t \geq T_0$, we have*

$$\Delta_t = \mathcal{O}\Big(\Big(\frac{\Delta_{T_0}}{\gamma^{\frac{\rho}{\rho+\beta-2}}}\Big)^{\frac{\rho}{2-\beta}^{t-T_0}}\Big). \tag{9}$$

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

Proof 4 *We divide the convergence proof of theorem 1 into three cases depending on the value of β and ρ.*

Case I: When $\rho < 2-\beta$
This is equivalent as $\frac{2-\beta}{\rho} > 1$. Now denote $\theta = \frac{2-\beta}{\rho}$. Since $\theta > 1$, we have following inequalities hold

$$\begin{aligned}
&\Delta_{t+1} \leq \Delta_t \\
&\Delta_{t+1}^{\theta} \leq \Delta_t^{\theta} \\
&\Delta_{t+1}^{-\theta} \geq \Delta_t^{-\theta}.
\end{aligned} \tag{32}$$

Now define an auxiliary function $\Phi(t) = \frac{1}{\theta-1}t^{1-\theta}$. Its derivative can be computed via $\Phi'(t) = -t^{-\theta}$ We now divide the last inequality at equation 32 into two different cases for analysis. One is the case where $\Delta_{t+1}^{-\theta} \leq 2\Delta_t^{-\theta}$, Another is the case where $\Delta_{t+1}^{-\theta} \geq 2\Delta_t^{-\theta}$.

When $\Delta_t^{-\theta} \leq \Delta_{t+1}^{-\theta} \leq 2\Delta_t^{-\theta}$, we have

$$\begin{aligned}
\Phi(\Delta_{t+1}) - \Phi(\Delta_t) &= \int_{\Delta_t}^{\Delta_{t+1}} \Phi'(t)\mathrm{d}t = \int_{\Delta_{t+1}}^{\Delta_t} t^{-\theta}\mathrm{d}t \\
&\geq (\Delta_t - \Delta_{t+1})\Delta_t^{-\theta} \\
&\geq (\Delta_t - \Delta_{t+1})\frac{\Delta_{t+1}^{-\theta}}{2} \\
&\geq \frac{\gamma(2\mu)^{\theta}}{4}\Delta_t^{\theta}\frac{\Delta_{t+1}^{-\theta}}{2} \\
&\geq \frac{\gamma(2\mu)^{\theta}}{4}\Delta_{t+1}^{\theta}\frac{\Delta_{t+1}^{-\theta}}{2} = \frac{\gamma(2\mu)^{\theta}}{8}.
\end{aligned}$$

The first inequality is using mean value theorem such that $\Phi(\Delta_{t+1}) - \Phi(\Delta_t) = |\Delta_{t+1} - \Delta_t||\Phi'(\xi)|$, where $\xi \in [\Delta_t, \Delta_{t+1}]$. Since $\Phi(\Delta_{t+1}) - \Phi(\Delta_t) \geq 0$, taking absolute value has no effect. Since $\theta > 0$, $|\Phi'(t)| = t^{-\theta}$ is monotone decreasing. Thus, we always have $\Delta_t^{-\theta} \leq \Phi'(\xi) \leq \Delta_{t+1}^{-\theta}$ for any $\xi \in [\Delta_t, \Delta_{t+1}]$; The second inequality uses the fact $\Delta_{t+1}^{-\theta} \leq 2\Delta_t^{-\theta}$; The third inequality is due to the recursion $\Delta_t - \Delta_{t+1} \geq \frac{\gamma(2\mu)^{\theta}}{4}\Delta_t^{\theta}$; The last inequality uses the fact that $\Delta_t^{\theta} > \Delta_{t+1}^{\theta}$ for all $\theta > 0$.

When $\Delta_{t+1}^{-\theta} > 2\Delta_t^{-\theta}$, it holds that $\Delta_{t+1}^{1-\theta} = (\Delta_{t+1}^{-\theta})^{\frac{1-\theta}{-\theta}} > 2^{\frac{1-\theta}{-\theta}}\Delta_t^{1-\theta}$. Then, we have

$$\begin{aligned}
\Phi(\Delta_{t+1}) - \Phi(\Delta_t) &= \frac{1}{\theta-1}(\Delta_{t+1}^{1-\theta} - \Delta_t^{1-\theta}) \\
&\geq \frac{1}{\theta-1}\big((2)^{\frac{\theta-1}{\theta}} - 1\big)\Delta_t^{1-\theta} \\
&\geq \frac{1}{\theta-1}\big((2)^{\frac{\theta-1}{\theta}} - 1\big)\Delta_0^{1-\theta},
\end{aligned}$$

where the first inequality is due to the recursion $\Delta_{t+1}^{1-\theta} = (\Delta_{t+1}^{-\theta})^{\frac{1-\theta}{-\theta}} > 2^{\frac{1-\theta}{-\theta}}\Delta_t^{1-\theta}$; the last inequality is due to the fact the sequence $\{\Delta_t\}_{t=1}^{T}$ is non-increasing. Now put the expression of θ in and denote

$$C = \min\Big\{\frac{\gamma(2\mu)^{\frac{2-\beta}{\rho}}}{8}, \frac{\rho}{2-\beta-\rho}(2^{\frac{2-\beta-\rho}{2-\beta}} - 1)\Delta_0^{\frac{2-\beta-\rho}{\rho}}\Big\}.$$

We conclude for all t, we have

$$\Phi(\Delta_t) \geq \sum_{i=0}^{t-1}\Phi(\Delta_{i+1}) - \Phi(\Delta_i) \geq Ct,$$

Thus, we have

$$\Delta_t \leq \Big(\frac{\rho}{(2-\beta-\rho)Ct}\Big)^{\frac{\rho}{2-\rho-\beta}} = \mathcal{O}\Big(\frac{\rho}{(2-\beta-\rho)\gamma t}\Big)^{\frac{\rho}{2-\rho-\beta}}, \tag{33}$$

When $C = \frac{\rho}{2-\beta-\rho}(2^{(2-\beta-\rho)/(2-\beta)} - 1)\Delta_0^{(2-\beta-\rho)/\rho}$*, in order to make* $\Delta_t \leq \epsilon$*, we have*

$$\rho/(2-\rho-\beta)\log\big((2-\beta-\rho)Ct/\rho\big) = \log\big(1/\epsilon\big),$$

which indicates $T = \mathcal{O}\big((\frac{1}{\epsilon})^{\frac{2-\rho-\beta}{\rho}}\big)$.

When $C = \widetilde{C}\epsilon^{\beta/\rho} = \Theta(\epsilon^{\beta/\rho})$*, in order to make* $\Delta_t \leq \epsilon$*, taking logarithm we have*

$$\log\big(\frac{(2-\beta-\rho)\widetilde{C}\epsilon^{\beta/\rho}t}{\rho}\big) = \log\big((1/\epsilon)^{\frac{2-\beta-\rho}{\rho}}\big).$$

Re-arrange above equality, we have $T = \mathcal{O}\big((\frac{1}{\epsilon})^{\frac{2-\rho-\beta}{\rho}} + (\frac{1}{\epsilon})^{\frac{\beta}{\rho}}\big)$*. Thus, when* $0 \leq \rho \leq 2-2\beta$*, we have* $T = \mathcal{O}(\frac{1}{\epsilon})^{\frac{2-\rho-\beta}{\rho}}$*; when* $2-2\beta < \rho \leq 2-\beta$*, we have* $T = \mathcal{O}((\frac{1}{\epsilon})^{\frac{\beta}{\rho}})$.

Case II: When $\rho = 2-\beta$*, It is equivalent to claim* β, ρ *satisfies* $\frac{2-\beta}{\rho} = 1$*, descent inequality equation 31 reduces to*

$$\Delta_{t+1} \leq \Delta_t - \frac{\gamma\mu}{2}\Delta_t = (1-\frac{\gamma\mu}{2})\Delta_t,$$

As long as $\mu < \frac{2}{\gamma}$*, the* Δ_t *converges to* 0.

$$\Delta_t \leq (1-\frac{\gamma\mu}{2})^t\Delta_0 = \mathcal{O}\Big((1-\frac{\gamma\mu}{2})^t\Big).$$

However, since the step-size rule of γ *includes target accuracy* ϵ*. The convergence rate is not a standard linear convergence. To obtain a* ϵ*-stationary point, we have*

$$\Delta_t \leq (1-\frac{\gamma\mu}{2})^t\Delta_0 \leq \exp(-\frac{\gamma\mu t}{2})\Delta_0 \leq \epsilon, \tag{34}$$

which gives us iteration complexity

$$T = \frac{2}{\gamma\mu}\log(\frac{\Delta_0}{\epsilon}) = \mathcal{O}\big((\frac{1}{\epsilon})^{\frac{\beta}{\rho}}\log(\frac{1}{\epsilon})\big).$$

Case III: When $\rho > 2-\beta$ *This case is equivalent to* $\frac{2-\beta}{\rho} < 1$*. For simplicity, denote* $C = \frac{(2\mu)^{(2-\beta)/\rho}}{4}$ *and* $\omega = \frac{\rho}{2-\beta}$*. The sequence generated by recursion equation 31 is guaranteed to converge to 0 when* $\epsilon \downarrow 0$.

For simplicity, rewriting equation 31 as $\Delta_{t+1} \leq \Delta_t - C\gamma\Delta_t^{1/\omega}$*. Notice* $\Delta_t \geq 0$*,* $C > 0$*,* $\{\Delta\}_t$ *is non-increasing. Now suppose the sequence* $\{\Delta\}_t$ *converge to a positive constant, denoted as* D*. There must exists* $0 < \tilde{\varepsilon} < D$ *such that* $\Delta_t > \tilde{\varepsilon}$ *for all* t*. Then we have*

$$\Delta_{t+1} \leq \Delta_t - C\gamma\Delta_t^{\frac{1}{\omega}} \leq \Delta_t - C\gamma\tilde{\varepsilon}^{\frac{1}{\omega}}.$$

Re-organize above recursion, we have $TC\gamma\tilde{\varepsilon}^{1/\omega} \leq \sum_{t=0}^{T-1}\Delta_t - \Delta_{t+1} \leq \Delta_0$*, which is equivalent as* $T \leq \frac{\Delta_0}{C\gamma\tilde{\varepsilon}^{1/\omega}} < \infty$*. This fact contradicts to* $\Delta > \tilde{\varepsilon}$ *for arbitrary* t*. In conclusion, as long as equation 31 holds, the sequence* $\{\Delta_t\}$ *converges to 0 as* $\epsilon \downarrow 0$.

Next, we determine the local convergence rate. When Δ_t *is small enough,* $\Delta_{t+1}^{1/\omega}$ *will dominate* Δ_{t+1} *order-wisely since* $1/\omega < 1$*. This leads to refined recursion*

$$C\gamma\Delta_{t+1}^{\frac{1}{\omega}} \leq \Delta_{t+1} + C\gamma\Delta_{t+1}^{\frac{1}{\omega}} \leq \Delta_{t+1} + C\gamma\Delta_t^{\frac{1}{\omega}} \leq \Delta_t.$$

The first inequality is due to non-negativity of Δ_{t+1}*, the second inequality is due to* $\Delta_{t+1} \leq \Delta_t$*, the third inequality is a re-organization of equation 31. Denote* $T_0 = \inf\{t \in \mathbf{N} | \Delta_t/(C\gamma)^{\omega/\omega-1} < 1\}$*, then we have*

$$\begin{aligned}\Delta_{t+1} \leq (C\gamma)^{-\omega}\Delta_t^{\omega} &= (C\gamma)^{-\omega-\omega^2-\ldots-\omega^{t-T_0}}\Delta_{T_0}^{\omega^{t-T_0}} \\ &= (C\gamma)^{\frac{\omega(1-\omega^{t-T_0})}{\omega-1}}\Delta_{T_0}^{\omega^{t-T_0}} \\ &= (C\gamma)^{\omega/\omega-1}\big((C\gamma)^{\omega/\omega-1}\big)^{\omega^{-t-T_0}}\Delta_{T_0}^{\omega^{t-T_0}}.\end{aligned} \tag{35}$$

Since $(C\gamma)^{\omega/\omega-1}$ *only effects order of convergence up to a constant. To simplify analysis, denote* $\hat{C} = (\frac{C(2\mu)^{\beta/\rho}}{8(L_0+L_1)+1})^{\omega/\omega-1}$ *and then we have* $(C\gamma)^{\omega/\omega-1} = \hat{C}\epsilon^{\beta/\rho+\beta-2} \leq \hat{C}$. *since* $0 \leq \epsilon \leq \min\{1, 1/2\mu\}$, *we further reduce the recursion to*

$$\begin{aligned}\Delta_{t+1} &\leq (C\gamma)^{\omega/\omega-1}\big((C\gamma)^{\omega/\omega-1}\big)^{\omega^{t-T_0}}\Delta_{T_0}^{\omega^{t-T_0}} \\ &\leq \hat{C}\Big((C\gamma)^{\frac{\omega}{\omega-1}}\Big)^{-\omega^{t-T_0}}\Delta_{T_0}^{\omega^{t-T_0}} \\ &= \mathcal{O}\bigg(\Big(\frac{\Delta_{T_0}}{\gamma^{\omega/\omega-1}}\Big)^{\omega^{t-T_0}}\bigg). \end{aligned} \tag{36}$$

Taking logarithm and multiply negative sign on both sides of equation 36. We have

$$\log(\frac{\hat{C}}{\epsilon}) = \omega^{t-T_0} \cdot \log\big(\frac{(C\gamma)^{\frac{\omega}{\omega-1}}}{\Delta_{T_0}}\big).$$

Now, extract $\epsilon^{\beta/(\rho+\beta-2)}$ *from* $(C\gamma)^{\omega/\omega-1}/\Delta_{T_0}$. *We have*

$$\begin{aligned}\log\bigg((C\gamma)^{\frac{\omega}{\omega-1}}/\Delta_{T_0}\bigg) &= \log\bigg(\big(\hat{C}/\Delta_{T_0}\big)\cdot\epsilon^{\frac{\beta}{\rho+\beta-2}}\bigg) \\ &\leq \big(\hat{C}/\Delta_{T_0}\big)\cdot\epsilon^{\frac{\beta}{\rho+\beta-2}},\end{aligned}$$

where the last inequality is due to the fact $\log(x) \leq x, \forall x > 0$. *Taking logarithm again, we have*

$$t - T_0 = \Omega\big(\log\big((\frac{1}{\epsilon})^{\frac{\beta}{\rho+\beta-2}}\big)\big).$$

F Proof of Proposition 3

Proposition 3 (Convergence of GD) *Let Assumptions 1 and 2 hold. Assume there exists a positive constant* G *such that* $||\nabla f(x_t)|| \leq G, \forall t \in T$, *When* $\rho = 1$ *and setting* $\alpha = \beta = 1$, *by selecting* $\gamma \leq \min\{\frac{1}{L_0}, \frac{1}{2L_1G}\}$, *gradient descent converges to an* ϵ*-stationary point within* $T = \Omega(\frac{G}{\epsilon})$ *iterations.*

Proof 5 *When* $\alpha = 1$, *putting the update rule of gradient descent* $w_{t+1} = w_t - \gamma\nabla f(w_t)$ *into descent lemma equation 1 yields*

$$\begin{aligned}&f(w_{t+1}) \\ &\overset{(i)}{\leq} f(w_t) - \gamma\big\|\nabla f(w_t)\big\|^2 + \frac{\gamma^2}{2}(L_0 + L_1\big\|\nabla f(w_t)\big\|)\big\|\nabla f(w_t)\big\|^2 \\ &= f(w_t) - (\gamma - \frac{L_0\gamma^2}{2})\big\|\nabla f(w_t)\big\|^2 + \frac{L_1\gamma^2}{2}\big\|\nabla f(w_t)\big\|^3 \\ &\overset{(ii)}{\leq} f(w_t) - \frac{\gamma}{2}\big\|\nabla f(w_t)\big\|^2 + \frac{L_1\gamma^2}{2}\big\|\nabla f(w_t)\big\|^3 \\ &\overset{(iii)}{\leq} f(w_t) - \frac{\gamma}{4}\big\|\nabla f(w_t)\big\|^2 \end{aligned} \tag{37}$$

where (i) is due to descent lemma equation 1;(ii) and (iii) are due to the learning rate design $\gamma \leq \min\{\frac{1}{L_0}, \frac{1}{2L_1G}\}$, *which ensures* $-(\gamma - \frac{L_0\gamma^2}{2}) \leq -\frac{\gamma}{2}$ *and* $\frac{L_1\gamma^2}{2}||\nabla f(w_t)||^3 \leq \frac{\gamma}{4}||\nabla f(w_t)||^2$. *When applying assumption 2 with* $\rho = 1$ *and denote* $f(w_t) - f^*$ *as* Δ_t, *we have*

$$\Delta_{t+1} \leq \Delta_t - \frac{\gamma\mu}{2}\Delta_t^2 \tag{38}$$

The rest of proof is exactly the same as proof for Theorem 1 Case I, we omit discussion here. As a result, one can show equation 38 converges to a ϵ*-stationary point after* $\mathcal{O}(\frac{G}{\epsilon})$ *iterations.*

G Proof of Theorem 2

Theorem 2 (Convergence of IAN-SGD) *Let Assumptions 1, 3 and 4 hold. For the IAN-SGD 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* $\frac{1}{2}$*, IAN-SGD produces a sequence satisfying* $\min_{t\le T} ||\nabla F(w_t)|| \le \epsilon$ *if the total number of iteration* T *satisfies*

$$T \ge \Lambda \max\Big\{\frac{256\Lambda}{\epsilon^4}, \frac{640L_1}{\epsilon^{2-\beta}}, \frac{64(L_0+L_1) + 128L_1(3\tau_2/\tau_1)^\beta}{\epsilon^2}\Big\}. \tag{17}$$

Proof 6 *Start from descent lemma equation 3 and put the update rule of IAN-SGD, equation 14 in, we have*

$$\begin{aligned}
&F\big(w_{t+1}\big) - F\big(w_t\big)\\
\le&\nabla F\big(w_t\big)^\top\big(w_{t+1} - w_t\big) + \frac{1}{2}\big(L_0 + L_1\big\|\nabla F\big(w_t\big)\big\|^\alpha\big)\big\|w_{t+1} - w_t\big\|^2\\
=&-\gamma\frac{\nabla F\big(w_t\big)^\top\nabla f_{\xi_B}(w_t)}{h_t^\beta} + \frac{1}{2}\gamma^2\big(L_0 + L_1\big\|\nabla F\big(w_t\big)\big\|^\alpha\big)\frac{\big\|\nabla f_{\xi_B}(w_t)\big\|^2}{h_t^{2\beta}}.
\end{aligned} \tag{39}$$

Since the update rule using IAN-SGD formulates a random trajectory in terms of w_t*, taking expectation over* ξ_B *and* w_t*, using condition expectation rule, we have*

$$\begin{aligned}
&\mathbb{E}_{w_t}\big[\big[\mathbb{E}_{\xi_B}\big[F(w_{t+1}) - F(w_t)|w_t\big]\big]\\
\le&\ \mathbb{E}_{w_t}\Big[\frac{-\gamma\mathbb{E}_{\xi_B}[\big\|\nabla F(w_t)\big\|^2|w_t]}{h_t^\beta} + \frac{1}{2}\gamma^2(L_0 + L_1\big\|\nabla F(w_t)\big\|^\alpha)\frac{\mathbb{E}_{\xi_B}\big[\big\|\nabla f_{\xi_B}(w_t)\big\|^2|w_t\big]}{h_t^{2\beta}}\Big].
\end{aligned}$$

When the expectation is conditioned on w_t*, we can simplify* $\mathbb{E}_{\xi_B}\big[||\nabla F(w_t)||^2|w_t\big]$ *into* $||\nabla F(w_t)||^2$ *since* $\nabla F(w_t)$ *is deterministic over* ξ_B*. Additionally, by remarks induced by assumption 4, when conditioned over* w_t*, randomness only comes from* $\xi_{B'}$*, thus we have*

$$\mathbb{E}_{\xi_B}\big[\big\|\nabla f_{\xi_B}(w_t)\big\|^2|w_t\big] \le \underbrace{(\frac{2\tau_1^2}{B} + 1)\big\|\nabla F(w_t)\big\|^2}_{\textit{See equation 51}} + \frac{2\tau_2^2}{B}.$$

Let $B = 2\tau_1^2$*, above inequality reduces to*

$$\mathbb{E}_{\xi_B}\big[\big\|\nabla f_{\xi_B}(w_t)\big\|^2|w_t\big] \le 2\big\|\nabla F(w)\big\|^2 + \frac{\tau_2^2}{\tau_1^2}. \tag{40}$$

Put equation 40 into above descent lemma, we have

$$\begin{aligned}
&\mathbb{E}_{w_t}\big[\mathbb{E}_{\xi_B}\big[F(w_{t+1}) - F(w_t)|w_t\big]\big]\\
\le&\ \mathbb{E}_{w_t}\Big[-\gamma\frac{\big\|\nabla F(w_t)\big\|^2}{h_t^\beta} + \frac{1}{2}\gamma^2(L_0 + L_1\big\|\nabla F(w_t)\big\|^\alpha)\frac{\mathbb{E}_{\xi_B}\big[\big\|\nabla f_{\xi_B}(w_t)\big\|^2|w_t\big]}{h_t^{2\beta}}\Big]\\
\le&\ \mathbb{E}_{w_t}\Big[-\gamma\frac{\big\|\nabla F(w_t)\big\|^2}{h_t^\beta} + \frac{1}{2}\gamma^2(L_0 + L_1\big\|\nabla F(w_t)\big\|^\alpha)\frac{2\big\|\nabla F(w_t)\big\|^2 + \tau_2^2/\tau_1^2}{h_t^{2\beta}}\Big].\\
=&\ \mathbb{E}_{w_t}\Big[\Big(\frac{\gamma}{h_t^\beta}\big(-1 + \gamma\frac{L_0 + L_1\big\|\nabla F(w_t)\big\|^\alpha}{h_t^\beta}\big)\Big)\big\|\nabla F(w_t)\big\|^2 + \frac{1}{2}\gamma^2\frac{L_0 + L_1\big\|\nabla F(w_t)\big\|^\alpha}{h_t^{2\beta}}\frac{\tau_2^2}{\tau_1^2}\Big].
\end{aligned} \tag{41}$$

By clipping structure and step size rule, from where we know $\frac{1}{h_t^\beta} = \min\big\{1, \frac{1}{4L_1\gamma(2||\nabla f_{\xi_{B'}}(w_t)|| + \frac{\tau_2}{\tau_1})^\beta}\big\} < 1$ *and* $\gamma \le \frac{1}{4L_0}$*, we have*

$$\frac{\gamma L_0}{h_t^\beta} < \gamma L_0 \le \frac{1}{4}, \tag{42}$$

$$\frac{\gamma L_1\big\|\nabla F(w_t)\big\|^\alpha}{h_t^\beta} \le \frac{1}{4}. \tag{43}$$

The last inequality in equation 43 utilizes lemma 2, from where we know

$$\begin{aligned}
\frac{1}{4}h_t^{\beta} &\overset{(i)}{\geq} \frac{1}{4}h_t^{\alpha} \\
&\overset{(ii)}{=} \frac{1}{4}(h_t^{\beta})^{\frac{\alpha}{\beta}} \\
&\overset{(iii)}{\geq} \frac{1}{4}(4\gamma L_1)^{\frac{\alpha}{\beta}}\big(2\big\|\nabla f_{\xi_{B'}}(w_t)\big\| + \frac{\tau_2}{\tau_1}\big)^{\alpha} \\
&\overset{(iv)}{\geq} \frac{1}{4}(4\gamma L_1)\big(2\big\|\nabla f_{\xi_{B'}}(w_t)\big\| + \frac{\tau_2}{\tau_1}\big)^{\alpha} \\
&\overset{(v)}{\geq} \gamma L_1\big\|\nabla F(w_t)\big\|^{\alpha},
\end{aligned} \tag{44}$$

where (i) utilizes the fact $h_t \geq 1$ and $\beta \geq \alpha$; (iii) utilize the fact that $h_t^{\beta} \geq 4L_1\gamma(2\big\|\nabla f_{\xi_{B'}}(w_t)\big\| + \frac{\tau_2}{\tau_1})^{\beta}$; (iv) utilizes the fact that $\gamma \leq \frac{1}{4L_1}$, thus $(4\gamma L_1) \leq (4\gamma L_1)^{\alpha/\beta}$ since $\beta \in [\alpha, 1]$; (v) utilizes the fact stated fact in Lemma 2. Combining equation 43, above descent lemma further reduces to

$$\begin{aligned}
&\mathbb{E}_{w_t}\Big[\mathbb{E}_{\xi_B}\big[F(w_{t+1}) - F(w_t)|w_t\big]\Big] \\
&\leq \mathbb{E}_{w_t}\Big[-\frac{\gamma}{2h_t^{\beta}}\big\|\nabla F(w_t)\big\|^2 + \underbrace{\frac{1}{2}\gamma^2\frac{L_0 + L_1\big\|\nabla F(w_t)\big\|^{\alpha}}{h_t^{2\beta}}\frac{\tau_2^2}{\tau_1^2}}_{\textit{Term 1, See Lemma 6}}\Big] \\
&\leq \mathbb{E}_{w_t}\Big[-\frac{\gamma}{2h_t^{\beta}}\big\|\nabla F(w_t)\big\|^2 + \frac{1}{2}\gamma^2(L_0 + L_1)(1 + \frac{\tau_2^2}{\tau_1^2})^2 + \frac{\gamma}{4h_t^{\beta}}\big\|\nabla F(w_t)\big\|^2\Big] \\
&= \mathbb{E}_{w_t}\Big[-\frac{\gamma}{4h_t^{\beta}}\big\|\nabla F(w_t)\big\|^2 + \frac{1}{2}\gamma^2(L_0 + L_1)(1 + \frac{\tau_2^2}{\tau_1^2})^2\Big],
\end{aligned} \tag{45}$$

where the last inequality utilize the inequality stated in Lemma 6, equation equation 53.
Re-organize the inequality by putting the negative term to LHS, we have

$$\mathbb{E}_{w_t}\Big[\frac{\gamma}{4h_t^{\beta}}\big\|\nabla F(w_t)\big\|^2\Big] \leq \mathbb{E}_{w_t}\Big[\mathbb{E}_{\xi_B}\big[F(w_t) - F(w_{t+1})|w_t\big]\Big] + \frac{1}{2}(L_0 + L_1)\gamma^2(1 + \frac{\tau_2^2}{\tau_1^2})^2. \quad \forall t \in [T] \tag{46}$$

In order to express the LHS into a more tractable form, we want to express $\frac{||\nabla F(w_t)||^2}{h_t^{\beta}}$ explicitly in a simpler form. Using the fact $\frac{1}{(a+b)^{\beta}} \geq \min\{\frac{1}{(2a)^{\beta}}, \frac{1}{(2b)^{\beta}}\}$. We have

$$\begin{aligned}
&\gamma\frac{\big\|\nabla F(w_t)\big\|^2}{h_t^{\beta}} \\
&\overset{(i)}{=} \gamma\min\Big\{1, \frac{1}{4L_1\gamma(2\big\|\nabla f_{\xi_{B'}}(w_t)\big\| + \frac{\tau_2}{\tau_1})^{\beta}}\Big\}\big\|\nabla F(w_t)\big\|^2 \\
&\overset{(ii)}{\geq} \gamma\min\Big\{1, \frac{1}{4L_1\gamma\big(\frac{5}{2}\big\|\nabla F(w_t)\big\| + \frac{3\tau_2}{2\tau_1}\big)^{\beta}}\Big\}\big\|\nabla F(w_t)\big\|^2 \\
&\overset{(iii)}{\geq} \gamma\min\Big\{1, \frac{1}{4L_1\gamma\big(5\big\|\nabla F(w_t)\big\|\big)^{\beta}}, \frac{1}{4L_1\gamma\big(\frac{3\tau_2}{\tau_1}\big)^{\beta}}\Big\}\big\|\nabla F(w_t)\big\|^2 \\
&\overset{(iv)}{=} \min\Big\{\gamma, \frac{1}{4L_1\big(5\big\|\nabla F(w_t)\big\|\big)^{\beta}}, \frac{1}{4L_1\big(\frac{3\tau_2}{\tau_1}\big)^{\beta}}\Big\}\big\|\nabla F(w_t)\big\|^2 \\
&\overset{(v)}{=} \min\Big\{\gamma, \frac{1}{4L_1\big(5\big\|\nabla F(w_t)\big\|\big)^{\beta}}\Big\}\big\|\nabla F(w_t)\big\|^2 \\
&\overset{(vi)}{\geq} \min\Big\{\gamma\big\|\nabla F(w_t)\big\|^2, \frac{\big\|\nabla F(w_t)\big\|^{2-\beta}}{20L_1}\Big\},
\end{aligned} \tag{47}$$

where (i) expands the expression of $\frac{1}{h_t^\beta}$*; (ii) utilizes the equation 50 to upper bounds* $||\nabla f_{\xi_{B'}}(w_t)||$ *by* $(\tau_1/\sqrt{16\tau_1^2}+1)||\nabla F(w_t)||+\tau_2/\sqrt{16\tau_1^2}$ *by setting* $B'=16\tau_1^2$*; (iii) utilizes the fact* $\frac{1}{a+b}\geq\min\{\frac{1}{2a},\frac{1}{2b}\}$ *where* $a=\frac{5}{2}\left\|\nabla F(w_t)\right\|$, $b=\frac{3\tau_2}{\tau_1}$*; (iv) puts* γ *inside the minimum operator. From step size rule,* $\gamma\leq\frac{1}{8L_1(3\tau_2/\tau_1)^\beta}$*, we can directly delete the third term* $\frac{1}{4L_1(3\tau_2/\tau_1)^\beta}$*, which reduces expressions in (v); (vi) further replaces* 5^β *by* 5 *in denominator.*

Since now equation 47 has no randomness induced from $\xi_{B'}$*. Summing the above descent lemma from* 0 *to* $T-1$*, we have*

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}_{w_t}\Big[\min\Big\{\frac{\gamma}{4}\left\|\nabla F(w_t)\right\|^2,\frac{\left\|\nabla F(w_t)\right\|^{2-\beta}}{80L_1}\Big\}\Big]\\
&\leq\sum_{t=0}^{T-1}\mathbb{E}_{w_t}\Big[\frac{\gamma}{4h_t^\beta}\left\|\nabla F(w_t)\right\|^2\Big]\\
&\leq\sum_{i=1}^{T}\mathbb{E}_{w_t}\Big[\mathbb{E}_{\xi_B}\left[F(w_t)-F(w_{t+1})|w_t\right]\Big]+T\frac{1}{2}\gamma^2(L_0+L_1)(1+\frac{\tau_2^2}{\tau_1^2})^2.
\end{aligned}$$

By step size rule, from where we know $\gamma\leq\frac{1}{\sqrt{T}}$*, we have*

$$\sum_{t=0}^{T-1}\mathbb{E}_{w_t}\Big[\min\Big\{\frac{\gamma}{4}\left\|\nabla F(w_t)\right\|^2,\frac{\left\|\nabla F(w_t)\right\|^{2-\beta}}{80L_1}\Big\}\Big]\leq F(w_0)-F^*+\frac{1}{2}(L_0+L_1)(1+\frac{\tau_2^2}{\tau_1^2})^2.$$

Denote $K=\{t|t\in[T]$ *such that* $\gamma\left\|\nabla F(w_t)\right\|^2\leq\frac{||\nabla F(w_t)||^{2-\beta}}{20L_1}\}$ *, then above descent lemma can be reduced to*

$$\sum_{t\in K}\mathbb{E}_{w_t}\Big[\frac{\gamma}{4}\left\|\nabla F(w_t)\right\|^2\Big]\leq F(w_0)-F^*+\frac{1}{2}(L_0+L_1)(1+\frac{\tau_2^2}{\tau_1^2})^2,$$

and

$$\sum_{t\in K^c}\mathbb{E}_{w_t}\Big[\frac{\left\|\nabla F(w_t)\right\|^{2-\beta}}{80L_1}\Big]\leq F(w_0)-F^*+\frac{1}{2}(L_0+L_1)(1+\frac{\tau_2^2}{\tau_1^2})^2.$$

Now denote RHS by $\Lambda=F(w_0)-F^*+\frac{1}{2}(L_0+L_1)(1+\frac{\tau_2^2}{\tau_1^2})^2$*, then we have*

$$\begin{aligned}
&\mathbb{E}_{w_t}\Big[\min_{t\in T}\left\|\nabla F(w_t)\right\|\Big]\\
&\leq\mathbb{E}_{w_t}\Big[\min\Big\{\frac{1}{|K|}\sum_{t\in K}\left\|\nabla F(w_t)\right\|,\frac{1}{|K^c|}\sum_{t\in K^c}\left\|\nabla F(w_t)\right\|\Big\}\Big]\\
&\overset{(i)}{\leq}\mathbb{E}_{w_t}\Big[\min\Big\{\sqrt{\frac{1}{|K|}\sum_{i=1}^{|K|}\left\|\nabla F(w_t)\right\|^2},\Big(\frac{1}{|K^c|}\sum_{i=1}^{|K^c|}\left\|\nabla F(w_t)\right\|^{2-\beta}\Big)^{\frac{1}{2-\beta}}\Big\}\Big]\\
&\overset{(ii)}{\leq}\max\Big\{\sqrt{(4\Lambda)\frac{4(L_0+L_1)+\sqrt{T}+8L_1(3\tau_2/\tau_1)^\beta}{T}},\big(\Lambda\frac{160L_1}{T}\big)^{\frac{1}{2-\beta}}\Big\},
\end{aligned}$$

where (i) comes from the concavity $y^{\frac{1}{2}}$ *and* $y^{\frac{1}{2-\beta}}$ *and inverse Jensen's inequality for concave function, and the last inequality follows from descent lemma as well as either* $K>\frac{T}{2}$ *or* $K^c>\frac{T}{2}$*. This implies, in order to find a point satisfies*

$$Pr(\min_{t\in[T]}||\nabla F(w_t)||\geq\epsilon)\leq\frac{1}{2}.$$

By Markov inequality, we must have $\mathbb{E}_{w_t}[\min_{t\in[T]}||\nabla F(w_t)||]\leq\frac{\epsilon}{2}$ *when* T *satisfies*

$$T\geq\Lambda\max\{\frac{256\Lambda}{\epsilon^4},\frac{640L_1}{\epsilon^{2-\beta}},\frac{64(L_0+L_1)+128L_1(3\tau_2/\tau_1)^\beta}{\epsilon^2}\}. \tag{48}$$

H Proof of Lemma 2

Before proving Lemma 2, let us proof the technical lemma to determine the upper bound of mini-batch stochastic gradient estimators given assumption 4

Lemma 5 *For mini-batch stochastic gradient estimator satisfying assumption 4, denote $\delta_B(w)$ as the approximation error $\delta_B(w) = \frac{1}{B}\sum_{i=1}^{B} \nabla f_i(w) - \nabla F(w)$, we have the upper bound*

$$\|\delta_B(w)\| \leq \frac{1}{\sqrt{B}}(\tau_1\|\nabla F(w)\| + \tau_2). \tag{49}$$

Proof 7 *The proof follows from applying Jensen's inequality for L2 norm.*

$$\begin{aligned}
\|\delta_B(w)\| &= \|\frac{1}{B}\sum_{i=1}^{B}\delta_i(w)\| \\
&= \frac{1}{B}\Big(\|\sum_{i=1}^{B}\delta_i(w)\|_2^2\Big)^{\frac{1}{2}} \\
&\leq \frac{1}{B}\Big(\sum_{i=1}^{B}\|\delta_i(w)\|_2^2\Big)^{\frac{1}{2}} \\
&\leq \frac{1}{B}\Big(\sum_{i=1}^{B}(\tau_1\|\nabla F(w)\| + \tau_2)^2\Big)^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{B}}(\tau_1\|\nabla F(w)\| + \tau_2).
\end{aligned}$$

where the first inequality uses Jensen's inequality and convexity of squared L2-norm; the second inequality uses the assumption equation 15.

This fact leads to

$$\|\nabla f_{\xi_B}(w)\| \leq (\frac{\tau_1}{\sqrt{B}} + 1)\|\nabla F(w)\| + \frac{\tau_2}{\sqrt{B}}. \tag{50}$$

Similarly, for variance of $\delta_B(w)$, we have the remark stated as following.

Remark 3 (Variance bound for mini-batch $\delta_B(w)$) *For variance of $\delta(w)$, following the same logic above, we have*

$$\begin{aligned}
Var(\|\delta_B(w)\|) &= \mathbb{E}\|\delta_B(w)\|^2 \\
&= \mathbb{E}\Big((\frac{1}{B}\sum_{i=1}^{B}\delta_i(w))^T(\frac{1}{B}\sum_{i=1}^{B}\delta_j(w))\Big) \\
&= \frac{1}{B^2}\mathbb{E}\Big[\sum_{i=1}^{B}\|\delta_i(w)\|^2\Big] \\
&\leq \frac{1}{B}(\tau_1\|\nabla F(w)\| + \tau_2)^2 \\
&\leq \frac{1}{B}(2\tau_1^2\|\nabla F(w)\| + 2\tau_2^2),
\end{aligned}$$

where the first inequality is due to equation 15. Thus, it is equivalent as

$$\mathbb{E}_{\xi_B}\left[\|\nabla f_{\xi_B}(w)\|^2\right] \leq (\frac{2\tau_1^2}{B} + 1)\|\nabla F(w)\|^2 + \frac{2\tau_2^2}{B}. \tag{51}$$

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*

$$\big\|\nabla f_{\xi_B}(w)\big\| \geq \frac{1}{2}\big\|\nabla F(w)\big\| - \frac{\tau_2}{2\tau_1}. \tag{16}$$

Proof 8 (Proof of Lemma 2) *When $\big\|\nabla F(w)\big\|$ is large such that $\big\|\nabla F(w)\big\| \geq \frac{\tau_2}{\tau_1}$, then equation 49 indicates*

$$\big\|\delta_B(w)\big\| \leq \frac{2\tau_1\big\|\nabla F(w)\big\|}{\sqrt{B}}.$$

In this case, if we choose $B = 16\tau_1^2$, we have $\big\|\delta_B(w)\big\| \leq \frac{1}{2}\big\|\nabla F(w)\big\|$. Since in this case, we assume, $\big\|\nabla F(w)\big\| \geq \frac{\tau_2}{\tau_1} \geq \frac{\tau_2}{2\tau_1}$, we have

$$\Big|\big\|\nabla f_{\xi_B}(w_t)\big\| - \big\|\nabla F(w)\big\|\Big| \leq \frac{1}{2}\big\|F(w)\big\|,$$

which is equivalent as

$$\big\|\nabla f_{\xi_B}(w)\big\| - \big\|\nabla F(w)\big\| \geq -\frac{1}{2}\big\|\nabla F(w)\big\|.$$

And this fact leads to

$$\frac{1}{2}\big\|\nabla F(w)\big\| \leq \big\|\nabla f_{\xi_B}(w)\big\| \leq \big\|\nabla f_{\xi_B}(w)\big\| + \frac{\tau_2}{2\tau_1}.$$

Re-organize the term gives us

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

Similarly, when $\big\|\nabla F(w)\big\| \leq \frac{\tau_2}{\tau_1}$, for single stochastic sample, by assumption 4, we have $\big\|\delta(w)\big\| \leq 2\tau_2$, from equation 49, for mini-batch stochastic gradient estimator, we have

$$\big\|\delta_B(w)\big\| \leq \frac{2\tau_2}{\sqrt{B}}.$$

By setting $B = 16\tau_1^2$, we have $\big\|\delta_B(w)\big\| \leq \frac{\tau_2}{2\tau_1}$. This fact leads to

$$\Big|\big\|\nabla f_{\xi_B}(w)\big\| - \big\|\nabla F(w)\big\|\Big| \leq \frac{\tau_2}{2\tau_1},$$

which is equivalent as

$$\big\|\nabla f_{\xi_B}(w)\big\| - \big\|\nabla F(w)\big\| \geq -\frac{\tau_2}{2\tau_1}.$$

Thus, we have

$$\begin{aligned}\frac{1}{2}\big\|\nabla F(w)\big\| &\leq \big\|\nabla F(w)\big\| \\ &= \big\|\nabla F(w)\big\| + \frac{\tau_2}{2\tau_1} - \frac{\tau_2}{2\tau_1} \\ &\leq \big\|\nabla f_{\xi_B}(w)\big\| + \frac{\tau_2}{2\tau_1},\end{aligned}$$

which leads to

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

Combine above, we conclude by choosing $B = 16\tau_1^2$, we always have

$$\big\|\nabla f_{\xi_B}(w)\big\| \geq \frac{1}{2}\big\|\nabla F(w)\big\| - \frac{\tau_2}{2\tau_1}. \tag{52}$$

I Lemma 6 and proof

Lemma 6 *For the "Term 1" defined in equation 45, we have upper bound*

$$\frac{1}{2}\gamma^2\frac{(L_0+L_1\|\nabla F(w_t)\|^{\alpha})}{h_t^{2\beta}}\frac{\tau_2^2}{\tau_1^2}\leq\frac{1}{2}\gamma^2(L_0+L_1)(1+\frac{\tau_2^2}{\tau_1^2})^2+\frac{\gamma}{4h_t^{\beta}}\|\nabla F(w_t)\|^2. \tag{53}$$

Proof 9 *When $\|\nabla F(w_t)\|\leq\sqrt{1+\tau_2^2/\tau_1^2}$, we have $\|\nabla F(w_t)\|^{\alpha}\leq(1+\tau_2^2/\tau_1^2)^{\frac{\alpha}{2}}$ for any $\alpha>0$. Since $(1+\tau_2^2/\tau_1^2)>1$ and $(1+\tau_2^2/\tau_1^2)>\tau_2^2/\tau_1^2$. These facts lead to*

$$\begin{aligned}
&\frac{1}{2}\gamma^2\frac{(L_0+L_1\|\nabla F(w_t)\|^{\alpha})}{h_t^{2\beta}}\frac{\tau_2^2}{\tau_1^2}\\
&\leq\frac{1}{2}\gamma^2(1+\frac{\tau_2^2}{\tau_1^2})^{\frac{\alpha}{2}}\frac{(L_0+L_1)}{h_t^{2\beta}}\frac{\tau_2^2}{\tau_1^2}\\
&\leq\frac{1}{2}\gamma^2(L_0+L_1)(1+\frac{\tau_2^2}{\tau_1^2})^{\frac{\alpha}{2}}(1+\frac{\tau_2^2}{\tau_1^2})\\
&\leq\frac{1}{2}\gamma^2(L_0+L_1)(1+\frac{\tau_2^2}{\tau_1^2})^2,
\end{aligned} \tag{54}$$

where the first inequality comes from $||\nabla F(w_t)||\leq\sqrt{1+\tau_2^2/\tau_1^2}$ and $1+\tau_2^2/\tau_1^2>1$; the second inequality comes from the fact that $\frac{1}{h_t}<1$, so does $\frac{1}{h_t^{2\beta}}$, and upper bound τ_2^2/τ_1^2 by $(1+\tau_2^2/\tau_1^2)$; the last inequality uses the fact that $0\leq\alpha\leq1$ and $(1+\tau_2^2/\tau_1^2)^{1+\alpha/2}\leq(1+\tau_2^2/\tau_1^2)^2$.

When $\|\nabla F(w_t)\|\geq\sqrt{1+\tau_2^2/\tau_1^2}$, we must have $\|\nabla F(w_t)\|^2\geq(1+\tau_2^2/\tau_1^2)\geq\tau_2^2/\tau_1^2$ for any $\alpha>0$, Thus, we conclude

$$\begin{aligned}
&\frac{\gamma^2}{2}\frac{L_1\|\nabla F(w_t)\|^{\alpha}}{h_t^{2\beta}}\cdot\frac{\tau_2^2}{\tau_1^2}\\
&=\frac{\gamma^2}{2}\frac{L_1\|\nabla F(w_t)\|^{\alpha}}{h_t^{\beta}}\frac{1}{h_t^{\beta}}\cdot\frac{\tau_2^2}{\tau_1^2}\\
&=\frac{\gamma^2}{2h_t^{\beta}}\frac{L_1\|\nabla F(w_t)\|^{\alpha}}{h_t^{\beta}}\cdot\frac{\tau_2^2}{\tau_1^2}\\
&\overset{(i)}{\leq}\frac{\gamma^2}{2h_t^{\beta}}\frac{L_1\|\nabla F(w_t)\|^{\alpha}}{h_t^{\beta}}\cdot\|\nabla F(w_t)\|^2\\
&\overset{(ii)}{\leq}\frac{\gamma^2}{2h_t^{\beta}}\frac{L_1\|\nabla F(w_t)\|^{\alpha}}{(4L_1\gamma)(2\|\nabla f_{\xi_{B'}}(w_t)\|+\frac{\tau_2}{\tau_1})^{\beta}}\cdot\|\nabla F(w_t)\|^2\\
&\overset{(iii)}{\leq}\frac{\gamma^2}{2h_t^{\beta}}\frac{L_1\|\nabla F(w_t)\|^{\alpha}}{(4L_1\gamma)\|\nabla F(w_t)\|^{\beta}}\|\nabla F(w_t)\|^2\\
&=\frac{\gamma^2}{2h_t^{\beta}}\frac{L_1}{4L_1\gamma\|\nabla F(w_t)\|^{\beta-\alpha}}\|\nabla F(w_t)\|^2\\
&\overset{(iv)}{\leq}\frac{\gamma^2}{2h_t^{\beta}}\frac{L_1}{4L_1\gamma}\|\nabla F(w_t)\|^2\\
&=\frac{\gamma}{8h_t^{\beta}}\|\nabla F(w_t)\|^2,
\end{aligned} \tag{55}$$

where (i) comes from the fact that $||\nabla F(w_t)||\geq\sqrt{1+\tau_2^2/\tau_1^2}$; (ii) comes from the fact $\frac{1}{h_t^{\beta}}\leq\frac{1}{4L_1\gamma(2||\nabla f_{\xi_{B'}}(w_t)||+\tau_2/\tau_1)^{\beta}}$; (iii) comes to the fact $||\nabla F(w_t)||\leq2||\nabla f_{\xi_{B'}}(w_t)||+\tau_2/\tau_1$; (iv) comes from the

fact that now $||\nabla F(w_t)|| > 1$*, thus* $\frac{1}{||\nabla F(w_t)||^{\beta-\alpha}} < 1$*. Similarly, when* $\left\|\nabla F(w_t)\right\| \geq \sqrt{1+\tau_2^2/\tau_1^2}$*, we can upper bound* $\frac{1}{2}\gamma^2 L_0 \frac{\tau_2^2}{\tau_1^2}$ *by*

$$\begin{aligned}
&\frac{1}{2h_t^{2\beta}}\gamma^2 L_0 \cdot \frac{\tau_2^2}{\tau_1^2} \\
&\leq \frac{\gamma}{2h_t^{\beta}}\gamma L_0 \left\|\nabla F(w_t)\right\|^2 \\
&\leq \frac{\gamma}{8h_t^{\beta}}\left\|\nabla F(w_t)\right\|^2,
\end{aligned} \tag{56}$$

where the first inequality uses the fact $\left\|\nabla F(w_t)\right\| \geq \sqrt{(1+\tau_2^2/\tau_1^2)}$ *and* $\frac{1}{h_t^{\beta}} \leq 1$*, second inequality uses the fact* $\gamma L_0 \leq \frac{1}{4}$*.*

Combine equation 54, equation 55, equation 56 give us desired result.