Generalized-Smooth Nonconvex Optimization is As Efficient As Smooth Nonconvex Optimization

Ziyi Chen[1] **Yi Zhou**[1] **Yingbin Liang**[2] **Zhaosong Lu**[3]

Abstract

Various optimal gradient-based algorithms have been developed for smooth nonconvex optimization. However, many nonconvex machine learning problems do not belong to the class of smooth functions and therefore the existing algorithms are sub-optimal. Instead, these problems have been shown to satisfy certain generalized-smooth conditions, which have not been well understood in the existing literature. In this paper, we propose a notion of α-symmetric generalized-smoothness that extends the existing notions and covers many important functions such as high-order polynomials and exponential functions. We study the fundamental properties and establish descent lemmas for the functions in this class. Then, to solve such a large class of nonconvex problems, we design a special deterministic normalized gradient descent algorithm that achieves the optimal iteration complexity $\mathcal{O}(\epsilon^{-2})$, and also prove that the popular SPIDER variance reduction algorithm achieves the optimal sample complexity $\mathcal{O}(\epsilon^{-3})$ in the stochastic setting. Our results show that solving generalized-smooth nonconvex problems is as efficient as solving smooth nonconvex problems.

1. Introduction

In many modern machine learning applications, training machine learning model requires solving a nonconvex optimization problem with big data, for which many efficient gradient-based optimization algorithms have been developed, e.g., gradient descent (GD) (Carmon et al., 2020), stochastic gradient descent (SGD) (Ghadimi and Lan, 2013) and many advanced stochastic variance reduction algorithms (Fang et al., 2018; Wang et al., 2019). In particular, the complexities of these algorithms have been extensively studied in nonconvex optimization. Specifically, under the standard assumption that the objective function is L-smooth (i.e., has Lipschitz continuous gradient), it has been shown that the basic GD algorithm (Carmon et al., 2020) and many advanced stochastic variance reduction algorithms (Fang et al., 2018; Cutkosky and Orabona, 2019) achieve the complexity lower bounds of finding an approximate stationary point of deterministic nonconvex optimization and stochastic nonconvex optimization, respectively. [1]

Although the class of smooth nonconvex problems can be effectively solved by the above provably optimal algorithms, it does not include many important modern machine learning applications, e.g., distributionally robust optimization (DRO) (Jin et al., 2021) and language model learning (Zhang et al., 2019), etc. Specifically, for the problems involved in these applications, they are not globally smooth but have been shown to satisfy certain generalized-smooth conditions, in which the smoothness parameters scale with the gradient norm in various ways (see the formal definitions in Section 2). To solve these generalized-smooth-type nonconvex problems, the existing works have developed various gradient-based algorithms, but only with sub-optimal complexity results for stochastic optimization. Therefore, we are motivated to *systematically build a comprehensive understanding of generalized-smooth functions and develop algorithms with improved complexities*.

To achieve this overarching goal, we need to address several fundamental challenges. First, the existing generalized-smooth conditions are proposed for specific application examples. Therefore, they define relatively restricted classes of functions that do not cover many popular ones such as high-order polynomials and exponential functions. Thus, we are motivated to consider the following question.

- *Q1: How to extend the existing notion of generalized-smoothness to cover a broad range of functions used in*

[1]Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, UT, US [2]Department of Electrical and Computer Engineering, Ohio State University, Columbus, OH, US [3]Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN, US. Correspondence to: Ziyi Chen <u1276972@utah.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] Deterministic and stochastic optimization problems are formulated respectively as $\min_w f(w)$ and $\min_w \mathbb{E}_{\xi\sim\mathbb{P}} f_\xi(w)$.

machine learning practice? What are the fundamental properties of the functions in this class?

Second, for such an extended class of generalized-smooth problems, it is expected that first-order algorithms may generally suffer from higher computation complexity (as compared to solving smooth problems). On the other hand, it is unclear how to design first-order algorithms that can efficiently solve these more challenging problems. Therefore, we aim to answer the following question.

- *Q2: Can first-order algorithms solve generalized-smooth nonconvex problems as efficiently as solving smooth nonconvex problems? In particular, what algorithms can achieve the optimal complexities?*

1.1. Our Contribution

In this paper, we provide comprehensive and affirmative answers to the aforementioned fundamental questions. Our contributions are summarized as follows.

- We propose a class of α-symmetric generalized-smooth functions, denoted by $\mathcal{L}^*_{\text{sym}}(\alpha)$, which we show strictly contains the popular class of L-smooth functions (i.e., functions with Lipschitz continuous gradient), the class of asymmetric generalized-smooth functions (Levy et al., 2020; Jin et al., 2021) and the class of Hessian-based generalized-smooth functions (Zhang et al., 2019) (see the definitions in Section 2). In particular, we show that our proposed function class $\mathcal{L}^*_{\text{sym}}(\alpha)$ includes a wide range of popular machine learning problems and functions used in practice, including distributionally robust optimization (Levy et al., 2020; Jin et al., 2021), objective function of language models (Zhang et al., 2019), high-order polynomials and exponential functions.

- We study the fundamental properties of functions in the class $\mathcal{L}^*_{\text{sym}}(\alpha)$ and establish new decent lemmas for functions in $\mathcal{L}^*_{\text{sym}}(\alpha)$ with different values of α (See Proposition 1). These technical tools play an important role later in designing new gradient-based algorithms and developing their corresponding convergence analysis.

- We develop a β-normalized gradient descent (named β-GD) algorithm for solving nonconvex problems in $\mathcal{L}^*_{\text{sym}}(\alpha)$, which normalizes the gradient $\nabla f(w_t)$ with the factor $\|\nabla f(w_t)\|^\beta$ in each iteration. We show that β-GD finds an approximate stationary point $\mathbb{E}\|\nabla f(w)\| \le \epsilon$ with iteration complexity $\mathcal{O}(\epsilon^{-2})$ as long as $\alpha \le \beta \le 1$, which matches the iteration complexity lower bound for deterministic smooth nonconvex optimization and hence is an optimal algorithm. On the other hand, we show that it may diverge when $0 < \beta < \alpha$ is used.

- For nonconvex stochastic optimization, we propose a class of expected α-symmetric generalized-smooth functions, denoted by $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$, which substantially generalizes the popular class of expected smooth functions. Interestingly, we prove that the original SPIDER algorithm still achieves the optimal sample complexity $\mathcal{O}(\epsilon^{-3})$ for solving nonconvex stochastic problems in $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$.

In summary, our work reveals that generalized-smooth nonconvex (stochastic) optimization is as efficient as smooth nonconvex (stochastic) optimization, and the optimal complexities can be achieved by β-GD (for deterministic case) and SPIDER (for stochastic case), respectively.

1.2. Related Work

L-smooth Functions $\mathcal{L}$: For deterministic nonconvex L-smooth problems, it is well-known that GD achieves the optimal iteration complexity $\mathcal{O}(\epsilon^{-2})$ (Carmon et al., 2020). For stochastic nonconvex problems $f(w) := \mathbb{E}_{\xi\sim\mathbb{P}} f_\xi(w)$, SGD achieves $\mathcal{O}(\epsilon^{-4})$ sample complexity (Ghadimi and Lan, 2013) which has been proved optimal for first-order stochastic algorithms if only the population loss f is L-smooth (Arjevani et al., 2022). (Fang et al., 2018) proposed the first variance reduction algorithm named SPIDER that achieves the optimal sample complexity $\mathcal{O}(\epsilon^{-3})$ under the stronger expected smoothness assumption (see eq. (17) for its definition). At the same time, several other variance reduction algorithms have been developed for stochastic nonconvex optimization that achieve the optimal sample complexity. For example, SARAH (Nguyen et al., 2017) and SpiderBoost (Wang et al., 2019) can be seen as unnormalized versions of SPIDER. STORM further improved the practical efficiency of these algorithms by using single-loop updates with adaptive learning rates (Cutkosky and Orabona, 2019). (Zhou et al., 2020) proposed the SNVRG algorithm by adjusting the SVRG variance reduction technique (Johnson and Zhang, 2013; Reddi et al., 2016) using multiple nested reference points, which also converge to a second-order stationary point.

Hessian-based Generalized-smooth Functions $\mathcal{L}^*_{\mathbf{H}}$: (Zhang et al., 2019) extended the L-smooth function class to a Hessian-based generalized-smooth function class $\mathcal{L}^*_{\text{H}}$ which allows the Lipschitz constant to linearly increase with the gradient norm (see Definition 2) and thus includes higher-order polynomials and many language models that are not L-smooth. For objective function on $\mathcal{L}^*_{\text{H}}$, (Zhang et al., 2019) also proposed clipped GD and normalized GD which keep the optimal iteration complexity $\mathcal{O}(\epsilon^{-2})$, and proposed clipped SGD which also achieves sample complexity $\mathcal{O}(\epsilon^{-4})$. (Zhang et al., 2020) proposed a general framework for clipped GD/SGD with momentum acceleration and obtained the same complexities for both deterministic and stochastic optimization. (Zhao et al., 2021) obtained sample complexity $\mathcal{O}(\epsilon^{-4})$ for normalized SGD with both small constant stepsize and diminishing stepsize. A contemporary work (Reisizadeh et al., 2023) reduced the sample complexity to $\mathcal{O}(\epsilon^{-3})$ by combining SPIDER variance reduction technique with gradient clipping.

Asymmetric Generalized-Smooth Functions $\mathcal{L}^*_{\text{asym}}$: Variants of clipped/normalized GD and SGD have been proposed on the asymmetric generalized-smooth function class $\mathcal{L}^*_{\text{asym}}$, which looks like a first-order variant of $\mathcal{L}^*_{\text{H}}$ (see Definition 1). For example, (Jin et al., 2021) applied mini-batch normalized SGD with momentum proposed by (Cutkosky and Mehta, 2020) to distributionally robust optimization problem which has been proved equivalent to minimizing a function in $\mathcal{L}^*_{\text{asym}}$ (Levy et al., 2020; Jin et al., 2021), and also obtained sample complexity $\mathcal{O}(\epsilon^{-4})$. (Yang et al., 2022) made normalized and clipped SGD differentially private by adding Gaussian noise. (Crawshaw et al., 2022) proposed generalized signSGD with ADAM-type normalization and obtained sample complexity $\mathcal{O}(\epsilon^{-4})$ on a smaller coordinate-wise version of $\mathcal{L}^*_{\text{asym}}$.

2. Existing Notions of Generalized-Smoothness

The class of L-smooth functions, which we denote as $\mathcal{L}$, includes all continuously differentiable functions with Lipschitz continuous gradient. Specifically, for any $f \in \mathcal{L}$, there exists $L_0 > 0$ such that

$$\|\nabla f(w') - \nabla f(w)\| \le L_0\|w' - w\|, \ \ \forall w, w' \in \mathbb{R}^d. \quad (1)$$

Many useful functions fall into this class, e.g., quadratic functions, logistic functions, etc. Nevertheless, $\mathcal{L}$ is a restricted function class that cannot efficiently model a broad class of functions, including higher-order polynomials, exponential functions, etc. For example, consider the one-dimensional polynomial function $f(x) = x^4$ in the range $x \in [-10, 10]$. According to (1), its smoothness parameter L_0 can be as large as 1200, leading to an ill-conditioned problem that hinders optimization.

To address this issue and provide a better model for optimization, previous works have introduced various notions of generalized-smoothness, which cover a broader class of functions that are used in machine learning applications. For example, distributionally robust optimization (DRO) is an important machine learning problem, and recently it has been proved that DRO can be reformulated as another problem whose objective function belongs to the following asymmetric generalized-smooth function class ($\mathcal{L}^*_{\text{asym}}$) (Levy et al., 2020; Jin et al., 2021).

Definition 1 ($\mathcal{L}^*_{\text{asym}}$ function class)**.** *The asymmetric generalized-smooth function class $\mathcal{L}^*_{asym}$ is the class of differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ that satisfy the following condition for all $w, w' \in \mathbb{R}^d$ and some constants $L_0, L_1 > 0$.*

$$\|\nabla f(w') - \nabla f(w)\| \le \big(L_0 + L_1\|\nabla f(w')\|\big)\|w' - w\|. \quad (2)$$

To elaborate, we name the above function class asymmetric generalized-smooth as the definition in (2) takes an asymmetric form. In particular, the smoothness parameter of the functions in $\mathcal{L}^*_{\text{asym}}$ scales with the gradient norm $\|\nabla f(w')\|$. This implies that the nonconvex problem can be ill-conditioned in the initial optimization stage when the gradient is relatively large.

On the other hand, (Zhang et al., 2019) showed that high-order polynomials and many language models belong to the following Hessian-based generalized-smooth function class $\mathcal{L}^*_{\text{H}}$.

Definition 2 ($\mathcal{L}^*_{\text{H}}$ function class)**.** *The Hessian-based generalized-smooth function class $\mathcal{L}^*_{H}$ is the class of twice-differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ that satisfy the following condition for all $w \in \mathbb{R}^d$ and some constants $L_0, L_1 > 0$.*

$$\|\nabla^2 f(w)\| \le L_0 + L_1\|\nabla f(w)\|. \quad (3)$$

In addition to the above notions of generalized-smoothness, many other works have developed optimization algorithms for minimizing the class of higher-order smooth functions, i.e., functions with Lipschitz continuous higher-order gradients (Nesterov and Polyak, 2006; Carmon et al., 2020; 2021). However, the resulting algorithms usually require either computing higher-order gradients or solving higher-order subproblems, which are not suitable for machine learning applications with big data. In the following subsection, we propose a so-called α-symmetric generalized-smooth function class, which we show substantially generalizes the existing generalized-smooth function classes and covers a wide range of functions used in many important machine learning applications.

3. The α-Symmetric Generalized-Smooth Function Class

We propose the following class of α-symmetric generalized-smooth functions $\mathcal{L}^*_{\text{sym}}(\alpha)$, which we show later covers the aforementioned generalized-smooth function classes and includes many important machine learning problems. Throughout the whole paper, we define $0^0 = 1$.

Definition 3 ($\mathcal{L}^*_{\text{sym}}(\alpha)$ function class)**.** *For $\alpha \in [0, 1]$, the α-symmetric generalized-smooth function class $\mathcal{L}^*_{sym}(\alpha)$ is the class of differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ that satisfy the following condition for all $w, w' \in \mathbb{R}^d$ and some constants $L_0, L_1 > 0$.*

$$\begin{aligned}&\|\nabla f(w') - \nabla f(w)\| \\ &\le \big(L_0 + L_1 \max_{\theta\in[0,1]} \|\nabla f(w_\theta(w, w'))\|^\alpha\big)\|w' - w\|, \quad (4)\end{aligned}$$

where $w_\theta(w, w') := \theta w' + (1 - \theta)w$.

Remark: we use $w_\theta(w,w')$ to emphasize its dependence on w,w'. Later whenever w,w' is given, we will use the abbreviation w_θ.

It can be seen that the above function class $\mathcal{L}^*_{\text{sym}}(\alpha)$ covers the aforementioned function classes $\mathcal{L}$ (corresponds to $L_1=0$) and $\mathcal{L}^*_{\text{asym}}$ (with $L_1>0,\alpha=1$ and $\max_{\theta\in[0,1]}\|\nabla f(w_\theta(w,w'))\|$ being replaced with the smaller term $\|\nabla f(w)\|$). In particular, compared to the asymmetric generalized-smooth function class $\mathcal{L}^*_{\text{asym}}$, our proposed function class $\mathcal{L}^*_{\text{sym}}(\alpha)$ generalizes it in two aspects. First, $\mathcal{L}^*_{\text{sym}}(\alpha)$ defines generalized-smoothness in a symmetric way with regard to the points w and w' since it considers the maximum gradient norm over the line segment $\{w_\theta:\theta\in[0,1]\}$. As a comparison, $\mathcal{L}^*_{\text{asym}}$ defines generalized-smoothness in an asymmetric way. Second, $\mathcal{L}^*_{\text{sym}}(\alpha)$ covers the functions whose smoothness parameter can scale polynomially as $\max_{\theta\in[0,1]}\|\nabla f(w_\theta)\|^\alpha$, whereas $\mathcal{L}^*_{\text{asym}}$ only considers the special case $\alpha=1$.

Next, we show connections among all these generalized-smooth function classes, and prove that our proposed function class $\mathcal{L}^*_{\text{sym}}(\alpha)$ is substantially bigger than others.

Theorem 1 (Function class comparison)**.** *The generalized-smooth function classes $\mathcal{L}^*_{asym}$, $\mathcal{L}^*_{\text{H}}$ and $\mathcal{L}^*_{sym}(\alpha)$ satisfy the following properties.*

1. $\mathcal{L}^*_{asym}\subset\mathcal{L}^*_{sym}(1)$*;*
2. $\mathcal{L}^*_{\text{H}}\subset\mathcal{L}^*_{sym}(1)$*. Moreover, they are equivalent when restricted to the set of twice-differentiable functions;*
3. *The polynomial function $f(w)=|w|^{\frac{2-\alpha}{1-\alpha}},w\in\mathbb{R},\alpha\in(0,1)$ satisfies $f\in\mathcal{L}^*_{sym}(\alpha)$. However, $f\notin\mathcal{L}^*_{sym}(\widetilde{\alpha})$ for all $\widetilde{\alpha}\in(0,\alpha)$ and $f\notin\mathcal{L}^*_{asym}$;*
4. *The exponential function $f(w)=e^w+e^{-w},w\in\mathbb{R}$ satisfies $f\in\mathcal{L}^*_{sym}(1)$. However, $f\notin\mathcal{L}^*_{sym}(\widetilde{\alpha})$ for all $\widetilde{\alpha}\in(0,1)$ and $f\notin\mathcal{L}^*_{asym}$.*

Remark: The functions in items 3 & 4 can be generalized to high-dimensional case $w\in\mathbb{R}^d$ by using $f(w)=\|w\|^{\frac{2-\alpha}{1-\alpha}}$ and $f(w)=e^{\|w\|}+e^{-\|w\|}$ respectively.

To elaborate, items 1 & 2 show that a special case of our proposed α-symmetric generalized-smooth function class $\mathcal{L}^*_{\text{sym}}(1)$ includes the other existing generalized-smooth function classes $\mathcal{L}^*_{\text{asym}},\mathcal{L}^*_{\text{H}}$. In particular, when f is restricted to be twice-differentiable, the class $\mathcal{L}^*_{\text{H}}$ is equivalent to $\mathcal{L}^*_{\text{sym}}(1)$. Moreover, items 3 & 4 show that our proposed generalized-smooth function class $\mathcal{L}^*_{\text{sym}}(\alpha)$ includes a wide range of 'fast-growing' functions, including high-order polynomials and even exponential functions, which are not included in $\mathcal{L}^*_{\text{asym}}$. In summary, our proposed α-symmetric generalized-smooth function class $\mathcal{L}^*_{\text{sym}}(\alpha)$ extends the existing boundary of smooth functions in nonconvex optimization.

Next, for the functions in $\mathcal{L}^*_{\text{sym}}(\alpha)$, we establish various important technical tools that are leveraged later to develop efficient algorithms and their convergence analysis.

Proposition 1 (Technical tools)**.** *The function class $\mathcal{L}^*_{sym}(\alpha)$ can be equivalently defined as follows.*

1. *For any $\alpha\in(0,1)$, function f belongs to $\mathcal{L}^*_{sym}(\alpha)$ if and only if for any $w,w'\in\mathbb{R}^d$,*
$$\|\nabla f(w')-\nabla f(w)\|\le\|w'-w\|\cdot\big(K_0+K_1\|\nabla f(w)\|^\alpha+K_2\|w'-w\|^{\frac{\alpha}{1-\alpha}}\big).\tag{5}$$
where $K_0:=L_0\big(2^{\frac{\alpha^2}{1-\alpha}}+1\big)$, $K_1:=L_1\cdot 2^{\frac{\alpha^2}{1-\alpha}}\cdot 3^\alpha$, $K_2:=L_1^{\frac{1}{1-\alpha}}\cdot 2^{\frac{\alpha^2}{1-\alpha}}\cdot 3^\alpha(1-\alpha)^{\frac{\alpha}{1-\alpha}}$.
2. *For $\alpha=1$, function f belongs to $\mathcal{L}^*_{sym}(1)$ if and only if for any $w,w'\in\mathbb{R}^d$,*
$$\|\nabla f(w')-\nabla f(w)\|\le\|w'-w\|\cdot\big(L_0+L_1\|\nabla f(w)\|\big)\exp\big(L_1\|w'-w\|\big).\tag{6}$$

Consequently, the following descent lemmas hold.

3. *If $f\in\mathcal{L}^*_{sym}(\alpha)$ for $\alpha\in(0,1)$, then for any $w,w'\in\mathbb{R}^d$,*
$$f(w')\le f(w)+\nabla f(w)^\top(w'-w)+\frac{1}{2}\|w'-w\|^2\cdot\big(K_0+K_1\|\nabla f(w)\|^\alpha+2K_2\|w'-w\|^{\frac{\alpha}{1-\alpha}}\big).\tag{7}$$
4. *If $f\in\mathcal{L}^*_{sym}(1)$, then for any $w,w'\in\mathbb{R}^d$,*
$$f(w')\le f(w)+\nabla f(w)^\top(w'-w)+\frac{1}{2}\|w'-w\|^2\cdot\big(L_0+L_1\|\nabla f(w)\|\big)\exp\big(L_1\|w'-w\|\big).\tag{8}$$

Technical Novelty. Proving the above items 1 & 2 turns out to be non-trivial and critical, because they directly imply the items 3 & 4[2], which play an important role in the convergence analysis of the algorithms proposed later in this paper. Specifically, there are two major steps to prove the equivalent definitions in items 1 & 2. First, we prove another equivalent definition, i.e., $f\in\mathcal{L}^*_{\text{sym}}(\alpha)$ if and only if for any $w,w\in\mathbb{R}^d$,
$$\|\nabla f(w')-\nabla f(w)\|\le\Big(L_0+L_1\int_0^1\|\nabla f(w_\theta)\|^\alpha d\theta\Big)\|w'-w\|.\tag{9}$$

Please refer to (25) in Lemma 2 in Appendix A for the details. To prove this, we uniformly divide the line segment

[2] Items 3 & 4 of Proposition 1 can be obtained by substituting items 1 & 2 into the inequality that $f(w')-f(w)-\nabla f(w)^\top(w'-w)\le\int_0^1|\nabla f(w_\theta)-\nabla f(w)||w'-w|d\theta$

between w and w' into n pieces with the end points $\{w_\theta : \theta = \frac{k}{n}\}_{k=0}^{n}$. Then, we obtain the following bound.

$$\begin{aligned}\|\nabla f(w')-\nabla f(w)\| &\le \sum_{k=0}^{n-1} \|\nabla f(w_{(k+1)/n})-\nabla f(w_{k/n})\| \\ &\le \|w'-w\| \sum_{k=0}^{n-1} \frac{1}{n} \max_{\theta\in[k/n,(k+1)/n]} h(\theta),\end{aligned}$$

where $w_{k/n}$ and $w_{(k+1)/n}$ denote w_θ with $\theta = k/n$ and $\theta = (k+1)/n$ respectively, and $h(\theta) := L_0 + L_1\|\nabla f(w_\theta)\|^\alpha$. As $n \to +\infty$, the summation in the above inequality converges to the desired integral $\int_0^1 h(\theta)d\theta$. Second, to prove sufficiency, i.e., (9) implies (5) & (6), we derive and solve an ordinary differential equation (ODE) of the function $H(\theta) := \int_0^\theta h(\theta')d\theta'$. This ODE is obtained by substituting $w' = w_\theta$ into the above equivalent definition (9). Then, to prove necessity, i.e., (5) & (6) imply (9), we use a similar dividing technique so that averaging the terms $K_0 + K_1\|\nabla f(w_{k/n})\|^\alpha$ and $L_0 + L_1\|\nabla f(w_{k/n})\|$ over $k = 0, 1, \ldots, n-1$ yields the desired integral as $n \to +\infty$, while at the same time the other terms vanish as $\|w_{(k+1)/n} - w_{k/n}\|^{\frac{\alpha}{1-\alpha}} \to 0$ and $\exp\big(L_1\|w_{(k+1)/n} - w_{k/n}\|\big) \to 1$.

Next, we present some nonconvex machine learning examples that belong to the proposed function class $\mathcal{L}^*_{\text{sym}}(\alpha)$.

Example 1: Phase Retrieval. Phase retrieval is a classic nonconvex machine learning and signal processing problem that arises in X-ray crystallography and coherent diffraction imaging applications (Drenth, 1994; Miao et al., 1999). In this problem, we aim to recover the structure of a molecular object from far-field diffraction intensity measurements when the object is illuminated by a source light. Mathematically, denote the underlying true object as $x \in \mathbb{R}^d$ and suppose we take m intensity measurements, i.e., $y_r = |a_r^\top x|^2, r = 1, 2, ..., m$ where $a_r \in \mathbb{R}^d$ and $\top$ denotes transpose. Then, phase retrieval proposes to recover the signal by solving the following nonconvex problem.

$$\min_{z\in\mathbb{R}^d} f(z) := \frac{1}{2m}\sum_{r=1}^{m}(y_r - |a_r^\top z|^2)^2. \tag{10}$$

The above nonconvex objective function is a high-order polynomial in the high-dimensional space. Therefore, it does not belong to the L-smooth function class $\mathcal{L}$. In the following result, we formally prove that the above phase retrieval problem can be effectively modeled by our proposed function class $\mathcal{L}^*_{\text{sym}}(\alpha)$.

Proposition 2. *The nonconvex phase retrieval objective function $f(z)$ in (10) belongs to $\mathcal{L}^*_{sym}(\frac{2}{3})$.*

Example 2: Distributionally Robust Optimization. In many practical machine learning applications, there is usually a gap between training data distribution and test data distribution. Therefore, it is much desired to train a model that is robust to distribution shift. Distributionally robust optimization (DRO) is such a popular optimization framework for training robust models. Specifically, DRO aims to solve the following problem

$$\min_{x\in\mathcal{X}} f(x) := \sup_{Q}\big\{\mathbb{E}_{\xi\sim Q}[\ell_\xi(x)] - \lambda d_\psi(Q,P)\big\}, \tag{11}$$

where the ψ-divergence term $\lambda d_\psi(Q,P)$ ($\lambda > 0$) penalizes the distribution shift between the training data distribution Q and the target distribution P, and it takes the form $d_\psi(Q,P) := \int \psi(\frac{dQ}{dP})dP$. Under mild assumptions on the nonconvex sample loss function ℓ_ξ (e.g., smooth and bounded variance) and the divergence function ψ, the above DRO problem is proven to be equivalent to the following minimization problem (Levy et al., 2020; Jin et al., 2021).

$$\min_{x\in\mathcal{X},\eta\in\mathbb{R}} L(x,\eta) := \lambda\mathbb{E}_{\xi\sim P}\psi^*\left(\frac{\ell_\xi(x)-\eta}{\lambda}\right) + \eta, \tag{12}$$

where ψ^* denotes the convex conjugate function of ψ. In particular, the objective function $L(x,\eta)$ in the above equivalent form has been shown to belong to the function class $\mathcal{L}^*_{\text{asym}}$ (Jin et al., 2021). Therefore, by item 1 of Theorem 1, we can make the following conclusion.

Lemma 1. *Regarding the equivalent form (12) of the DRO problem (11), its objective function L belongs to the function class $\mathcal{L}^*_{sym}(1)$.*

4. Optimal Method for Solving Nonconvex Problems in $\mathcal{L}^*_{\text{sym}}(\alpha)$

In this section, we develop an efficient and optimal deterministic gradient-based algorithm for minimizing nonconvex functions in $\mathcal{L}^*_{\text{sym}}(\alpha)$ and analyze its iteration complexity.

The challenge for optimizing the functions in $\mathcal{L}^*_{\text{sym}}(\alpha)$ is that the generalized-smoothness parameter scales with $\max_{\theta\in[0,1]}\|\nabla f(w_\theta)\|^\alpha$. To address this issue, we need to use a specialized gradient normalization technique, and this motivates us to consider the β-normalized gradient descent (β-GD) algorithm as shown in Algorithm 1. To elaborate, β-GD simply normalizes the gradient update by the gradient norm term $\|\nabla f(w_t)\|^\beta$ for some $\beta \ge 0$. Such a normalized update is closely related to some existing gradient-type algorithms, including the clipped GD algorithm that uses the normalization term $\max\{\|\nabla f(w_t)\|, C\}$ and the normalized GD that uses the normalization term $\|\nabla f(w_t)\| + C$ (Zhang et al., 2019), where $C > 0$ is a certain constant. We obtain the following convergence result of β-GD on minimizing functions in $\mathcal{L}^*_{\text{sym}}(\alpha)$.

Theorem 2 (Convergence of β-GD). *Apply the β-GD algorithm to minimize any function $f \in \mathcal{L}^*_{sym}(\alpha)$ with $\beta \in [\alpha, 1]$. Choose $\gamma = \frac{\epsilon^\beta}{12(K_0+K_1+2K_2)+1}$[3] if $\alpha \in (0,1)$*

[3] See the definition of K_0, K_1, K_2 in Proposition 1.

Algorithm 1 β-Normalized GD

```
Input: Iteration number $T$, initialization $w_0$, learning rate $\gamma$, normalization parameter $\beta$.
for $t = 0, 1, 2, \dots, T-1$ do
    Update $w_{t+1} = w_t - \gamma \frac{\nabla f(w_t)}{\|\nabla f(w_t)\|^{\beta}}$.
end
Output: $w_{\widetilde{T}}$ where $\widetilde{T}$ is sampled from $\{0, 1, \dots, T-1\}$ uniformly at random.
```

and $\gamma = \frac{\epsilon^{\beta}}{4L_0+1}$ *if* $\alpha = 1$ *(*ϵ *is the target accuracy). Then, the following convergence rate result holds.*

$$\mathbb{E}_{\widetilde{T}}\|\nabla f(w_{\widetilde{T}})\| \le \Big(\frac{2}{T\gamma}\Big)^{\frac{1}{2-\beta}}\big(f(w_0) - f^*\big)^{\frac{1}{2-\beta}} + \frac{1}{2}\epsilon. \quad (13)$$

Consequently, to achieve $\mathbb{E}_{\widetilde{T}}\|\nabla f(w_{\widetilde{T}})\| \le \epsilon$*, the required overall iteration complexity is* $T = \frac{4}{\gamma \epsilon^{2-\beta}} = \mathcal{O}(\epsilon^{-2})$.

Theorem 2 shows that β-GD achieves the iteration complexity $\mathcal{O}(\epsilon^{-2})$ when minimizing functions in $\mathcal{L}^*_{\text{sym}}(\alpha)$. Such a complexity result matches the iteration complexity lower bound for deterministic smooth nonconvex optimization and hence is optimal. In particular, Theorem 2 shows that to minimize any function $f \in \mathcal{L}^*_{\text{sym}}(\alpha)$, it suffices to apply β-GD with any $\beta \in [\alpha, 1]$ and a proper learning rate $\gamma = \mathcal{O}(\epsilon^{\beta})$. Intuitively, with a larger α, the gradient norm of function f in the class $\mathcal{L}^*_{\text{sym}}(\alpha)$ increases faster as $\|w\| \to +\infty$, and therefore we need to use a larger normalization parameter β and a smaller learning rate $\mathcal{O}(\epsilon^{\beta})$ to alleviate gradient explosion. Interestingly, the convergence and iteration complexity of β-GD remain the same as long as $\beta \ge \alpha$ is used, i.e., over-normalization does not affect the complexity order. In practice, when α is unknown a priori for the function class $\mathcal{L}^*_{\text{sym}}(\alpha)$, one can simply use the conservative choice $\beta = 1$ and is guaranteed to converge.

Technical Novelty. In the proof of Theorem 2, a major challenge is that due to the β-normalization term in Algorithm 1, the generalized-smoothness of functions in the class $\mathcal{L}^*_{\text{sym}}(\alpha)$ introduces additional higher-order terms to the Taylor expansion upper bounds, as can be seen from the descent lemmas shown in (7) (for $\alpha \in (0, 1)$) and (8) (for $\alpha = 1$). In the convergence proof, these terms contribute to certain fast-increasing terms that reduce the overall optimization progress. For example, when $\alpha \in (0, 1)$, substituting $w = w_t$ and $w' = w_{t+1}$ into (7) yields that [4]

$$\begin{aligned} f(w_{t+1}) - f(w_t) &\le -\gamma\|\nabla f(w_t)\|^{2-\beta} \\ &+ \frac{\gamma}{6}\Big(\mathcal{O}(\gamma)\|\nabla f(w_t)\|^{2-2\beta} + \mathcal{O}(\gamma)\|\nabla f(w_t)\|^{2+\alpha-2\beta} \\ &+ \mathcal{O}(\gamma^{\frac{1}{1-\alpha}})\|\nabla f(w_t)\|^{\frac{(2-\alpha)(1-\beta)}{1-\alpha}}\Big). \end{aligned} \quad (14)$$

The above key inequality bounds the optimization progress $f(w_{t+1}) - f(w_t)$ using gradient norm terms with very different exponents. This makes it challenging to achieve the desired level of optimization progress, as compared with the analysis of minimizing other (generalized) smooth functions in $\mathcal{L}$, $\mathcal{L}^*_{\text{asym}}$ and $\mathcal{L}^*_{\text{H}}$ (Zhang et al., 2019; Jin et al., 2021). To address this issue and homogenize the diverse exponents, we develop a technical tool in Lemma 5 in Appendix A to bridge polynomials with different exponents. With this technique, we further obtain the following optimization progress bound

$$f(w_{t+1}) - f(w_t) \le -\frac{\gamma}{2}\|\nabla f(w_t)\|^{2-\beta} + \mathcal{O}(\gamma^{\frac{2}{\beta}}), \quad (15)$$

which leads to the desired result with proper telescoping.

We also obtain the following complementary result to Theorem 2, which shows that β-GD may diverge in general with under-normalization.

Theorem 3. *(Divergence of* β*-GD) For the* β*-GD algorithm with* $\beta \in [0, \alpha)$*, there always exists a convex function* $f \in \mathcal{L}^*_{sym}(\alpha)$ *with a unique minimizer such that for any learning rate* $\gamma > 0$*,* β*-GD diverges for all initialization* $\|w_0\| > C$ *for some constant* $C > 0$.

5. Expected α-Symmetric Generalized-Smooth Functions in Stochastic Optimization

In this section, we propose a class of expected α-symmetric generalized-smooth functions and study their properties in stochastic optimization. Specifically, we consider the following nonconvex stochastic optimization problem

$$\min_{w \in \mathbb{R}^d} f(w) := \mathbb{E}_{\xi \sim \mathbb{P}}[f_{\xi}(w)], \quad (16)$$

where $\mathbb{P}$ denotes the distribution of the data sample ξ. Throughout, we adopt the following standard assumption on the stochastic gradients (Ghadimi and Lan, 2013; Fang et al., 2018; Jin et al., 2021; Arjevani et al., 2022).

Assumption 1. *The stochastic gradient is unbiased, i.e.,* $\mathbb{E}_{\xi \sim \mathbb{P}}[\nabla f_{\xi}(w)] = \nabla f(w)$ *and satisfies the following variance bound for some* $\Gamma, \Lambda > 0$.

$$\mathbb{E}_{\xi \sim \mathbb{P}}\|\nabla f_{\xi}(w) - \nabla f(w)\|^2 \le \Gamma^2\|\nabla f(w)\|^2 + \Lambda^2. \quad (17)$$

If only the population loss f is smooth, i.e., $f \in \mathcal{L}$, a recent work has established a sample complexity lower bound $\mathcal{O}(\epsilon^{-4})$ for first-order stochastic algorithms (Arjevani et al., 2022), which can be achieved by the standard stochastic gradient descent (SGD) algorithm (Ghadimi and Lan, 2013) and its clipped and normalized versions (Zhang et al., 2019). Therefore, one should not expect an improved sample complexity when optimizing the larger class of generalized-smooth functions $\mathcal{L}^*_{\text{sym}}(\alpha)$. To overcome this sample complexity barrier, many existing works consider the subclass of

[4] See (i) of (45) in Appendix E for the full expression of $\mathcal{O}$ in eq. (14).

expected smooth functions $\mathbb{E}\mathcal{L}$, in which there exists a constant $L_0 > 0$ such that for all $w, w' \in \mathbb{R}^d$, all the functions f_ξ satisfy

$$\mathbb{E}_\xi \|\nabla f_\xi(w') - \nabla f_\xi(w)\|^2 \le L_0^2 \|w' - w\|^2. \tag{18}$$

Many variance-reduced algorithms, e.g., SPIDER (Fang et al., 2018) and STORM (Cutkosky and Orabona, 2019), have been proved to achieve the near-optimal sample complexity $\mathcal{O}(\epsilon^{-3})$ for optimizing functions in $\mathbb{E}\mathcal{L}$. Therefore, we are inspired to propose and study the following expected α-symmetric generalized-smooth function class $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$.

Definition 4 ($\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$ function class)**.** *For $\alpha \in [0, 1]$, the expected α-symmetric generalized-smooth function class $\mathbb{E}\mathcal{L}^*_{sym}(\alpha)$ is the class of differentiable stochastic functions $f = \mathbb{E}_\xi[f_\xi]$ that satisfy the following condition for all $w, w' \in \mathbb{R}^d$ and some constants $L_0, L_1 > 0$.*

$$\begin{aligned}&\mathbb{E}_{\xi\sim\mathbb{P}} \|\nabla f_\xi(w') - \nabla f_\xi(w)\|^2 \\ &\le \|w' - w\|^2 \mathbb{E}_{\xi\sim\mathbb{P}} \big(L_0 + L_1 \max_{\theta\in[0,1]} \|\nabla f_\xi(w_\theta)\|^\alpha\big)^2\end{aligned} \tag{19}$$

where $w_\theta := \theta w' + (1-\theta)w$.

Remark: It is clear that the function class $\mathbb{E}\mathcal{L}^*_{\text{sym}}(0)$ is equivalent to the function class $\mathbb{E}\mathcal{L}$. Also, a sufficient condition to guarantee $f \in \mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$ is that $f_\xi \in \mathcal{L}^*_{\text{sym}}(\alpha)$ for every sample ξ.

Proposition 3. *Both the aforementioned phase retrieval problem and DRO problem belong to $\mathbb{E}\mathcal{L}^*_{sym}(\alpha)$ with $\alpha = \frac{2}{3}, 1$ respectively.*

We further develop the following technical tools associated with the function class $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$, which are used later to analyze a stochastic algorithm.

Proposition 4 (Technical tools)**.** *Under Assumption 1, the following statements hold.*

1. *For any $\alpha \in (0, 1)$, function $f = \mathbb{E}_\xi[f_\xi]$ belongs to $\mathbb{E}\mathcal{L}^*_{sym}(\alpha)$ if and only if for any $w, w' \in \mathbb{R}^d$,*

$$\begin{aligned}&\mathbb{E}_\xi \|\nabla f_\xi(w') - \nabla f_\xi(w)\|^2 \le \|w' - w\|^2 \\ &\quad\cdot \big(\overline{K}_0 + \overline{K}_1 \mathbb{E}_\xi \|\nabla f_\xi(w)\|^\alpha + \overline{K}_2 \|w' - w\|^{\frac{\alpha}{1-\alpha}}\big)^2,\end{aligned} \tag{20}$$

where $\overline{K}_0 = 2^{\frac{2-\alpha}{1-\alpha}} L_0$, $\overline{K}_1 = 2^{\frac{2-\alpha}{1-\alpha}} L_1$, $\overline{K}_2 = (5L_1)^{\frac{1}{1-\alpha}}$;

2. *For $\alpha = 1$, function $f = \mathbb{E}_\xi[f_\xi]$ belongs to $\mathbb{E}\mathcal{L}^*_{sym}(\alpha)$ if and only if for any $w, w' \in \mathbb{R}^d$,*

$$\begin{aligned}&\mathbb{E}_\xi \|\nabla f_\xi(w') - \nabla f_\xi(w)\|^2 \le 2\|w' - w\|^2 \\ &\quad\cdot (L_0^2 + 2L_1^2 \mathbb{E}_\xi \|\nabla f_\xi(w)\|^2) \exp(12 L_1^2 \|w' - w\|^2).\end{aligned} \tag{21}$$

3. *$\mathbb{E}\mathcal{L}^*_{sym}(\alpha) \subset \mathcal{L}^*_{sym}(\alpha)$.*

Remark: Item 3 of Proposition 4 implies that we can apply the descent lemmas (items 3 & 4 of Proposition 1) to the population loss $f = \mathbb{E}_\xi[f_\xi]$. This is very useful later in the convergence analysis of our proposed stochastic algorithm.

6. Optimal Method for Solving Nonconvex Problems in $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$

In this section, we explore stochastic algorithms for solving nonconvex problems in the function class $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$ and see if any algorithm can achieve the optimal sample complexity.

In the existing literature, many stochastic variance reduction algorithms, e.g., SPIDER (Fang et al., 2018) and STORM (Cutkosky and Orabona, 2019), have been developed and proved to achieve the optimal sample complexity $\mathcal{O}(\epsilon^{-3})$ for minimizing the class of expected-smooth stochastic nonconvex problems (i.e., $\mathbb{E}\mathcal{L}$). However, for the extended class $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$, it is unclear what is the sample complexity lower bound and the optimal stochastic algorithm design. Inspired by the existing literature and the structures of functions in $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$, a good algorithm design must apply both variance reduction and a proper normalization to the stochastic updates in order to combat the generalized-smoothness and achieve an improved sample complexity. Interestingly, we discover that the original SPIDER algorithm design already well balances these two techniques and can be directly applied to solve problems in $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$. The original SPIDER algorithm is summarized in Algorithm 2 below.

Algorithm 2 SPIDER (Fang et al., 2018)

```
Input: Iteration number $T$, epoch size $q$, initialization $w_0$,
learning rate $\gamma$, batchsize $|S_t|$.
for $t = 0, 1, 2, \dots, T-1$ do
    Sample a minibatch of data $S_t$.
    if $t \mod q = 0$ then
        Compute $v_t = \nabla f_{S_t}(w_t)$
    else
        Compute $v_t = v_{t-1} + \nabla f_{S_t}(w_t) - \nabla f_{S_t}(w_{t-1})$
    end
    Update $w_{t+1} = w_t - \gamma \frac{v_t}{\|v_t\|}$.
end
Output: $w_{\widetilde{T}}$ where $\widetilde{T}$ is sampled from $\{0, 1, \dots, T-1\}$
uniformly at random..
```

However, establishing the convergence of SPIDER for the extended function class $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$ is fundamentally more challenging. Intuitively, this is because the characterization of variance of the stochastic update v_t is largely affected by the generalized-smoothness structure, and it takes a complex form that needs to be treated carefully. Please refer to the elaboration on technical novelty later for more details.

Surprisingly, by choosing proper hyper-parameters that are adapted to the function class $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$, we are able to prove that SPIDER achieves the optimal sample complexity as formally stated in the following theorem.

Theorem 4 (Convergence of SPIDER)**.** *Apply the SPIDER algorithm to minimize any function $f = \mathbb{E}_\xi[f_\xi] \in \mathbb{E}\mathcal{L}^*_{sym}(\alpha)$ and assume Assumption 1 hold. Set $|S_t| = B$ when*

$t \mod q = 0$ *and* $|S_t| = B'$ *otherwise, and let* $B \geq \Omega(\max\{\Lambda^2\epsilon^{-2}, \Gamma^2 q^2\}), B' \geq \Omega(\max\{q, q^2\epsilon^2\})$. *Choose* $\gamma = \frac{\epsilon}{2\overline{K}_0+4\overline{K}_2+2\overline{K}_1(\Lambda^\alpha+\Gamma^\alpha+1)+1}$ *when* $\alpha \in (0,1)$ *and* $\gamma = \frac{\epsilon}{5L_1\sqrt{\Gamma^2+1}+8\sqrt{L_0^2+2L_1^2\Lambda^2}}$ *when* $\alpha = 1$ *(*ϵ *is the target accuracy). Then, the following result holds for* $T = qK$ *iterations where* $K \in \mathbb{N}^+$.

$$\mathbb{E}\|\nabla f(w_{\widetilde{T}})\| \leq \frac{16}{5T\gamma}\big(\mathbb{E}f(w_0) - f^*\big) + \frac{4\epsilon}{5}. \qquad (22)$$

In particular, to achieve $\mathbb{E}\|\nabla f(w_{\widetilde{T}})\| \leq \epsilon$, *we can choose* $B = \mathcal{O}(\epsilon^{-2})$, $B' = q = \mathcal{O}(\epsilon^{-1})$, $\gamma = \mathcal{O}(\epsilon)$ *and* $T = \mathcal{O}(\epsilon^{-2})$[5] *so that the above conditions are satisfied. Consequently, the overall sample complexity is* $\mathcal{O}(\epsilon^{-3})$.

Theorem 4 proves that SPIDER achieves an overall sample complexity $\mathcal{O}(\epsilon^{-3})$ when solving nonconvex problems in $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$ for any $\alpha \in (0, 1]$. Note that such a sample complexity matches the well-known sample complexity lower bound for the class of expected-smooth nonconvex optimization problems (Fang et al., 2018), which is a subset of $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$. Consequently, we can make two important observations: (i) this implies that the sample complexity lower bound of $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$ is actually $\mathcal{O}(\epsilon^{-3})$; and (ii) the SPIDER algorithm is provably optimal for solving nonconvex problems in such an extended function class.

Technical Novelty. Compared with the original analysis of SPIDER for minimizing expected-smooth functions (Fang et al., 2018), our proof of the above theorem needs to address a major challenge on bounding the expected bias error $\mathbb{E}\|\delta_t\|$ where $\delta_t = v_t - \nabla f(x_t)$. To elaborate more specifically, in the original analysis of SPIDER, (Fang et al., 2018) established the following key lemma (see their Lemma 1) that bounds the martingale variance of the update v_t.

$$\mathbb{E}\|\delta_t\|^2 \leq \mathbb{E}\|\delta_0\|^2 + \frac{1}{B'}\sum_{k=0}^{t-1}\mathbb{E}_\xi\|\nabla f_\xi(w_{t+1}) - \nabla f_\xi(w_t)\|^2.$$

The above inequality only depends on the variance reduction structure of SPIDER and hence still holds in our case. However, to further bound the term $\mathbb{E}_\xi\|\nabla f_\xi(w_{t+1}) - \nabla f_\xi(w_t)\|^2$ for functions in the class $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$, we need to leverage the expected generalized-smoothness properties in (20) & (21) and the update rule $\|w_{t+1} - w_t\| = \gamma = \mathcal{O}(\epsilon)$. We then obtain that $\mathbb{E}_\xi\|\nabla f_\xi(w_{t+1}) - \nabla f_\xi(w_t)\|^2 \leq \mathcal{O}(\epsilon^2) + \mathcal{O}(\epsilon^2\|\nabla f(w_t)\|^{2\alpha})$, and consequently, the above martingale variance bound becomes

$$\mathbb{E}\|\delta_t\|^2 \leq \mathbb{E}\|\delta_0\|^2 + \mathcal{O}(\epsilon^2) + \frac{\mathcal{O}(\epsilon^2)}{B'}\sum_{k=0}^{t-1}\mathbb{E}\|\nabla f(w_t)\|^{2\alpha}. \quad (23)$$

When $\alpha > \frac{1}{2}$, the term $\mathbb{E}\|\nabla f(w_t)\|^{2\alpha}$ in (23) cannot be upper bounded by any functional of $\mathbb{E}\|\nabla f(w_t)\|$, so taking square root of (23) cannot yield the desired bound $\mathbb{E}\|\delta_t\| \leq \mathcal{O}(\epsilon) + \frac{\mathcal{O}(\epsilon)}{\sqrt{B'}}\sum_{k=0}^{t-1}\mathbb{E}\|\nabla f(w_t)\|$ used in the original analysis of SPIDER. To address this issue for functions in $\mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$, we consider the more refined conditional error recursion

$$\mathbb{E}\big(\|\delta_{t+1}\|^2\big|S_{1:t}\big) \leq \|\delta_t\|^2 + \frac{\epsilon^2}{B'}\big(1 + \|\nabla f(w_t)\|^2\big), \quad (24)$$

where there is no randomness in $\|\nabla f(w_t)\|^2$ since we are conditioning on the minibatches $S_{1:t} := \{S_1, \ldots, S_t\}$ (see (58) in Appendix A). Therefore, by taking square root of (24) followed by further taking iterated expectation, we can obtain the desired term $\mathbb{E}\|\nabla f(w_t)\|$ in the upper bound of $\mathbb{E}\|\delta_t\|$. After that, we iterate the resulting bound over t via a non-trivial induction argument to complete the analysis (see the proof of (60) in Lemma 58 in Appendix A).

7. Experiments[6]

7.1. Application to Nonconvex Phase Retrieval

In this section, we test our algorithms via solving the nonconvex phase retrieval problem in (10). We set the problem dimension $d = 100$ and sample size $m = 3000$. The measurement vector $a_r \in \mathbb{R}^d$ and the underlying true parameter $z \in \mathbb{R}^d$ are generated entrywise using Gaussian distribution $\mathcal{N}(0, 0.5)$. The initialization $z_0 \in \mathbb{R}^d$ is generated entrywise using Gaussian distribution $\mathcal{N}(5, 0.5)$. Then, we generate $y_i = |a_r^\top z|^2 + n_i$ for $i = 1, ..., m$, where $n_i \sim \mathcal{N}(0, 4^2)$ is the additive Gaussian noise.

We first compare deterministic algorithms with fine-tuned learning rate γ over 500 iterations. This includes the basic GD with $\gamma = 8 \times 10^{-4}$, clipped GD (Zhang et al., 2019) with $\gamma = 0.9$ and normalization term $\max(\|\nabla f(x_t)\|, 100)$, and our β-GD with $\beta = \frac{1}{3}, \frac{2}{3}, 1$ and $\gamma = 0.03, 0.1, 0.2$, respectively. Figure 1 (top left) plots the comparison result on objective function value v.s. iteration. It can be seen that our proposed β-GD with $\beta = \frac{1}{3}, \frac{2}{3}$ converges faster than the existing GD, normalized GD (1-GD) and clipped GD algorithms, which shows the advantage of using a proper normalization parameter β.

We further compare stochastic algorithms with fine-tuned learning rate γ and fixed batch size $b = 50$ over 500 iterations. This includes the basic SGD with $\gamma = 2 \times 10^{-4}$, normalized SGD with $\gamma = 2 \times 10^{-3}$, normalized SGD with momentum (Jin et al., 2021) with $\gamma = 3 \times 10^{-3}$ and momentum coefficient 10^{-4}, clipped SGD (Zhang et al., 2019) with $\gamma = 0.3$ and normalization term $\max(\|\nabla f(z_t)\|, 10^3)$, and SPIDER with $\gamma = 0.01$, epoch size $q = 5$ and batchsizes $B = 3000, B' = 50$. We generate the initialization

[5] See eqs. (65)-(70) in Appendix H for the full expression of these hyperparameters.

[6] The code can be downloaded from https://github.com/changy12/Generalized-Smooth-Nonconvex-Optimization-is-As-Efficient-As-Smooth-Nonconvex-Optimization

by running $\frac{2}{3}$-GD with $\gamma = 0.1$ for 100 iterations from z_0. Figure 1 (top right) plots the comparison result on objective function value v.s. sample complexity. It can be seen that SPIDER uses slightly more samples at the beginning but converges to a much better solution than the other SGD-type algorithms. This demonstrates the advantage of applying both variance reduction and proper normalization to solve generalized-smooth nonconvex stochastic problems.

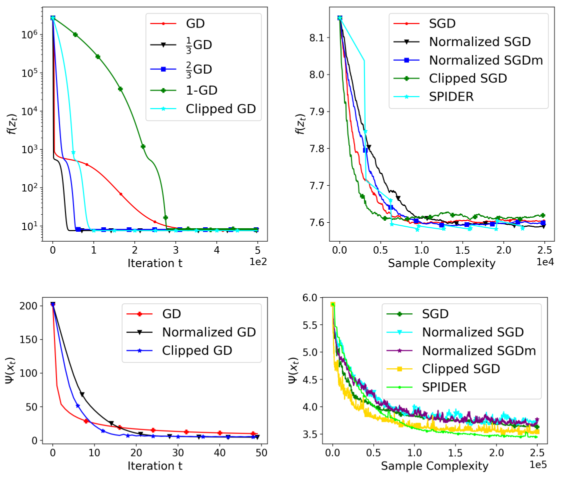

Figure 1: Experimental results (two subfigures above for phase retrieval and two below for DRO).

7.2. Application to DRO

In this section, we test our algorithms via solving the nonconvex DRO problem in (12) on the life expectancy data[7], which collected the life expectancy (target) and its influencing factors (features) of 2413 people for regression analysis. We preprocess the data by filling the missing values with the median of the corresponding variables, censorizing and standardizing all the variables[8], removing two categorical variables ("country" and "status"), and adding standard Gaussian noise to the target to ensure model robustness. We select the first 2000 samples $\{x_i, y_i\}_{i=1}^{2000}$ as the training samples where $x_i \in \mathbb{R}^{34}$ and $y_i \in \mathbb{R}$ are feature and target respectively. In the DRO problem (12), we set $\lambda = 0.01$ and select $\psi^*(t) = \frac{1}{4}(t+2)_+^2 - 1$ which corresponds to χ^2 divergence. For any sample pair x_ξ, y_ξ, we adopt the regularized mean square loss function $\ell_\xi(w) = \frac{1}{2}(y_\xi - x_\xi^\top w)^2 + 0.1\sum_{j=1}^{34} \ln\left(1 + |w^{(j)}|\right)$ with parameter $w = [w^{(1)}; \ldots; w^{(34)}] \in \mathbb{R}^{34}$. We initialize $\eta_0 = 0.1$ and randomly initialize $w_0 \in \mathbb{R}^{34}$ entrywise using standard Gaussian distribution.

We first compare deterministic algorithms with fine-tuned learning rate γ over 50 iterations. This includes the basic GD with $\gamma = 10^{-4}$, clipped GD (Zhang et al., 2019) with $\gamma = 0.3$ and normalization term $\max(\|\nabla L(x_t, \eta_t)\|, 10)$, and normalized GD (our β-GD with $\beta = 1$) with $\gamma = 0.2$, respectively. Figure 1 (bottom left) plots the comparison result on the objective function value $\Psi(x_t) := \min_{\eta\in\mathbb{R}} L(x_t, \eta)$ (L is defined in eq. (12)) v.s. iteration. It can be seen that normalized GD and clipped GD converge to comparable function values and both outperform standard GD.

We further compare stochastic algorithms with fine-tuned learning rate γ and fixed minibatch size $b = 50$ over 5000 iterations. This includes the basic SGD with $\gamma = 2 \times 10^{-4}$, normalized SGD with $\gamma = 8 \times 10^{-3}$, normalized SGD with momentum with $\gamma = 8 \times 10^{-3}$ and momentum coefficient 10^{-4}, clipped SGD [1] with $\gamma = 0.05$ and normalization term $\max(\|\nabla L(x_t, \eta_t)\|, 100)$, and SPIDER with $\gamma = 4 \times 10^{-3}$, epoch size $q = 20$ and batchsizes $B = 2000, B' = 50$. We generate the initialization by running normalized GD with $\gamma = 0.2$ for 30 iterations from w_0, η_0. Figure 1 (bottom right) plots the comparison result on objective function value $\Psi(x_t)$ v.s. sample complexity. It can be seen that SPIDER takes slightly more samples at the beginning but converges to a better solution than the other SGD-type algorithms. This demonstrates the advantage of applying both variance reduction and proper normalization to solve generalized-smooth nonconvex stochastic problems.

8. Conclusion

In this work, we proposed a new class of generalized-smooth functions that extends the existing ones. We developed both deterministic and stochastic gradient-based algorithms for solving problems in this class and obtained the optimal complexities. 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 other popular variance reduction algorithms such as STORM and SpiderBoost can be normalized to solve generalized-smooth nonconvex stochastic problems.

Acknowledgement

The work of Ziyi and Yi Zhou was supported in part by U.S. National Science Foundation under the grants CCF-2106216, DMS-2134223 and CAREER-2237830.

The work of Yingbin was supported in part by the U.S. National Science Foundation under the grants CCF-1900145 and CCF-1909291.

The work of Zhaosong was supported in part by U.S. National Science Foundation under the grant IIS-2211491.

[7] https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?resource=download

[8] The detailed process of filling missing values and censorization can be seen in https://thecleverprogrammer.com/2021/01/06/life-expectancy-analysis-with-python/

References

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2022). Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pages 1–50.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2020). Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2021). Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 185(1-2):315–355.

Crawshaw, M., Liu, M., Orabona, F., Zhang, W., and Zhuang, Z. (2022). Robustness to unbounded smoothness of generalized signsgd. In *Advances in Neural Information Processing Systems*.

Cutkosky, A. and Mehta, H. (2020). Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR.

Cutkosky, A. and Orabona, F. (2019). Momentum-based variance reduction in non-convex sgd. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 15236–15245.

Drenth, J. (1994). Principles of protein x-ray crystallography. *Springer Science & Business Media*.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. (2018). Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31.

Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

Jin, J., Zhang, B., Wang, H., and Wang, L. (2021). Non-convex distributionally robust optimization: Non-asymptotic analysis. *Advances in Neural Information Processing Systems*, 34:2771–2782.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. (2020). Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860.

Miao, J., Charalambous, P., Kirz, J., and Sayre, D. (1999). Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400(6742):342–344.

Nesterov, Y. and Polyak, B. (2006). Cubic regularization of newton's method and its global performance. *Mathematical Programming*.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. (2017). Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR.

Reddi, S. J., Hefny, A., Sra, S., Póczos, B., and Smola, A. (2016). Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR.

Reisizadeh, A., Li, H., Das, S., and Jadbabaie, A. (2023). Variance-reduced clipping for non-convex optimization. *ArXiv:2303.00883*.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. (2019). Spiderboost and momentum: Faster variance reduction algorithms. *Advances in Neural Information Processing Systems*, 32.

Yang, X., Zhang, H., Chen, W., and Liu, T.-Y. (2022). Normalized/clipped sgd with perturbation for differentially private non-convex optimization. *ArXiv:2206.13033*.

Zhang, B., Jin, J., Fang, C., and Wang, L. (2020). Improved analysis of clipping algorithms for non-convex optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 15511–15521.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. (2019). Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*.

Zhao, S.-Y., Xie, Y.-P., and Li, W.-J. (2021). On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64(3):1–13.

Zhou, D., Xu, P., and Gu, Q. (2020). Stochastic nested variance reduction for nonconvex optimization. *The Journal of Machine Learning Research*, 21(1):4130–4192.

Appendix

Table of Contents

A. Supporting Lemmas

Lemma 2. *$f\in\mathcal{L}_{sym}^*(\alpha)$ if and only if for any $w,w\in\mathbb{R}^d$,*

$$\|\nabla f(w')-\nabla f(w)\|\le\Big(L_0+L_1\int_0^1\|\nabla f(w_\theta)\|^\alpha d\theta\Big)\|w'-w\| \tag{25}$$

where $w_\theta:=\theta w'+(1-\theta)w$.

Lemma 2 provides an equivalent definition of $f\in\mathcal{L}_{\text{asym}}^*(1)$ which is sometimes more convenient to use than Definition 3, for example, in the proof in Section B.2.

Proof. Eq. (25) directly implies eq. (4) (i.e., $f\in\mathcal{L}_{\text{sym}}^*(\alpha)$) since

$$\int_0^1\|\nabla f(w_\theta)\|^\alpha d\theta\le\max_{\theta\in[0,1]}\|\nabla f(w_\theta)\|^\alpha.$$

Then it remains to prove eq. (25) given eq. (4). For any $n\in\mathbb{N}^+$, we have

$$\begin{aligned}&\|\nabla f(w')-\nabla f(w)\|\\&\le\sum_{k=0}^{n-1}\|\nabla f(w_{(k+1)/n})-\nabla f(w_{k/n})\|\\&\overset{(i)}{\le}\sum_{k=0}^{n-1}\|w_{(k+1)/n}-w_{k/n}\|\Big(L_0+L_1\max_{\theta\in[0,1]}\|\nabla f(w_{\theta(k+1)/n+(1-\theta)k/n})\|^\alpha\Big)\\&\overset{(ii)}{=}\|w'-w\|\sum_{k=0}^{n-1}\frac{1}{n}\max_{\theta\in[k/n,(k+1)/n]}h(\theta)\end{aligned}$$

where (i) uses eq. (4) with w,w' replaced by $w_{k/n},w_{(k+1)/n}$ respectively ($w_{k/n}$ and $w_{(k+1)/n}$ denote w_θ with $\theta=k/n$ and $\theta=(k+1)/n$ respectively) and (ii) denotes $h(\theta):=L_0+L_1\|\nabla f(w_\theta)\|^\alpha$. Since $h(\cdot)$ is continuous, letting $n\to+\infty$ in the above inequality proves eq. (25) as follows.

$$\|\nabla f(w')-\nabla f(w)\|\le\|w'-w\|\int_0^1h(\theta)d\theta=\Big(L_0+L_1\int_0^1\|\nabla f(w_\theta)\|^\alpha d\theta\Big)\|w'-w\|$$

□

Lemma 3. *$f\in\mathbb{E}\mathcal{L}_{sym}^*(\alpha)$ if and only if for any $w,w\in\mathbb{R}^d$,*

$$\mathbb{E}_\xi\|\nabla f_\xi(w')-\nabla f_\xi(w)\|^2\le\|w'-w\|^2\mathbb{E}_\xi\int_0^1\big(L_0+L_1\|\nabla f_\xi(w_\theta)\|^\alpha\big)^2d\theta \tag{26}$$

where $w_\theta:=\theta w'+(1-\theta)w$.

Lemma 3 provides an equivalent definition of $f\in\mathbb{E}\mathcal{L}_{\text{asym}}^*(1)$ which is sometimes more convenient to use than Definition 4.

Proof. It sufficies to prove the equivalence between eqs. (26) & (19).

Eq. (26) directly implies eq. (19) since

$$\int_0^1\big(L_0+L_1\|\nabla f_\xi(w_\theta)\|^\alpha\big)^2d\theta\le\max_{\theta\in[0,1]}\big(L_0+L_1\|\nabla f_\xi(w_\theta)\|^\alpha\big)^2=\big(L_0+L_1\max_{\theta\in[0,1]}\|\nabla f_\xi(w_\theta)\|^\alpha\big)^2.$$

Then it remains to prove eq. (26) given eq. (19). For any $w,w'\in\mathbb{R}^d$, denote $w_\theta:=\theta w'+(1-\theta)w$. Then for any $\theta\in[0,1]$ and $n\in\mathbb{N}^+$, we have

$$\mathbb{E}_\xi\|\nabla f_\xi(w_\theta)-\nabla f_\xi(w)\|^2$$

$$
\begin{aligned}
&= \mathbb{E}_\xi \Big\| \sum_{k=0}^{n-1} \big(\nabla f_\xi(w_{\theta(k+1)/n}) - \nabla f_\xi(w_{\theta k/n})\big) \Big\|^2 \\
&\overset{(i)}{\le} n \sum_{k=0}^{n-1} \mathbb{E}_\xi \big\| \nabla f_\xi(w_{\theta(k+1)/n}) - \nabla f_\xi(w_{\theta k/n}) \big\|^2 \\
&\overset{(ii)}{\le} n \sum_{k=0}^{n-1} \|w_{\theta(k+1)/n} - w_{\theta k/n}\|^2 \mathbb{E}_\xi \Big(L_0 + L_1 \max_{\theta' \in [0,1]} \|\nabla f_\xi\big(\theta' w_{\theta(k+1)/n} + (1-\theta') w_{\theta k/n}\big)\|^\alpha \Big)^2 \\
&= \theta^2 \|w' - w\|^2 \mathbb{E}_\xi \sum_{k=0}^{n-1} \frac{1}{n} \max_{\theta' \in [0,1]} \big(L_0 + L_1 \|\nabla f_\xi\big(w_{\theta'\theta(k+1)/n + (1-\theta')\theta k/n}\big)\|^\alpha \big)^2 \\
&\overset{(iii)}{=} \theta^2 \|w' - w\|^2 \mathbb{E}_\xi \sum_{k=0}^{n-1} \frac{1}{n} \max_{u \in [k/n, (k+1)/n]} h(u)
\end{aligned}
$$

where (i) applies Jensen' inequality to the convex function $\|\cdot\|^2$, (ii) uses eq. (19), and (iii) denotes $h(u) := \big(L_0 + L_1 \|\nabla f_\xi(w_{\theta u})\|^\alpha\big)^2$. Since h is a continuous function, letting $n \to +\infty$ in the above inequality yields that

$$
\mathbb{E}_\xi \|\nabla f_\xi(w_\theta) - \nabla f_\xi(w)\|^2 \le \theta^2 \|w' - w\|^2 \mathbb{E}_\xi \int_0^1 h(u) du \le \theta^2 \|w' - w\|^2 \mathbb{E}_\xi \int_0^1 \big(L_0 + L_1 \|\nabla f_\xi(w_{\theta u})\|^\alpha\big)^2 du. \tag{27}
$$

Substituting $\theta = 1$ into the above inequality proves eq. (26). □

Lemma 4. *Under Assumption 1, the stochastic gradient $\nabla f_\xi(w)$ and true gradient $\nabla f(w)$ satisfy the following inequalities for any $\tau \in [0, 2]$,*

$$
\mathbb{E}_{\xi \sim \mathbb{P}} \|\nabla f_\xi(w)\|^\tau \le (\Gamma^\tau + 1) \|\nabla f(w)\|^\tau + \Lambda^\tau. \tag{28}
$$

Proof. First, when $\tau = 2$, Assumption 1 implies eq. (28) as follows.

$$
\mathbb{E}_{\xi \sim \mathbb{P}} \|\nabla f_\xi(w)\|^2 \overset{(i)}{=} \mathbb{E}_{\xi \sim \mathbb{P}} \|\nabla f_\xi(w) - \nabla f(w)\|^2 + \|\nabla f(w)\|^2 \le (\Gamma^2 + 1) \|\nabla f(w)\|^2 + \Lambda^2 \tag{29}
$$

where (i) uses $f(w) = \mathbb{E}_\xi f_\xi$. Then, when $\tau \in [0, 2)$, we prove eq. (28) as follows.

$$
\mathbb{E}_{\xi \sim \mathbb{P}} \|\nabla f_\xi(w)\|^\tau \overset{(i)}{\le} (\mathbb{E}_{\xi \sim \mathbb{P}} \|\nabla f_\xi(w)\|^2)^{\tau/2} \overset{(ii)}{\le} \big((\Gamma^2 + 1)\|\nabla f(w)\|^2 + \Lambda^2\big)^{\tau/2} \overset{(iii)}{=} (\Gamma^\tau + 1) \|\nabla f(w)\|^\tau + \Lambda^\tau,
$$

where (i) applies Jensen's inequality to the concave function $g(s) = s^{\tau/2}$, (ii) uses eq. (29), and (iii) uses the inequality that $(a + b)^{\tau/2} \le a^{\tau/2} + b^{\tau/2}$ for any $a, b \ge 0$ and $\tau/2 \in [0, 1]$. □

Note that the only randomness of Algorithm 2 comes from S_t, so we can consider the filtration $\mathcal{F}(S_{1:t}) := \mathcal{F}(S_1, \ldots, S_t)$ which monotonically increases with larger t. Then, it can be easily seen from Algorithm 2 that

$$
v_t, w_{t+1}, \delta_{t+1} \in \mathcal{F}(S_{1:t}) / \mathcal{F}(S_{1:(t-1)}) \tag{30}
$$

B. Proof of Theorem 1

B.1. Proof of Item 1

On one hand, $f \in \mathcal{L}^*_{\text{asym}}$ means $\|\nabla f(w') - \nabla f(w)\| \le \big(L_0 + L_1 \|\nabla f(w')\|\big) \|w' - w\|$ for all w, w', which directly implies eq. (4) with $\alpha = 1$, i.e., $f \in \mathcal{L}^*_{\text{asym}}(1)$. On the other hand, we will prove item 4 which shows that $f(w) = e^w + e^{-w}, w \in \mathbb{R}$ belongs to $\mathcal{L}^*_{\text{sym}}(1)$ but not $\mathcal{L}^*_{\text{asym}}$. Therefore, $\mathcal{L}^*_{\text{asym}} \subset \mathcal{L}^*_{\text{sym}}(1)$.

B.2. Proof of Item 2

Note that if a function is not twice-differentiable, it cannot belong to $\mathcal{L}_{\rm H}^*$ but may still belong to $\mathcal{L}_{\rm sym}^*(1)$. For example, for the function $f(w)=w|w|$ whose derivative $f'(w)=2|w|$ is not differentiable (so $f\notin\mathcal{L}_{\rm H}^*$), we have $f\in\mathcal{L}\subset\mathcal{L}_{\rm sym}^*(1)$ since $|f'(w')-f'(w)|\le 2\big||w'|-|w|\big|\le 2|w'-w|$.

Therefore, it remains to prove for twice-differentiable functions f the equivalence between eq. (31) below (definition of $\mathcal{L}_{\rm H}^*$) and eq. (25) with $\alpha=1$ (equivalent definition of $\mathcal{L}_{\rm sym}^*(1)$).

$$\|\nabla^2 f(w')\|\le L_0+L_1\|\nabla f(w)\|. \tag{31}$$

Eq. (31) implies eq. (25) as proved below.

$$\begin{aligned}\|\nabla f(w')-\nabla f(w)\| &= \Big\|\int_0^1 \nabla^2 f(w_\theta)(w'-w)d\theta\Big\| \\ &\le \int_0^1 \|\nabla^2 f(w_\theta)\|\|w'-w\|d\theta \\ &\overset{(i)}{\le} \|w'-w\|\int_0^1 \big(L_0+L_1\|\nabla f(w_\theta)\|^\alpha\big)d\theta \\ &= \|w'-w\|\Big(L_0+L_1\int_0^1 \|\nabla f(w_\theta)\|^\alpha d\theta\Big),\end{aligned}$$

where (i) uses eq. (31). Finally, it remains prove eq. (31) given eq. (25).

Note that of the symmetric Hessian matrix $\nabla^2 f(w)$ has eigenvalue $\|\nabla^2 f(w)\|$ or $-\|\nabla^2 f(w)\|$. Denote s as the corresponding eigenvector with $\|s\|=1$, i.e., $\nabla^2 f(w)s=\pm\|\nabla^2 f(w)\|s$. In eq. (25) , we adopt $w':=w+\theta' s$ ($\theta'\in(0,1)$), so $w_\theta:=\theta w'+(1-\theta)w=\theta(w+\theta' s)+(1-\theta)w=w+\theta\theta' s$ and thus eq. (25) becomes

$$\|\nabla f(w+\theta' s)-\nabla f(w)\| \le \theta'\Big(L_0+L_1\int_0^1 \|\nabla f(w+\theta\theta' s)\|^\alpha d\theta\Big) \tag{32}$$

The left side of eq. (32) can be rewritten as follows.

$$\begin{aligned}\|\nabla f(w+\theta' s)-\nabla f(w)\| &= \theta'\Big\|\int_0^1 \nabla^2 f\big(\theta(w+\theta' s)+(1-\theta)w\big)d\theta\Big\| \\ &= \Big\|\int_0^1 \nabla^2 f(w+\theta\theta' s)\theta' d\theta\Big\| \\ &\overset{(i)}{=} \Big\|\int_0^{\theta'} \nabla^2 f(w+us)sdu\Big\|,\end{aligned} \tag{33}$$

where (i) uses change of variables $u=\theta'\theta$. The right side of eq. (32) can be rewritten as follows.

$$\theta'\Big(L_0+L_1\int_0^1 \|\nabla f(w+\theta\theta' s)\|^\alpha d\theta\Big) \overset{(i)}{=} L_0\theta'+L_1\int_0^{\theta'} \|\nabla f(w+us)\|^\alpha du, \tag{34}$$

where (i) also uses change of variables $u=\theta'\theta$. Substituting eqs. (33) & (34) into eq. (32) and multiplying both sides by $1/\theta'>0$, we obtain that

$$\Big\|\frac{1}{\theta'}\int_0^{\theta'} \nabla^2 f(w+us)sdu\Big\| \le L_0+\frac{L_1}{\theta'}\int_0^{\theta'} \|\nabla f(w+us)\|^\alpha du.$$

Letting $\theta'\to+0$ in the above inequality, we obtain eq. (31) as follows.

$$\|\nabla^2 f(w)\|=\|\nabla^2 f(w)s\|\le L_0+L_1\|\nabla f(w)\|^\alpha.$$

B.3. Proof of Item 3

The polynomial function $f(w) = |w|^{\frac{2-\alpha}{1-\alpha}}, w \in \mathbb{R}$ is twice-differentiable with first and second order derivatives below.

$$f'(w) = \frac{2-\alpha}{1-\alpha}|w|^{\frac{1}{1-\alpha}}\text{sgn}(w), \quad f''(w) = \frac{2-\alpha}{(1-\alpha)^2}|w|^{\frac{\alpha}{1-\alpha}}.$$

Therefore, for any $w, w' \in \mathbb{R}$, we have

$$\begin{aligned}
|f'(w') - f'(w)| &\le |w' - w| \max_{\theta\in[0,1]} |f''(w_\theta)| \\
&\le \frac{2-\alpha}{(1-\alpha)^2}|w' - w| \max_{\theta\in[0,1]} |w_\theta|^{\frac{\alpha}{1-\alpha}} \cdot 1 \\
&= \frac{2-\alpha}{(1-\alpha)^2}|w' - w| \max_{\theta\in[0,1]} \Big|\frac{1-\alpha}{2-\alpha} f'(w_\theta)\Big|^{\alpha} \\
&\le \frac{(2-\alpha)^{1-\alpha}}{(1-\alpha)^{2-\alpha}}|w' - w| \max_{\theta\in[0,1]} |f'(w_\theta)|^{\alpha} \qquad (35)
\end{aligned}$$

where $w_\theta := \theta w' + (1-\theta)w$. This verifies eq. (4) and thus proves that $f \in \mathcal{L}_{\text{sym}}(\alpha)$.

Next, we prove that $f \notin \mathcal{L}^*_{\text{sym}}(\widetilde{\alpha})$ for all $\widetilde{\alpha} \in (0, \alpha)$. Suppose $f \in \mathcal{L}^*_{\text{sym}}(\widetilde{\alpha})$, i.e., the following inequality holds for all $w, w' \in \mathbb{R}^d$.

$$|f'(w') - f'(w)| \le |w' - w|\big(L_0 + L_1 \max_{\theta\in[0,1]} |f'(w_\theta)|^{\widetilde{\alpha}}\big),$$

where $w_\theta := \theta w' + (1-\theta)w$. Substituting $w' = 0$ into the above inequality, we obtain the following inequality for all $w \in \mathbb{R}^d$.

$$\frac{2-\alpha}{1-\alpha}|w|^{\frac{1}{1-\alpha}} \le |w|\Big(L_0 + L_1\Big(\frac{2-\alpha}{1-\alpha}|w|^{\frac{1}{1-\alpha}}\Big)^{\widetilde{\alpha}}\Big).$$

As $|w| \to +\infty$, the left side of the above inequality is $\mathcal{O}(|w|^{\frac{1}{1-\alpha}})$ whereas the right side has strictly smaller order $\mathcal{O}(|w|^{\frac{1-\alpha+\widetilde{\alpha}}{1-\alpha}})$. Hence, the above inequality cannot hold for sufficiently large $|w|$, which means the assumption that $f \in \mathcal{L}^*_{\text{sym}}(\widetilde{\alpha})$ does not hold.

Finally, we prove that $f \notin \mathcal{L}^*_{\text{asym}}$. Suppose $f \in \mathcal{L}^*_{\text{asym}}$, i.e., the following inequality holds for all $w, w' \in \mathbb{R}^d$.

$$|f'(w') - f'(w)| \le |w' - w|\big(L_0 + L_1|f'(w)|\big).$$

Substituting $w = 0$ into the above inequality, we obtain the following inequality for all $w' \in \mathbb{R}^d$.

$$\frac{2-\alpha}{1-\alpha}|w'|^{\frac{1}{1-\alpha}} \le L_0|w'|,$$

which implies that $|w'| \le \big(\frac{L_0(1-\alpha)}{2-\alpha}\big)^{\frac{1-\alpha}{\alpha}} < +\infty$. Hence, the above inequality cannot for all sufficiently large $|w'|$, which means the assumption that $f \in \mathcal{L}^*_{\text{asym}}$ does not hold.

B.4. Proof of Item 4

The exponential function $f(w) = e^w + e^{-w}, w \in \mathbb{R}$ is twice-differentiable with first and second order derivatives below.

$$f'(w) = e^w - e^{-w} = \text{sgn}(w)\big(e^{|w|} - e^{-|w|}\big), \quad f''(w) = e^w + e^{-w} = e^{|w|} + e^{-|w|}.$$

When $|w| \le 1$, $|f''(w)| \le e + e^{-1} < 4$; When $|w| > 1$, $|f'(w)| + 4 = e^{|w|} + e^{-|w|} - 2e^{-|w|} + 4 > |f''(w)| - 2e^{-1} + 4 > |f''(w)|$. Combining the two cases yields that $|f''(w)| < |f'(w)| + 4$, which implies that $f \in \mathcal{L}^*_{\text{sym}}$. Since f is twice-differentiable, we have $f \in \mathcal{L}^*_{\text{sym}}(1)$ based on item 2 of Theorem 1.

Next, we prove that $f\notin\mathcal{L}^*_{\text{sym}}(\widetilde{\alpha})$ for all $\widetilde{\alpha}\in(0,1)$. Suppose $f\in\mathcal{L}^*_{\text{sym}}(\widetilde{\alpha})$, i.e., the following inequality holds for all $w,w'\in\mathbb{R}^d$.

$$|f'(w')-f'(w)|\le|w'-w|\big(L_0+L_1\max_{\theta\in[0,1]}|f'(w_\theta)|^{\widetilde{\alpha}}\big),$$

where $w_\theta:=\theta w'+(1-\theta)w$. Substituting $w'=0$ into the above inequality, we obtain the following inequality.

$$e^{|w|}-e^{-|w|}\le|w|\big(L_0+L_1(e^{|w|}-e^{-|w|})^{\widetilde{\alpha}}\big),\forall w\in\mathbb{R}^d,$$

which implies that

$$\frac{(e^{|w|}-e^{-|w|})^{1-\widetilde{\alpha}}}{|w|}\le\frac{L_0+L_1(e^{|w|}-e^{-|w|})^{\widetilde{\alpha}}}{(e^{|w|}-e^{-|w|})^{\widetilde{\alpha}}},\forall w\in\mathbb{R}^d/\{\mathbf{0}\}.$$

As $|w|\to+\infty$, the left side of the above inequality goes to $+\infty$ while the right sides converges to $L_1<+\infty$. Hence, the above inequality cannot hold for sufficiently large $|w|$, which means the assumption that $f\in\mathcal{L}^*_{\text{sym}}(\widetilde{\alpha})$ does not hold.

Finally, we prove that $f\notin\mathcal{L}^*_{\text{asym}}$. Suppose $f\in\mathcal{L}^*_{\text{asym}}$, i.e., the following inequality holds for all $w,w'\in\mathbb{R}^d$.

$$|f'(w')-f'(w)|\le|w'-w|\big(L_0+L_1|f'(w)|\big).$$

Substituting $w=0$ into the above inequality and rearranging it, we obtain the following inequality for all $w'\in\mathbb{R}^d/\{\mathbf{0}\}$.

$$\frac{e^{|w'|}-e^{-|w'|}}{|w'|}\le L_0,$$

As $|w|\to+\infty$, the left side of the above inequality goes to $+\infty$, so the above inequality cannot for all sufficiently large $|w'|$, which means the assumption that $f\in\mathcal{L}^*_{\text{asym}}$ does not hold.

C. Proof of Proposition 1

C.1. Proof of Item 1

First, we prove eq. (5) for $f\in\mathcal{L}^*_{\text{sym}}(\alpha)$ with $\alpha\in(0,1)$. Note that eq. (25) holds for all $w,w'\in\mathbb{R}^d$. Hence, for any $\theta'\in[0,1]$, we can replace w' with $w_{\theta'}:=\theta'w'+(1-\theta')w$ in eq. (25), so w_θ becomes $\theta'w_\theta+(1-\theta)w=\theta'\theta w'+(1-\theta'\theta)w=w_{\theta'\theta}$. Therefore, eq. (25) becomes

$$\begin{aligned}\|\nabla f(w_{\theta'})-\nabla f(w)\|&\le\Big(L_0+L_1\int_0^1\|\nabla f(w_{\theta'\theta})\|^\alpha d\theta\Big)\|w_{\theta'}-w\|\\&=\Big(L_0\theta'+L_1\int_0^1\|\nabla f(w_{\theta'\theta})\|^\alpha\theta' d\theta\Big)\|w'-w\|\\&\overset{(i)}{=}H(\theta')\|w'-w\|\end{aligned}\tag{36}$$

where (i) denotes $H(\theta'):=L_0\theta'+L_1\int_0^1\|\nabla f(w_{\theta'\theta})\|^\alpha\theta'd\theta=L_0\theta'+L_1\int_0^{\theta'}\|\nabla f(w_u)\|^\alpha du$. Then its derivative $H'(\theta)$ can be bounded as follows,

$$\begin{aligned}H'(\theta')&=L_0+L_1\|\nabla f(w_{\theta'})\|^\alpha\\&\le L_0+L_1\|\nabla f(w_{\theta'})-\nabla f(w)\|^\alpha+L_1\|\nabla f(w)\|^\alpha\\&\overset{(i)}{\le}L_0+L_1\|w'-w\|^\alpha H(\theta')^\alpha+L_1\|\nabla f(w)\|^\alpha\\&\overset{(ii)}{\le}3L_1\Big(\frac{1}{3}\|w'-w\|H(\theta')+\frac{1}{3}\|\nabla f(w)\|+\frac{L_0^{\frac{1}{\alpha}}}{3L_1^{\frac{1}{\alpha}}}\Big)^\alpha.\end{aligned}\tag{37}$$

where (i) uses eq. (36) and (ii) applies Jensen's inequality to the concave function $g(x) = x^{\alpha}$. Rearranging the above inequality yields that

$$\begin{aligned}
3^{1-\alpha}L_1(1-\alpha)\|w'-w\| &\ge (1-\alpha)\|w'-w\|\Big(\|w'-w\|H(\theta')+\|\nabla f(w)\|+\frac{L_0^{\frac{1}{\alpha}}}{L_1^{\frac{1}{\alpha}}}\Big)^{-\alpha}H'(\theta')\\
&=\frac{d}{d\theta'}\Big(\|w'-w\|H(\theta')+\|\nabla f(w)\|+\frac{L_0^{\frac{1}{\alpha}}}{L_1^{\frac{1}{\alpha}}}\Big)^{1-\alpha}.
\end{aligned}$$

Integrating the above inequality over $\theta' \in [0,\theta]$ yields that

$$\begin{aligned}
&\Big(\|w'-w\|H(\theta)+\|\nabla f(w)\|+\frac{L_0^{\frac{1}{\alpha}}}{L_1^{\frac{1}{\alpha}}}\Big)^{1-\alpha}\\
&\le 3^{1-\alpha}L_1(1-\alpha)\|w'-w\|\theta+\Big(\|w'-w\|H(0)+\|\nabla f(w)\|+\frac{L_0^{\frac{1}{\alpha}}}{L_1^{\frac{1}{\alpha}}}\Big)^{1-\alpha}\\
&\overset{(i)}{\le} 2^{\alpha}\Big(3\big(L_1(1-\alpha)\|w'-w\|\theta\big)^{\frac{1}{1-\alpha}}+\|\nabla f(w)\|+\frac{L_0^{\frac{1}{\alpha}}}{L_1^{\frac{1}{\alpha}}}\Big)^{1-\alpha}
\end{aligned}$$

where (i) uses $H(0) = 0$ and applies Jensen's inequality to the concave function $g(x) = x^{1-\alpha}$. Therefore,

$$\|w'-w\|H(\theta)\le 2^{\frac{\alpha}{1-\alpha}}\Big(3\big(L_1(1-\alpha)\|w'-w\|\theta\big)^{\frac{1}{1-\alpha}}+\|\nabla f(w)\|+\frac{L_0^{\frac{1}{\alpha}}}{L_1^{\frac{1}{\alpha}}}\Big)-\|\nabla f(w)\|-\frac{L_0^{\frac{1}{\alpha}}}{L_1^{\frac{1}{\alpha}}}.$$

Substituting the above inequality into eq. (36), we obtain that

$$\begin{aligned}
\|\nabla f(w_\theta)\| &\le \|\nabla f(w)\|+\|\nabla f(w_\theta)-\nabla f(w)\|\\
&\le \|\nabla f(w)\|+\|w'-w\|H(\theta)\\
&\le 2^{\frac{\alpha}{1-\alpha}}\Big(3\big(L_1(1-\alpha)\|w'-w\|\theta\big)^{\frac{1}{1-\alpha}}+\|\nabla f(w)\|+\frac{L_0^{\frac{1}{\alpha}}}{L_1^{\frac{1}{\alpha}}}\Big).
\end{aligned}$$

Then, substituting the above inequality into eq. (4), we obtain that

$$\begin{aligned}
&\|\nabla f(w')-\nabla f(w)\|\\
&\le \big(L_0+L_1\max_{\theta\in[0,1]}\|\nabla f(w_\theta)\|^{\alpha}\big)\|w'-w\|\\
&\le \Big(L_0+L_1\cdot 2^{\frac{\alpha^2}{1-\alpha}}\Big(3\big(L_1(1-\alpha)\|w'-w\|\big)^{\frac{1}{1-\alpha}}+\|\nabla f(w)\|+\frac{L_0^{\frac{1}{\alpha}}}{L_1^{\frac{1}{\alpha}}}\Big)^{\alpha}\Big)\|w'-w\|\\
&\overset{(i)}{\le} \Big(L_0+L_1\cdot 2^{\frac{\alpha^2}{1-\alpha}}\Big(3^{\alpha}\big(L_1(1-\alpha)\|w'-w\|\big)^{\frac{\alpha}{1-\alpha}}+\|\nabla f(w)\|^{\alpha}+\frac{L_0}{L_1}\Big)\Big)\|w'-w\|\\
&= \|w'-w\|\big(K_0+K_1\|\nabla f(w)\|^{\alpha}+K_2\|w'-w\|^{\frac{\alpha}{1-\alpha}}\big)
\end{aligned}$$

where (i) uses the inequality that $(a+b+c)^{\alpha} \le a^{\alpha}+b^{\alpha}+c^{\alpha}$ for any $a,b,c \ge 0$ and $\alpha \in [0,1]$, and (ii) denotes that $K_0 := L_0\big(2^{\frac{\alpha^2}{1-\alpha}}+1\big)$, $K_1 := L_1\cdot 2^{\frac{\alpha^2}{1-\alpha}}\cdot 3^{\alpha}$, $K_2 := L_1^{\frac{1}{1-\alpha}}\cdot 2^{\frac{\alpha^2}{1-\alpha}}\cdot 3^{\alpha}(1-\alpha)^{\frac{\alpha}{1-\alpha}}$.

Next, we prove $f \in \mathcal{L}^*_{\text{sym}}(\alpha)$ given eq. (5). For any $w, w' \in \mathbb{R}^d$ and $n \in \mathbb{N}^+$, we have

$$\begin{aligned}
&\|\nabla f(w')-\nabla f(w)\|\\
&\overset{(i)}{\le}\sum_{k=0}^{n-1}\|\nabla f(w_{(k+1)/n})-\nabla f(w_{k/n})\|
\end{aligned}$$

$$
\begin{aligned}
&\overset{(ii)}{\leq} \sum_{k=0}^{n-1} \|w_{(k+1)/n} - w_{k/n}\| \big(K_0 + K_1\|\nabla f(w_{k/n})\|^{\alpha} + K_2\|w_{(k+1)/n} - w_{k/n}\|^{\frac{\alpha}{1-\alpha}}\big)\\
&\overset{(iii)}{=} \|w' - w\| \sum_{k=0}^{n-1} \Big(\frac{1}{n}h\Big(\frac{k}{n}\Big) + \frac{1}{n}\cdot\frac{K_2}{n^{\frac{\alpha}{1-\alpha}}}\|w' - w\|^{\frac{\alpha}{1-\alpha}}\Big)\\
&= \|w' - w\|\Big(\frac{K_2}{n^{\frac{\alpha}{1-\alpha}}}\|w' - w\|^{\frac{\alpha}{1-\alpha}} + \sum_{k=0}^{n-1}\frac{1}{n}h\Big(\frac{k}{n}\Big)\Big),
\end{aligned}
$$

where (i) denotes $w_\theta := \theta w' + (1-\theta)w$, (ii) uses eq. (5) with w, w' replaced by $w_{k/n}, w_{(k+1)/n}$ respectively and (iii) denotes $h(\theta) := K_0 + K_1\|\nabla f(w_\theta)\|^{\alpha}$. Since $h(\cdot)$ is continuous, letting $n \to +\infty$ in the above inequality proves eq. (25) as follows, which implies $f \in \mathcal{L}^*_{\text{sym}}(\alpha)$ by Lemma 2.

$$
\|\nabla f(w') - \nabla f(w)\| \leq \|w' - w\| \int_0^1 h(\theta)d\theta = \Big(L_0 + L_1\int_0^1 \|\nabla f(w_\theta)\|^{\alpha}d\theta\Big)\|w' - w\|
$$

C.2. Proof of Item 2

Note that eq. (37) holds for any function $f \in \mathcal{L}^*_{\text{sym}}(\alpha)$ with $\alpha \in [0,1]$. Substituting $\alpha = 1$ into eq. (37), we obtain that

$$
H'(\theta) \leq L_0 + L_1\|w' - w\|H(\theta) + L_1\|\nabla f(w)\|,
$$

where $H(\theta') := L_0\theta' + L_1\int_0^{\theta'}\|\nabla f(w_u)\|du$. Rearranging the above inequality yields that

$$
L_1\|w' - w\| \geq \frac{L_1\|w' - w\|H'(\theta')}{L_0 + L_1\|w' - w\|H(\theta') + L_1\|\nabla f(w)\|} = \frac{d}{d\theta'}\ln\big(L_0 + L_1\|w' - w\|H(\theta') + L_1\|\nabla f(w)\|\big).
$$

Integrating the above inequality over $\theta' \in [0,\theta]$ yields that (note that $H(0) = 0$)

$$
\ln\big(L_0 + L_1\|w' - w\|H(\theta) + L_1\|\nabla f(w)\|\big) \leq \ln\big(L_0 + L_1\|\nabla f(w)\|\big) + L_1\|w' - w\|,
$$

which implies that

$$
L_1\|w' - w\|H(\theta) \leq \big(L_0 + L_1\|\nabla f(w)\|\big)\exp\big(L_1\|w' - w\|\big) - L_0 - L_1\|\nabla f(w)\|.
$$

Substituting the above inequality and $\alpha = 1$ into eq. (36), we obtain that

$$
\begin{aligned}
\|\nabla f(w_\theta)\| &\leq \|\nabla f(w)\| + \|\nabla f(w_\theta) - \nabla f(w)\|\\
&\leq \|\nabla f(w)\| + \|w' - w\|H(\theta)\\
&\leq \|\nabla f(w)\| + \frac{1}{L_1}\Big(\big(L_0 + L_1\|\nabla f(w)\|\big)\exp\big(L_1\|w' - w\|\big) - L_0 - L_1\|\nabla f(w)\|\Big)\\
&= \Big(\frac{L_0}{L_1} + \|\nabla f(w)\|\Big)\exp\big(L_1\|w' - w\|\big) - \frac{L_0}{L_1}
\end{aligned}
$$

Then, substituting the above inequality and $\alpha = 1$ into eq. (4), we prove eq. (6) as follows.

$$
\begin{aligned}
&\|\nabla f(w') - \nabla f(w)\|\\
&\leq \big(L_0 + L_1\max_{\theta\in[0,1]}\|\nabla f(w_\theta)\|\big)\|w' - w\|\\
&\leq \big(L_0 + L_1\|\nabla f(w)\|\big)\exp\big(L_1\|w' - w\|\big)\|w' - w\|
\end{aligned}
$$

Next, we prove $f \in \mathcal{L}^*_{\text{sym}}(\alpha)$ given eq. (6). For any $w, w' \in \mathbb{R}^d$ and $n \in \mathbb{N}^+$, we have

$$
\|\nabla f(w') - \nabla f(w)\|
$$

$$
\begin{aligned}
&\overset{(i)}{\le} \sum_{k=0}^{n-1} \|\nabla f(w_{(k+1)/n}) - \nabla f(w_{k/n})\| \\
&\overset{(ii)}{\le} \sum_{k=0}^{n-1} \|w_{(k+1)/n} - w_{k/n}\| \big(L_0 + L_1\|\nabla f(w_{k/n})\|\big) \exp\big(L_1\|w_{(k+1)/n} - w_{k/n}\|\big) \\
&\overset{(iii)}{=} \|w' - w\| \sum_{k=0}^{n-1} \frac{1}{n} h\Big(\frac{k}{n}\Big) \exp\Big(\frac{L_1}{n}\|w' - w\|\Big) \\
&= \|w' - w\| \sum_{k=0}^{n-1} \frac{1}{n} h\Big(\frac{k}{n}\Big) + \|w' - w\| \sum_{k=0}^{n-1} \frac{1}{n} h\Big(\frac{k}{n}\Big) \Big[\exp\Big(\frac{L_1}{n}\|w' - w\|\Big) - 1\Big] \\
&\le \|w' - w\| \sum_{k=0}^{n-1} \frac{1}{n} h\Big(\frac{k}{n}\Big) + \|w' - w\| \max_{\theta\in[0,1]} h(\theta) \Big[\exp\Big(\frac{L_1}{n}\|w' - w\|\Big) - 1\Big]
\end{aligned}
$$

where (i) denotes $w_\theta := \theta w' + (1-\theta)w$, (ii) uses eq. (6) with w, w' replaced by $w_{k/n}, w_{(k+1)/n}$ respectively and (iii) denotes $h(\theta) := L_0 + L_1\|\nabla f(w_\theta)\|$. Since $h(\cdot)$ is continuous, letting $n \to +\infty$ in the above inequality proves eq. (25) with $\alpha = 1$ as follows, which implies $f \in \mathcal{L}_{\text{sym}}^*(1)$ by Lemma 2.

$$
\|\nabla f(w') - \nabla f(w)\| \le \|w' - w\| \int_0^1 h(\theta) d\theta = \Big(L_0 + L_1 \int_0^1 \|\nabla f(w_\theta)\| d\theta\Big) \|w' - w\|
$$

C.3. Proof of Item 3

Since $f \in \mathcal{L}_{\text{sym}}^*(\alpha)$ for $\alpha \in (0,1)$, eq. (5) holds based on item 1 of Proposition 1. Hence, for any $w, w' \in \mathbb{R}^d$, we prove eq. (7) as follows.

$$
\begin{aligned}
& f(w') - f(w) - \nabla f(w)^\top (w' - w) \\
&= \int_0^1 \big(\nabla f(w_\theta) - \nabla f(w)\big)^\top (w' - w) d\theta \\
&\le \int_0^1 \|\nabla f(w_\theta) - \nabla f(w)\| \|w' - w\| d\theta \\
&\overset{(i)}{\le} \int_0^1 \|w_\theta - w\| \big(K_0 + K_1\|\nabla f(w)\|^\alpha + K_2\|w_\theta - w\|^{\frac{\alpha}{1-\alpha}}\big) \|w' - w\| d\theta \\
&= \int_0^1 \theta \|w' - w\|^2 \big(K_0 + K_1\|\nabla f(w)\|^\alpha + K_2 \theta^{\frac{\alpha}{1-\alpha}} \|w' - w\|^{\frac{\alpha}{1-\alpha}}\big) d\theta \\
&= \frac{1}{2}\|w' - w\|^2 \big(K_0 + K_1\|\nabla f(w)\|^\alpha\big) + K_2 \|w' - w\|^{\frac{2-\alpha}{1-\alpha}} \int_0^1 \theta^{\frac{1}{1-\alpha}} d\theta \\
&\le \frac{1}{2}\|w' - w\|^2 \big(K_0 + K_1\|\nabla f(w)\|^\alpha + 2K_2\|w' - w\|^{\frac{\alpha}{1-\alpha}}\big),
\end{aligned}
$$

where (i) uses eq. (5) with w' replaced by $w_\theta := \theta w' + (1-\theta)w$.

C.4. Proof of Item 4

Since $f \in \mathcal{L}_{\text{sym}}^*(\alpha)$ for $\alpha \in (0,1)$, eq. (6) holds based on item 2 of Proposition 1. Hence, for any $w, w' \in \mathbb{R}^d$, we prove eq. (7) as follows.

$$
\begin{aligned}
& f(w') - f(w) - \nabla f(w)^\top (w' - w) \\
&= \int_0^1 \big(\nabla f(w_\theta) - \nabla f(w)\big)^\top (w' - w) d\theta \\
&\le \int_0^1 \|\nabla f(w_\theta) - \nabla f(w)\| \|w' - w\| d\theta
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(i)}{\le} \int_0^1 \|w_\theta - w\|\big(L_0 + L_1\|\nabla f(w)\|\big)\exp\big(L_1\|w_\theta - w\|\big)\|w' - w\|d\theta \\
&\le \int_0^1 \theta\|w' - w\|^2\big(L_0 + L_1\|\nabla f(w)\|\big)\exp\big(L_1\|w' - w\|\big)d\theta \\
&= \frac{1}{2}\|w' - w\|^2\big(L_0 + L_1\|\nabla f(w)\|\big)\exp\big(L_1\|w' - w\|\big),
\end{aligned}
$$

where (i) uses eq. (6) with w' replaced by $w_\theta := \theta w' + (1-\theta)w$.

D. Proof of Proposition 2 and Proposition 3

D.1. Proof for Phase Retrieval Problem

The objective function (10) of phase retrieval problem can be rewritten in the stochastic form $f(z) = \mathbb{E}_\xi f_\xi(z)$ where ξ is obtained from $\{1, 2, \ldots, m\}$ uniformly at random and

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

To prove that $f \in \mathcal{L}^*_{\rm sym}(\frac{2}{3})$ and $f \in \mathbb{E}\mathcal{L}^*_{\rm sym}(\frac{2}{3})$ respectively required by Proposition 2 and Proposition 3, it suffices to prove that $f_\xi \in \mathcal{L}^*_{\rm sym}(\frac{2}{3})$ for every sample ξ.

For any $z \in \mathbb{R}^d$ and sample ξ, the gradient $\nabla f_\xi(z) = \frac{1}{2}(|a_\xi^\top z|^2 - y_\xi)(a_\xi a_\xi^\top)z$ satisfies

$$
\begin{aligned}
\|\nabla f_\xi(z)\|^{\frac{2}{3}} &= \frac{1}{2^{\frac{2}{3}}}\big\|(|a_\xi^\top z|^2 - y_\xi)(a_\xi a_\xi^\top)z\big\|^{\frac{2}{3}} \\
&\ge \frac{1}{2}\big||a_\xi^\top z|^3 - y_\xi|a_\xi^\top z|\big|^{\frac{2}{3}}\|a_\xi\|^{\frac{2}{3}} \\
&\overset{(ii)}{\ge} \frac{1}{2}\big(|a_\xi^\top z|^2 - |y_\xi||a_\xi^\top z|^{\frac{2}{3}}\big)\big|\|a_\xi\|^{\frac{2}{3}} \\
&\overset{(iii)}{\ge} \frac{1}{3}\big(|a_\xi^\top z|^2 - |y_\xi|^{\frac{3}{2}}\big)\big|\|a_\xi\|^{\frac{2}{3}}
\end{aligned} \tag{38}
$$

where (i) applies Jensen's inequality, (ii) uses the inequality that $|a-b|^{\frac{2}{3}} \ge |a|^{\frac{2}{3}} - |b|^{\frac{2}{3}}$ for any $a, b \in \mathbb{R}$, (iii) uses $|y_\xi|a^{\frac{2}{3}} \le \frac{1}{3}a^2 + \frac{2}{3}|y_\xi|^{\frac{3}{2}}$ for any $a \ge 0$ based on Young's inequality.

For any $z, z' \in \mathbb{R}^d$, we obtain the following inequality which proves that $f_\xi \in \mathcal{L}^*_{\rm sym}(\frac{2}{3})$ as desired.

$$
\begin{aligned}
&\|\nabla f_\xi(z') - \nabla f_\xi(z)\| \\
&= \frac{1}{2}\big\|(|a_\xi^\top z'|^2 - y_\xi)(a_\xi a_\xi^\top)z' - (|a_\xi^\top z|^2 - y_\xi)(a_\xi a_\xi^\top)z\big\| \\
&\le \frac{1}{4}\big\|(|a_\xi^\top z'|^2 + |a_\xi^\top z|^2 - 2y_\xi)(a_\xi a_\xi^\top)(z' - z) + (|a_\xi^\top z'|^2 - |a_\xi^\top z|^2)(a_\xi a_\xi^\top)(z' + z)\big\| \\
&\overset{(i)}{\le} \frac{1}{4}\|a_\xi\|^2\big(|a_\xi^\top z'|^2 + |a_\xi^\top z|^2 + 2|y_\xi|\big)\|z' - z\| + \frac{1}{4}\|a_\xi\|^2(|a_\xi^\top z'| + |a_\xi^\top z|)^2\|z' - z\| \\
&\overset{(ii)}{\le} \frac{1}{4}\|z' - z\|\|a_\xi\|^2\big(3|a_\xi^\top z'|^2 + 3|a_\xi^\top z|^2 + 2|y_\xi|\big) \\
&\le \frac{1}{4}\|z' - z\|\|a_\xi\|^{\frac{4}{3}}\|a_\xi\|^{\frac{2}{3}}\big(3|a_\xi^\top z'|^2 + 3|a_\xi^\top z|^2 - 3|y_\xi| - 3|y_\xi| + 8|y_\xi|\big) \\
&\overset{(iii)}{\le} \|z' - z\|\Big(\frac{9}{4}a_{\max}^{\frac{4}{3}}\|\nabla f_\xi(z')\|^{\frac{2}{3}} + \frac{9}{4}a_{\max}^{\frac{4}{3}}\|\nabla f_\xi(z)\|^{\frac{2}{3}} + 2y_{\max}a_{\max}^2\Big) \\
&\le \|z' - z\|\Big(\frac{9}{4}a_{\max}^{\frac{4}{3}}\max_{\theta\in[0,1]}\big\|\nabla f_\xi\big(\theta z' + (1-\theta)z\big)\big\|^{\frac{2}{3}} + 2y_{\max}a_{\max}^2\Big)
\end{aligned} \tag{39}
$$

where (i) uses trianagular inequality, $\|a_\xi a_\xi^\top\| = \|a_\xi\|^2$, $|y_\xi| \le 1$ and the following inequality, (ii) uses $(|a_\xi^\top z'| + |a_\xi^\top z|)^2 \le 2|a_\xi^\top z'|^2 + 2|a_\xi^\top z|^2$, (iii) uses eq. (38) and denotes that $y_{\max} := \max_{1\le r\le m}|y_r|$ and that $a_{\max} := \max_{1\le r\le m}\|a_r\|$.

$$\big||a_\xi^\top z'|^2 - |a_\xi^\top z|^2\big| = (|a_\xi^\top z'|+|a_\xi^\top z|)(|a_\xi^\top z'|-|a_\xi^\top z|) \le (|a_\xi^\top z'|+|a_\xi^\top z|)\|a_\xi^\top(z'-z)\| \le \|a_\xi^\top\|(|a_\xi^\top z'|+|a_\xi^\top z|)\|z'-z\|.$$

D.2. Proof for DRO Problem

We adopt the following assumptions from (Jin et al., 2021):

- ℓ_ξ is G-Lipschitz continuous and L-smooth.
- $\mathbb{E}\big(\ell_\xi(x) - \ell(x)\big)^2 \le \sigma^2$ where $\ell(x) := \mathbb{E}\ell_\xi(x)$
- ψ is a non-negative convex function with $\psi(1) = 0$ and $\psi(t) = +\infty$ for all $t < 0$, and ψ^* is M-smooth.

Then we rewrite the objective function (12) as $L(x,\eta) = \mathbb{E}L_\xi(x,\eta)$ where

$$L_\xi(x,\eta) := \lambda\psi^*\left(\frac{\ell_\xi(x)-\eta}{\lambda}\right) + \eta. \tag{40}$$

The gradient $\nabla L_\xi = \big[\nabla_x L_\xi; \frac{\partial}{\partial\eta}L_\xi\big]$ can be computed as follows.

To prove that $L \in \mathbb{E}\mathcal{L}^*_{\text{sym}}(1)$ required by Proposition 3, it suffices to prove that $L_\xi \in \mathcal{L}^*_{\text{sym}}(1)$ for every sample ξ.

$$\nabla_x L_\xi(x,\eta) = \psi^{*\prime}\Big(\frac{\ell_\xi(x)-\eta}{\lambda}\Big)\nabla\ell_\xi(x), \tag{41}$$

$$\frac{\partial}{\partial\eta}L_\xi(x,\eta) = 1 - \psi^{*\prime}\Big(\frac{\ell_\xi(x)-\eta}{\lambda}\Big). \tag{42}$$

Hence, for any $(x',\eta'), (x,\eta) \in \mathbb{R}^d \times \mathbb{R}$, $\nabla L_\xi(x',\eta') - \nabla L_\xi(x,\eta) = A + B$ where

$$A = \Big[\psi^{*\prime}\Big(\frac{\ell_\xi(x)-\eta}{\lambda}\Big)\big(\nabla\ell_\xi(x') - \nabla\ell_\xi(x)\big); 0\Big]$$

$$B = \Big[\psi^{*\prime}\Big(\frac{\ell_\xi(x')-\eta'}{\lambda}\Big) - \psi^{*\prime}\Big(\frac{\ell_\xi(x)-\eta}{\lambda}\Big)\Big]\big[\nabla\ell_\xi(x'); -1\big].$$

Therefore, we can prove that $L_\xi \in \mathcal{L}^*_{\text{sym}}(1)$ as follows.

$$\begin{aligned}
&\big\|\nabla L_\xi(x',\eta') - \nabla L_\xi(x,\eta)\big\| \\
&\le \|A\| + \|B\| \\
&\le \Big|\psi^{*\prime}\Big(\frac{\ell_\xi(x)-\eta}{\lambda}\Big)\Big|\|\nabla\ell_\xi(x') - \nabla\ell_\xi(x)\| + \Big|\psi^{*\prime}\Big(\frac{\ell_\xi(x')-\eta'}{\lambda}\Big) - \psi^{*\prime}\Big(\frac{\ell_\xi(x)-\eta}{\lambda}\Big)\Big|\sqrt{\|\nabla\ell_\xi(x')\|^2 + 1} \\
&\overset{(i)}{\le} \Big|1 - \frac{\partial}{\partial\eta}L_\xi(x,\eta)\Big|L\|x'-x\| + \frac{M}{\lambda}\big|\ell_\xi(x') - \eta' - \big(\ell_\xi(x) - \eta\big)\big|\sqrt{G^2+1} \\
&\overset{(ii)}{\le} \Big(L + L\Big|\frac{\partial}{\partial\eta}L_\xi(x,\eta)\Big|\Big)\|x'-x\| + \frac{M}{\lambda}\sqrt{G^2+1}\big[G\|x'-x\| + |\eta'-\eta|\big] \\
&\overset{(iii)}{\le} \Big(L + \frac{2M(G+1)^2}{\lambda} + L\|\nabla L_\xi(x,\eta)\|\Big)\|(x'-x, \eta'-\eta)\|
\end{aligned} \tag{43}$$

where (i) uses eq. (42), and the above assumptions that ℓ_ξ is G-Lipschitz, L-smooth and that ψ^* is M-smooth, (ii) uses the above assumptions that ℓ_ξ is G-Lipschitz, and (iii) uses $\|x'-x\| + \|\eta'-\eta\| \le \sqrt{2}\|(x'-x, \eta'-\eta)\|$ and $\Big|\frac{\partial}{\partial\eta}L_\xi(x,\eta)\Big| \le \|\nabla L_\xi(x,\eta)\|$.

E. Proof of Theorem 2

We will first prove the following lemma which will be used in the proof of Theorem 2.

Lemma 5. *For any $x \ge 0$, $C \in [0,1]$, $\Delta > 0$ and $0 \le \omega \le \omega'$ such that $\Delta \ge \omega' - \omega$, the following inequality holds*

$$Cx^{\omega} \le x^{\omega'} + C^{\frac{\omega'}{\Delta}} \tag{44}$$

Proof of Lemma 5. We consider three cases: $\omega = 0$, $\omega' = \omega > 0$ and $\omega' > \omega > 0$.

(Case I) When $\omega = 0$, $\Delta \ge \omega'$ and $\Delta > 0$ imply that $\frac{\omega'}{\Delta} \in [0,1]$, so $Cx^{\omega} = C \le C^{\frac{\omega'}{\Delta}}$, which implies eq. (44).

(Case II) When $\omega' = \omega > 0$, $Cx^{\omega} \le x^{\omega} = x^{\omega'}$, which implies eq. (44).

(Case III) When $\omega' > \omega > 0$, by applying Young's inequality with $p = \frac{\omega'}{\omega} > 1$ and $q = \frac{\omega'}{\omega' - \omega} > 1$ which satisfy $\frac{1}{p} + \frac{1}{q} = 1$, we prove eq. (44) as follows.

$$Cx^{\omega} \le \frac{x^{p\omega}}{p} + \frac{C^q}{q} \le x^{\omega'} + C^{\frac{\omega'}{\omega'-\omega}} \le x^{\omega'} + C^{\frac{\omega'}{\Delta}}.$$

□

Now we will prove Theorem 2. We omit the well-known case of $\beta = \alpha = 0$ where GD is applied to L-smooth function $f \in \mathcal{L}$. Hence, we focus on the case of $\beta > 0$. We first bound $f(w_{t+1}) - f(w_t)$ in two cases: $\alpha \in (0,1)$ and $\alpha = 1$.

(Case I) When $\alpha \in (0,1)$, eq. (7) holds for $f \in \mathcal{L}^*_{\text{sym}}(\alpha)$. Hence, we have

$$\begin{aligned}
&f(w_{t+1}) - f(w_t) \\
&\le \nabla f(w_t)^\top (w_{t+1} - w_t) + \frac{1}{2}\big(K_0 + K_1\|\nabla f(w_t)\|^{\alpha}\big)\|w_{t+1} - w_t\|^2 + K_2\|w_{t+1} - w_t\|^{\frac{2-\alpha}{1-\alpha}} \\
&\overset{(i)}{=} -\gamma\|\nabla f(w_t)\|^{2-\beta} + \frac{\gamma}{6}\big(3K_0\gamma \cdot \|\nabla f(w_t)\|^{2-2\beta} + 3K_1\gamma \cdot \|\nabla f(w_t)\|^{2+\alpha-2\beta} + 6K_2\gamma^{\frac{1}{1-\alpha}} \cdot \|\nabla f(w_t)\|^{\frac{(2-\alpha)(1-\beta)}{1-\alpha}}\big) \\
&\overset{(ii)}{\le} -\gamma\|\nabla f(w_t)\|^{2-\beta} + \frac{\gamma}{6}\big(3\|\nabla f(w_t)\|^{2-\beta} + (3K_0\gamma)^{\frac{2}{\beta}-1} + (3K_1\gamma)^{\frac{2}{\beta}-1} + (6K_2\gamma)^{\frac{2}{\beta}-1}\big) \\
&\overset{(iii)}{\le} -\frac{\gamma}{2}\|\nabla f(w_t)\|^{2-\beta} + \gamma^{\frac{2}{\beta}}(3K_0 + 3K_1 + 6K_2)^{\frac{2}{\beta}-1} \\
&\overset{(iv)}{\le} -\frac{\gamma}{2}\|\nabla f(w_t)\|^{2-\beta} + \frac{\gamma}{4}\epsilon^{2-\beta}
\end{aligned}$$

where (i) uses the update rule $w_{t+1} = w_t - \gamma\frac{\nabla f(w_t)}{\|\nabla f(w_t)\|^{\beta}}$ of Algorithm 1 (β-GD), (ii) uses $\gamma^{\frac{1}{1-\alpha}} \le 1$ and applies Lemma 5 three times respectively with $x = \|\nabla f(w_t)\|$, $C = 3K_0\gamma, 3K_1\gamma, 6K_2\gamma$ ($C \in [0,1]$ since $\gamma = \frac{\epsilon^{\beta}}{12(K_0+K_1+2K_2)+1}$ and $\epsilon \in (0,1)$), $\Delta = \beta, \omega = 2-2\beta, 2+\alpha-2\beta, \frac{(2-\alpha)(1-\beta)}{1-\alpha}, \omega' = 2-\beta$, (iii) uses the inequality that $a^{\tau} + b^{\tau} + c^{\tau} \le (a+b+c)^{\tau}$ for $\tau = \frac{2}{\beta} - 1 > 1$ and any $a, b, c \ge 0$, and (iv) uses $\gamma = \frac{\epsilon^{\beta}}{12(K_0+K_1+2K_2)+1}$.

(Case II) When $\alpha = 1$, we have $\beta = 1$ and eq. (8) holds for $f \in \mathcal{L}^*_{\text{sym}}(1)$. Hence, we have

$$\begin{aligned}
&f(w_{t+1}) - f(w_t) \\
&\le \nabla f(w_t)^\top (w_{t+1} - w_t) + \frac{1}{2}\|w_{t+1} - w_t\|^2\big(L_0 + L_1\|\nabla f(w_t)\|\big)\exp\big(L_1\|w_{t+1} - w_t\|\big) \\
&\overset{(i)}{=} -\gamma\|\nabla f(w_t)\| + \frac{\gamma^2}{2}\big(L_0 + L_1\|\nabla f(w_t)\|\big)\exp(L_1\gamma) \\
&\overset{(ii)}{\le} -\frac{\gamma}{2}\|\nabla f(w_t)\| + L_0\gamma^2 \\
&\overset{(iii)}{\le} -\frac{\gamma}{2}\|\nabla f(w_t)\|^{2-\beta} + \frac{\gamma}{4}\epsilon
\end{aligned} \tag{45}$$

where (i) uses the update rule $w_{t+1} = w_t - \gamma\frac{\nabla f(w_t)}{\|\nabla f(w_t)\|}$ of Algorithm 1 (β-GD with $\beta = 1$) and (ii) and (iii) use $\gamma = \frac{\epsilon}{4L_0+1} \le \frac{1}{2L_1}$. Note that eq. (45) holds in both cases. Therefore, by telescoping eq. (45) and rearranging it, we obtain that

$$\begin{aligned}\mathbb{E}_{\widetilde{T}}\|\nabla f(w_{\widetilde{T}})\|^{2-\beta} &= \frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|^{2-\beta}\\ &\le \frac{2}{T\gamma}\big(f(w_0) - f^*\big) + \frac{1}{2}\epsilon^{2-\beta},\end{aligned}$$

where (i) uses $\gamma = \frac{\epsilon^\beta}{12(K_0+K_1+2K_2)+1}$ and $f(w_T) \ge f^* := \min_{w\in\mathbb{R}^d} f(w)$. By applying Lyapunov inequality, the above inequality implies convergence rate (13) as follows.

$$\begin{aligned}\mathbb{E}_{\widetilde{T}}\|\nabla f(w_{\widetilde{T}})\| &\le \big(\mathbb{E}_{\widetilde{T}}\|\nabla f(w_{\widetilde{T}})\|^{2-\beta}\big)^{\frac{1}{2-\beta}}\\ &\le \Big(\frac{2}{T\gamma}\big(f(w_0) - f^*\big) + \frac{1}{2}\epsilon^{2-\beta}\Big)^{\frac{1}{2-\beta}},\\ &\overset{(i)}{\le} \Big(\frac{2}{T\gamma}\Big)^{\frac{1}{2-\beta}}\big(f(w_0) - f^*\big)^{\frac{1}{2-\beta}} + \Big(\frac{1}{2}\Big)^{\frac{1}{2-\beta}}\epsilon\\ &\le \Big(\frac{2}{T\gamma}\Big)^{\frac{1}{2-\beta}}\big(f(w_0) - f^*\big)^{\frac{1}{2-\beta}} + \frac{1}{2}\epsilon\end{aligned}$$

where (i) uses $(a+b)^\tau \le a^\tau + b^\tau$ for $\tau = \frac{1}{2-\beta} \in [0,1]$ and any $a, b \ge 0$.

Then, substituting $T = \frac{4}{\gamma}$ into the above convergence rate, we obtain that $\mathbb{E}_{\widetilde{T}}\|\nabla f(w_{\widetilde{T}})\| \le \epsilon$.

F. Proof of Theorem 3

We consider the following two cases.

(Case I) When $\alpha \in (0,1)$, consider the convex function $f(w) := |w|^{\frac{2-\alpha}{1-\alpha}}$ with unique minimizer $w = 0$ and derivative $f'(w) = \frac{2-\alpha}{1-\alpha}|w|^{\frac{1}{1-\alpha}}\mathrm{sgn}(w)$. Based on item 3 of Proposition 1, $f \in \mathcal{L}^*_{\text{sym}}(\alpha)$. Applying β-GD to this function yields that

$$w_{t+1} = w_t - \frac{\gamma f'(w_t)}{|f'(w_t)|^\beta} = w_t - \gamma\Big(\frac{2-\alpha}{1-\alpha}\Big)^{1-\beta}|w_t|^{\frac{1-\beta}{1-\alpha}}\mathrm{sgn}(w_t).$$

Note that $0 \le \beta < \alpha < 1$. Hence, if $|w_t| > C$ with constant $C := \Big(\frac{3(1-\alpha)}{\gamma(2-\alpha)}\Big)^{\frac{1-\alpha}{\alpha-\beta}} > 0$, we have $\gamma\frac{2-\alpha}{1-\alpha}|w_t|^{\frac{1-\beta}{1-\alpha}} > 3|w_t|$ and thus $|w_{t+1}| > 2|w_t|$. Therefore, if $|w_0| > C$, by induction we obtain that $|w_t| > 2^tC$ for any t, and thus $|f'(w_t)| > 2^{\frac{t}{1-\alpha}}C^{\frac{1}{1-\alpha}}$, $f(w_t) > 2^{\frac{t(2-\alpha)}{1-\alpha}}C^{\frac{2-\alpha}{1-\alpha}}$, which means β-GD diverges.

(Case II) When $\alpha = 1$, consider the convex function $f(w) := e^w + e^{-w}$ with unique minimizer $w = 0$ and derivative $f'(w) := e^w - e^{-w} = (e^{|w|} - e^{-|w|})\mathrm{sgn}(w)$. Based on item 4 of Proposition 1, $f \in \mathcal{L}^*_{\text{sym}}(1)$. Applying β-GD to this function yields that

$$w_{t+1} = w_t - \frac{\eta f'(w_t)}{|f'(w_t)|^\beta} = w_t - \eta\big(e^{|w_t|} - e^{-|w_t|}\big)^{1-\beta}\mathrm{sgn}(w_t). \tag{46}$$

Since $\beta < \alpha = 1$, $|w_t|^{-1}\big(e^{|w_t|} - e^{-|w_t|}\big)^{1-\beta} \to +\infty$ as $|w_t| \to +\infty$. Hence, there exists a constant $C > 1$ such that $\big(e^{|w_t|} - e^{-|w_t|}\big)^{1-\beta} > 3|w_t|$ for $|w_t| > C$. Therefore, $|w_{t+1}| > 2|w_t|$. Therefore, if $|w_0| > C$, by induction we obtain that $|w_t| > 2^tC$ for any t, and thus $f(w_t) > |f'(w_t)| = e^{|w_t|} - e^{-|w_t|} \ge \frac{1}{2}e^{|w_t|} > \frac{1}{2}\exp(2^tC)$, which means β-GD diverges.

G. Proof of Proposition 4

G.1. Proof of Item 1

First, we will prove eq. (20) given $f \in \mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$. Note that eq. (27) holds for $f \in \mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$, i.e.,

$$\begin{aligned}
\mathbb{E}_\xi\|\nabla f_\xi(w_\theta) - \nabla f_\xi(w)\|^2 &\le \theta^2\|w' - w\|^2\mathbb{E}_\xi \int_0^1 \big(L_0 + L_1\|\nabla f_\xi(w_{\theta u})\|^\alpha\big)^2 du \\
&\overset{(i)}{=} \theta\|w' - w\|^2\mathbb{E}_\xi \int_0^\theta \big(L_0 + L_1\|\nabla f_\xi(w_{u'})\|^\alpha\big)^2 du' \\
&\overset{(ii)}{\le} G(\theta)\|w' - w\|^2 \qquad (47)
\end{aligned}$$

where (i) uses change of variables $u' = \theta u$ and (ii) denotes $G(\theta) := \mathbb{E}_\xi \int_0^\theta \big(L_0 + L_1\|\nabla f_\xi(w_{u'})\|^\alpha\big)^2 du'$ and uses $\theta \le 1$. Then

$$\begin{aligned}
G'(\theta) &= \mathbb{E}_\xi\big(L_0 + L_1\|\nabla f_\xi(w_\theta)\|^\alpha\big)^2 \\
&\overset{(i)}{\le} 2L_0^2 + 2L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w_\theta)\|^{2\alpha} \\
&\overset{(ii)}{\le} 2L_0^2 + 4L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w)\|^{2\alpha} + 4L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w_\theta) - \nabla f_\xi(w)\|^{2\alpha} \\
&\overset{(iii)}{\le} 2L_0^2 + 4L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w)\|^{2\alpha} + 4L_1^2\big(\mathbb{E}_\xi\|\nabla f_\xi(w_\theta) - \nabla f_\xi(w)\|^2\big)^\alpha \\
&\overset{(iv)}{\le} 2L_0^2 + 4L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w)\|^{2\alpha} + 4L_1^2 G(\theta)^\alpha\|w' - w\|^{2\alpha} \\
&\overset{(v)}{\le} 3\big(A + BG(\theta)\big)^\alpha, \qquad (48)
\end{aligned}$$

where (i) use the inequality that $(a+b)^2 \le 2a^2 + 2b^2$ for any $a, b \ge 0$, (ii) uses the inequality that $\|v' + v\|^{2\alpha} \le 2\|v'\|^{2\alpha} + 2\|v\|^{2\alpha}$ for any $v, v' \in \mathbb{R}^d$ and $\alpha \in [0,1]$, (iii) uses Jensen's inequality that $\mathbb{E}(X^\alpha) \le (\mathbb{E}X)^\alpha$ where $X = \|\nabla f_\xi(w_\theta) - \nabla f_\xi(w)\|^2$, (iv) uses eq. (47), and (v) uses Jensen's inequality that $a^\alpha + b^\alpha + c^\alpha \le 3(a+b+c)^\alpha$ for any $a, b, c \ge 0$ and denotes that $A := (2L_0^2)^{\frac{1}{\alpha}} + (4L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w)\|^{2\alpha})^{\frac{1}{\alpha}}$, $B := (4L_1^2)^{\frac{1}{\alpha}}\|w' - w\|^2$. When $\alpha \in (0,1)$, rearranging the above inequality yields that

$$3B(1-\alpha) \ge \frac{B(1-\alpha)G'(\theta)}{\big(A + BG(\theta)\big)^\alpha} = \frac{d}{d\theta}\big(A + BG(\theta)\big)^{1-\alpha}$$

Integrating the above inequality over $\theta \in [0,1]$ yields that

$$\big(A + BG(1)\big)^{1-\alpha} \le 3B(1-\alpha) + \big(A + BG(0)\big)^{1-\alpha} \le 3B + A^{1-\alpha} \le 2\big((3B)^{\frac{1}{1-\alpha}} + A\big)^{1-\alpha}.$$

where (i) applies Jensen's inequality to the concave function $x^{1-\alpha}$. Rearranging the above inequality yields that

$$\begin{aligned}
BG(1) &\le 2^{\frac{1}{1-\alpha}}\big((3B)^{\frac{1}{1-\alpha}} + A\big) - A \\
&\le 6^{\frac{1}{1-\alpha}}B^{\frac{1}{1-\alpha}} + A(2^{\frac{1}{1-\alpha}}) \\
&\le 6^{\frac{1}{1-\alpha}}(4L_1^2)^{\frac{1}{\alpha(1-\alpha)}}\|w' - w\|^{\frac{2}{1-\alpha}} + 2^{\frac{1}{1-\alpha}}(2L_0^2)^{\frac{1}{\alpha}} + 2^{\frac{1}{1-\alpha}}(4L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w)\|^{2\alpha})^{\frac{1}{\alpha}}.
\end{aligned}$$

Substituting the above inequality into eq. (47), we obtain that

$$\begin{aligned}
\mathbb{E}_\xi\|\nabla f_\xi(w_\theta) - \nabla f_\xi(w)\|^2 &\le G(\theta)\|w' - w\|^2 \\
&\le (4L_1^2)^{-\frac{1}{\alpha}}BG(1) \\
&\le (24L_1^2)^{\frac{1}{1-\alpha}}\|w' - w\|^{\frac{2}{1-\alpha}} + 2^{\frac{1}{1-\alpha}}\Big(\frac{L_0^2}{2L_1^2}\Big)^{\frac{1}{\alpha}} + 2^{\frac{1}{1-\alpha}}(\mathbb{E}_\xi\|\nabla f_\xi(w)\|^{2\alpha})^{\frac{1}{\alpha}} \qquad (49)
\end{aligned}$$

Therefore,

$$\mathbb{E}\|\nabla f_\xi(w_\theta)\|^{2\alpha} \overset{(i)}{\le} 2\mathbb{E}\|\nabla f_\xi(w_\theta) - \nabla f_\xi(w)\|^{2\alpha} + 2\mathbb{E}\|\nabla f_\xi(w)\|^{2\alpha}$$

$$\begin{aligned}&\overset{(ii)}{\le} 2\big(\mathbb{E}\|\nabla f_\xi(w_\theta)-\nabla f_\xi(w)\|^2\big)^\alpha+2\mathbb{E}\|\nabla f_\xi(w)\|^{2\alpha}\\&\overset{(iii)}{\le} 2(24L_1^2)^{\frac{\alpha}{1-\alpha}}\|w'-w\|^{\frac{2\alpha}{1-\alpha}}+2^{\frac{\alpha}{1-\alpha}}L_0^2L_1^{-2}+\big(2^{\frac{1}{1-\alpha}}+2\big)\mathbb{E}_\xi\|\nabla f_\xi(w)\|^{2\alpha}\end{aligned}\tag{50}$$

where (i) uses the inequality that $\|v'+v\|^{2\alpha}\le 2\|v'\|^{2\alpha}+2\|v\|^{2\alpha}$ for any $v,v'\in\mathbb{R}^d$ and $\alpha\in(0,1)$, (ii) uses Jensen's inequality that $\mathbb{E}(X^\alpha)\le(\mathbb{E}X)^\alpha$ where $X=\|\nabla f_\xi(w_\theta)-\nabla f_\xi(w)\|^2$, and (iii) uses eq. (49) and the inequality that $(a+b+c)^\alpha\le a^\alpha+b^\alpha+c^\alpha$ for any $a,b,c\ge 0$ and $\alpha\in[0,1]$. Note that (26) holds for $f\in\mathbb{E}\mathcal{L}_{\text{sym}}^*$, so we have

$$\begin{aligned}&\mathbb{E}_\xi\|\nabla f_\xi(w')-\nabla f_\xi(w)\|^2\\&\le\|w'-w\|^2\mathbb{E}_\xi\int_0^1\big(L_0+L_1\|\nabla f_\xi(w_\theta)\|^\alpha\big)^2d\theta\\&\le 2\|w'-w\|^2\int_0^1 L_0^2+L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w_\theta)\|^{2\alpha}d\theta\\&\overset{(i)}{\le}\|w'-w\|^2\mathbb{E}_\xi\big(\overline{K}_0^2+\overline{K}_1^2\|\nabla f_\xi(w)\|^{2\alpha}+\overline{K}_2^2\|w'-w\|^{\frac{2\alpha}{1-\alpha}}\big)\\&\overset{(ii)}{\le}\|w'-w\|^2\mathbb{E}_\xi\big(\overline{K}_0+\overline{K}_1\|\nabla f_\xi(w)\|^\alpha+\overline{K}_2\|w'-w\|^{\frac{\alpha}{1-\alpha}}\big)^2\end{aligned}$$

where (i) uses eq. (50) and denotes that $\overline{K}_0^2:=2^{\frac{4-2\alpha}{1-\alpha}}L_0^2\ge 2L_0^2(2^{\frac{1}{1-\alpha}}+1)$, $\overline{K}_1^2:=2^{\frac{4-2\alpha}{1-\alpha}}L_1^2\ge 2L_1^2(2^{\frac{1}{1-\alpha}}+2)$, $\overline{K}_2^2:=(25L_1^2)^{\frac{1}{1-\alpha}}\ge 4L_1^2(24L_1^2)^{\frac{\alpha}{1-\alpha}}$, and (ii) uses the inequality that $a^2+b^2+c^2\le(a+b+c)^2$ for any $a,b,c\ge 0$. This proves eq. (20).

Then, it remains to prove that $f\in\mathbb{E}\mathcal{L}_{\text{sym}}^*(\alpha)$ given eq. (20). Then for any $w,w'\in\mathbb{R}^d$ and $n\in\mathbb{N}^+$, we have

$$\begin{aligned}&\mathbb{E}_\xi\|\nabla f_\xi(w')-\nabla f_\xi(w)\|^2\\&=\mathbb{E}_\xi\Big\|\sum_{k=0}^{n-1}\big(\nabla f_\xi(w_{(k+1)/n})-\nabla f_\xi(w_{k/n})\big)\Big\|^2\\&\overset{(i)}{\le} n\sum_{k=0}^{n-1}\mathbb{E}_\xi\big\|\nabla f_\xi(w_{(k+1)/n})-\nabla f_\xi(w_{k/n})\big\|^2\\&\overset{(ii)}{\le} n\sum_{k=0}^{n-1}\|w_{(k+1)/n}-w_{k/n}\|^2\mathbb{E}_\xi\big(\overline{K}_0+\overline{K}_1\|\nabla f_\xi(w_{k/n})\|^\alpha+\overline{K}_2\|w_{(k+1)/n}-w_{k/n}\|^{\frac{\alpha}{1-\alpha}}\big)^2\\&=\|w'-w\|^2\mathbb{E}_\xi\sum_{k=0}^{n-1}\frac{1}{n}\big(\overline{K}_0+\overline{K}_1\|\nabla f_\xi(w_{k/n})\|^\alpha+\overline{K}_2n^{-\frac{\alpha}{1-\alpha}}\|w'-w\|^{\frac{\alpha}{1-\alpha}}\big)^2,\end{aligned}$$

where (i) applies Jensen' inequality to the convex function $\|\cdot\|^2$ and (ii) uses eq. (20). For any $\epsilon>0$, there exists $n_0>0$ such that $\overline{K}_2n^{-\frac{\alpha}{1-\alpha}}\|w'-w\|^{\frac{\alpha}{1-\alpha}}<\epsilon$ for any $n\ge n_0$. Therefore, taking limit superior of both sides of the above inequality, we obtain that

$$\begin{aligned}&\mathbb{E}_\xi\|\nabla f_\xi(w')-\nabla f_\xi(w)\|^2\\&\le\|w'-w\|^2\limsup_{n\to+\infty}\mathbb{E}_\xi\sum_{k=0}^{n-1}\frac{1}{n}\big(\overline{K}_0+\overline{K}_1\|\nabla f_\xi(w_{k/n})\|^\alpha+\overline{K}_2n^{-\frac{\alpha}{1-\alpha}}\|w'-w\|^{\frac{\alpha}{1-\alpha}}\big)^2\\&\overset{(i)}{\le}\|w'-w\|^2\mathbb{E}_\xi\limsup_{n\to+\infty}\sum_{k=0}^{n-1}\frac{1}{n}\big(\overline{K}_0+\epsilon+\overline{K}_1\|\nabla f_\xi(w_{k/n})\|^\alpha\big)^2\\&=\|w'-w\|^2\mathbb{E}_\xi\int_0^1\big(\overline{K}_0+\epsilon+\overline{K}_1\|\nabla f_\xi(w_\theta)\|^\alpha\big)^2d\theta\end{aligned}$$

where (i) uses Fatou's lemma. Letting $\epsilon\to+0$ in the above inequality, we obtain the following inequality, which proves that $f\in\mathbb{E}\mathcal{L}_{\text{sym}}^*(\alpha)$ based on Lemma 3

$$\mathbb{E}_\xi\|\nabla f_\xi(w')-\nabla f_\xi(w)\|^2\le\|w'-w\|^2\mathbb{E}_\xi\int_0^1\big(\overline{K}_0+\overline{K}_1\|\nabla f_\xi(w_\theta)\|^\alpha\big)^2d\theta\tag{51}$$

G.2. Proof of Item 2

First, we will prove eq. (21) given $f \in \mathbb{E}\mathcal{L}^*_{\text{sym}}(1)$. Note that eq. (48) holds for any $f \in \mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$ with $\alpha \in [0,1]$. Substituting $\alpha = 1$ into eq. (48), i.e.,

$$G'(\theta) \le 3A + 3BG(\theta)$$

where $G(\theta) := \mathbb{E}_\xi \int_0^\theta \big(L_0 + L_1\|\nabla f_\xi(w_u)\|\big)^2 du$, $w_u := uw' + (1-u)w$, $A := 2L_0^2 + 4L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w)\|^2$ and $B := 4L_1^2\|w'-w\|^2$. Rearranging the above inequality yields that

$$3B \ge \frac{BG'(\theta)}{A + BG(\theta)} = \frac{d}{d\theta}\ln\big(A + BG(\theta)\big).$$

Integrating the above inequality over $\theta \in [0,1]$, we obtain that

$$\ln\big(A + BG(1)\big) \le \ln\big(A + BG(0)\big) + 3B = \ln A + 3B.$$

Hence, we have $BG(\theta) \le BG(1) \le A(e^{3B} - 1)$. Substituting this inequality into eq. (47), we obtain that

$$\begin{aligned}
\mathbb{E}_\xi\|\nabla f_\xi(w_\theta) - \nabla f_\xi(w)\|^2 &\le G(\theta)\|w'-w\|^2 \\
&\le \frac{A}{4L_1^2}(e^{3B} - 1) \\
&\le \Big(\frac{L_0^2}{2L_1^2} + \mathbb{E}_\xi\|\nabla f_\xi(w)\|^2\Big)\big(\exp(12L_1^2\|w'-w\|^2) - 1\big).
\end{aligned} \tag{52}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\|\nabla f_\xi(w_\theta)\|^2 &\overset{(i)}{\le} 2\mathbb{E}\|\nabla f_\xi(w_\theta) - \nabla f_\xi(w)\|^2 + 2\mathbb{E}\|\nabla f_\xi(w)\|^2 \\
&\overset{(ii)}{\le} \Big(\frac{L_0^2}{L_1^2} + 2\mathbb{E}_\xi\|\nabla f_\xi(w)\|^2\Big)\big(\exp(12L_1^2\|w'-w\|^2) - 1\big) + 2\mathbb{E}\|\nabla f_\xi(w)\|^2
\end{aligned} \tag{53}$$

where (i) uses the inequality that $\|v'+v\|^2 \le 2\|v'\|^2 + 2\|v\|^2$ for any $v, v' \in \mathbb{R}^d$ and (ii) uses eq. (52). Note that (26) holds for $f \in \mathbb{E}\mathcal{L}^*_{\text{sym}}(1)$. Hence, we prove eq. (21) as follows.

$$\begin{aligned}
&\mathbb{E}_\xi\|\nabla f_\xi(w') - \nabla f_\xi(w)\|^2 \\
&\le \|w'-w\|^2\mathbb{E}_\xi\int_0^1 \big(L_0 + L_1\|\nabla f_\xi(w_\theta)\|\big)^2 d\theta \\
&\le 2\|w'-w\|^2\int_0^1 L_0^2 + L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w_\theta)\|^2 d\theta \\
&\overset{(i)}{\le} 2\|w'-w\|^2\Big(L_0^2 + (L_0^2 + 2L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w)\|^2)\big(\exp(12L_1^2\|w'-w\|^2) - 1\big) + 2L_1^2\mathbb{E}\|\nabla f_\xi(w)\|^2\Big) \\
&= 2\|w'-w\|^2(L_0^2 + 2L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w)\|^2)\exp(12L_1^2\|w'-w\|^2),
\end{aligned}$$

where (i) uses eq. (53). This proves eq. (21).

Finally, it remains to prove that $f \in \mathbb{E}\mathcal{L}^*_{\text{sym}}(\alpha)$ given eq. (21). Then for any $w, w' \in \mathbb{R}^d$ and $n \in \mathbb{N}^+$, we have

$$\begin{aligned}
&\mathbb{E}_\xi\|\nabla f_\xi(w') - \nabla f_\xi(w)\|^2 \\
&= \mathbb{E}_\xi\Big\|\sum_{k=0}^{n-1}\big(\nabla f_\xi(w_{(k+1)/n}) - \nabla f_\xi(w_{k/n})\big)\Big\|^2 \\
&\overset{(i)}{\le} n\sum_{k=0}^{n-1}\mathbb{E}_\xi\big\|\nabla f_\xi(w_{(k+1)/n}) - \nabla f_\xi(w_{k/n})\big\|^2
\end{aligned}$$

$$\begin{aligned}
&\overset{(ii)}{\le} 2n\sum_{k=0}^{n-1}\|w_{(k+1)/n}-w_{k/n}\|^2(L_0^2+2L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w_{k/n})\|^2)\exp(12L_1^2\|w_{(k+1)/n}-w_{k/n}\|^2)\\
&=\|w'-w\|^2\sum_{k=0}^{n-1}\frac{1}{n}(L_0^2+2L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w_{k/n})\|^2)\exp(12n^{-2}L_1^2\|w'-w\|^2),
\end{aligned}$$

where (i) applies Jensen' inequality to the convex function $\|\cdot\|^2$ and (ii) uses eq. (21). For any $\epsilon>0$, there exists $n_0>0$ such that $\exp(12n^{-2}L_1^2\|w'-w\|^2)<1+\epsilon$ for any $n\ge n_0$. Therefore, letting $n\to+\infty$ in the above inequality, we obtain that

$$\begin{aligned}
&\mathbb{E}_\xi\|\nabla f_\xi(w')-\nabla f_\xi(w)\|^2\\
&\le(1+\epsilon)\|w'-w\|^2\limsup_{n\to+\infty}\sum_{k=0}^{n-1}\frac{1}{n}(L_0^2+2L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w_{k/n})\|^2)\\
&=(1+\epsilon)\|w'-w\|^2\int_0^1(L_0^2+2L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w_\theta)\|^2)d\theta,
\end{aligned}$$

where (i) uses Fatou's lemma. Letting $\epsilon\to+0$ in the above inequality, we obtain the following inequality, which proves that $f\in\mathbb{E}\mathcal{L}_{\text{sym}}^*(1)$ based on Lemma 3

$$\mathbb{E}_\xi\|\nabla f_\xi(w')-\nabla f_\xi(w)\|^2\le\|w'-w\|^2\mathbb{E}_\xi\int_0^1(L_0^2+2L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w_\theta)\|^2)d\theta. \tag{54}$$

G.3. Proof of Item 3

For any $f\in\mathbb{E}\mathcal{L}_{\text{sym}}^*(\alpha)$, we will prove that $f\in\mathcal{L}_{\text{sym}}^*(\alpha)$ in two cases: $\alpha\in(0,1)$ and $\alpha=1$.

(Case I) When $\alpha\in(0,1)$, eq. (20) holds, so we have

$$\begin{aligned}
&\|\nabla f(w')-\nabla f(w)\|=\big\|\mathbb{E}_\xi\big(\nabla f_\xi(w')-\nabla f_\xi(w)\big)\big\|\\
&\le\sqrt{\mathbb{E}_\xi\|\nabla f_\xi(w')-\nabla f_\xi(w)\|^2}\\
&\le\|w'-w\|\big(\overline{K}_0+\overline{K}_1\mathbb{E}_\xi\|\nabla f_\xi(w)\|^\alpha+\overline{K}_2\|w'-w\|^{\frac{\alpha}{1-\alpha}}\big)\\
&\overset{(i)}{\le}\|w'-w\|\big(\overline{K}_0+\overline{K}_1\Lambda^\alpha+\overline{K}_1(\Gamma^\alpha+1)\|\nabla f(w)\|^\alpha+\overline{K}_2\|w'-w\|^{\frac{\alpha}{1-\alpha}}\big),
\end{aligned}$$

where (i) uses Lemma 4. The above inequality implies that $f\in\mathcal{L}_{\text{sym}}^*(\alpha)$ based on item 1 of Proposition 1.

(Case II) When $\alpha=1$, eq. (21) holds, so we have

$$\begin{aligned}
&\|\nabla f(w')-\nabla f(w)\|=\big\|\mathbb{E}_\xi\big(\nabla f_\xi(w')-\nabla f_\xi(w)\big)\big\|\\
&\le\sqrt{\mathbb{E}_\xi\|\nabla f_\xi(w')-\nabla f_\xi(w)\|^2}\\
&\le\|w'-w\|\sqrt{2L_0^2+4L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w)\|^2}\exp(6L_1^2\|w'-w\|^2)\\
&\overset{(i)}{\le}\|w'-w\|\sqrt{2L_0^2+4L_1^2\Lambda^2+4L_1^2(\Gamma^2+1)\|\nabla f(w)\|^2}\exp(6L_1^2\|w'-w\|^2)\\
&\overset{(ii)}{\le}\|w'-w\|\big(2L_0+2L_1\Lambda+2L_1(\Gamma+1)\|\nabla f(w)\|\big)\exp(6L_1^2\|w'-w\|^2)
\end{aligned}$$

where (i) uses Lemma 4, (ii) uses the inequality that $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ for any $a,b\ge0$. The above inequality implies that $f\in\mathcal{L}_{\text{sym}}^*(\alpha)$ based on item 2 of Proposition 1.

H. Proof of Theorem 4

We will first prove the following lemmas which will be used in the proof of Theorem 4.

Lemma 6. *Apply SPIDER algorithm (Algorithm 2) to $f \in \mathcal{L}^*_{sym}(\alpha)$ with stepsize $\gamma \le \frac{\epsilon}{2\overline{K}_0+2\overline{K}_2+2\overline{K}_1(\Lambda^\alpha+\Gamma^\alpha+1)+1}$ (when $\alpha \in (0,1)$) or $\gamma \le \frac{\epsilon}{3L_1\sqrt{\Gamma^2+1}+3\sqrt{L_0^2+2L_1^2\Lambda^2}}$ (when $\alpha = 1$) ($\epsilon \in (0,1)$ is the target accuracy). Then we have,*

$$\mathbb{E}_{\xi\sim\mathbb{P}}\big(\|\nabla f_\xi(w_{t+1}) - \nabla f_\xi(w_t)\|^2 \big| S_{1:t}\big) \le \epsilon^2\big(1+\|\nabla f(w_t)\|^2\big), \tag{55}$$

where $\xi \sim \mathbb{P}$ is independent from the minibatches $S_{1:t}$.

Proof of Lemma 6. Given $S_{1:t}$, w_t, w_{t+1} are non-random based on eq. (30). Hence, eq. (20) or (21) holds respectively when $\alpha \in (0,1)$ or $\alpha = 1$.

If $\alpha \in (0,1)$, eq. (55) can be proved as follows

$$\begin{aligned}
&\mathbb{E}_{\xi\sim\mathbb{P}}\big(\|\nabla f_\xi(w_{t+1}) - \nabla f_\xi(w_t)\|^2 \big| S_{1:t}\big)\\
&\overset{(i)}{\le} \|w_{t+1}-w_t\|^2\big(\overline{K}_0 + \overline{K}_1\mathbb{E}_\xi\big(\|\nabla f_\xi(w_t)\|^\alpha \big| S_{1:t}\big) + \overline{K}_2\|w_{t+1}-w_t\|^{\frac{\alpha}{1-\alpha}}\big)^2\\
&\overset{(ii)}{\le} \gamma^2\big(\overline{K}_0 + \overline{K}_1\Lambda^\alpha + \overline{K}_1(\Gamma^\alpha+1)\|\nabla f(w_t)\|^\alpha + \overline{K}_2\big)^2\\
&\overset{(iii)}{\le} 2\gamma^2(\overline{K}_0+\overline{K}_2+\overline{K}_1\Lambda^\alpha)^2 + 2\gamma^2\overline{K}_1^2(\Gamma^\alpha+1)^2\cdot\|\nabla f(w_t)\|^{2\alpha}\\
&\le 2\gamma^2(\overline{K}_0+\overline{K}_2+\overline{K}_1\Lambda^\alpha)^2 + 2\gamma^2\overline{K}_1^2(\Gamma^\alpha+1)^2\big(\|\nabla f(w_t)\|^2+1\big)\\
&\overset{(iv)}{\le} \epsilon^2\big(1+\|\nabla f(w_t)\|^2\big),
\end{aligned}$$

where (i) uses eq. (20), (ii) uses eq. (28) and $\|w_{t+1}-w_t\| = \gamma \le 1$ based on Algorithm 2, (iii) uses the inequality that $(a+b)^2 \le 2a^2+2b^2$ for any $a, b \ge 0$, (iv) uses $\gamma \le \frac{\epsilon}{2\overline{K}_0+2\overline{K}_2+2\overline{K}_1(\Lambda^\alpha+\Gamma^\alpha+1)} \le \frac{\epsilon}{2}\big((\overline{K}_0+\overline{K}_2+\overline{K}_1\Lambda^\alpha)^2+\overline{K}_1^2(\Gamma^\alpha+1)^2\big)^{-\frac{1}{2}}$.

If $\alpha = 1$, eq. (55) can be proved as follows

$$\begin{aligned}
&\mathbb{E}_{\xi\sim\mathbb{P}}\big(\|\nabla f_\xi(w_{t+1}) - \nabla f_\xi(w_t)\|^2 \big| S_{1:t}\big)\\
&\overset{(i)}{\le} 2\|w_{t+1}-w_t\|^2\cdot(L_0^2 + 2L_1^2\mathbb{E}_\xi\|\nabla f_\xi(w_t)\|^2)\exp(12L_1^2\|w_{t+1}-w_t\|^2)\\
&\overset{(ii)}{=} 2\gamma^2\exp(12L_1^2\gamma^2)\big(L_0^2+2L_1^2\Lambda^2+2L_1^2(\Gamma^2+1)\|\nabla f(w_t)\|^2\big)\\
&\overset{(iii)}{\le} \epsilon^2\big(1+\|\nabla f(w_t)\|^2\big),
\end{aligned}$$

where (i) uses eq. (21), (ii) uses eq. (28) and $\|w_{t+1}-w_t\| = \gamma \le 1$ based on Algorithm 2, (iii) uses $\gamma \le \frac{\epsilon}{3L_1\sqrt{\Gamma^2+1}+3\sqrt{L_0^2+2L_1^2\Lambda^2}} \le \min\big(\frac{1}{3L_1}, \frac{\epsilon}{3L_1\sqrt{\Gamma^2+1}+3\sqrt{L_0^2+2L_1^2\Lambda^2}}\big)$. □

Lemma 7. *Apply SPIDER algorithm (Algorithm 2) to $f \in \mathcal{L}^*_{sym}(\alpha)$ with stepsize γ given by Lemma 6 batchsize $|S_t| = B$ when $t \mod q = 0$ and $|S_t| = B'$ otherwise. Then the approximation error $\delta_t := v_t - \nabla f(w_t)$ has the following properties conditional on minibatches $S_{1:t} := \{S_1, \ldots, S_t\}$.*

$$\mathbb{E}\big(\delta_{t+1}\big|S_{1:t}\big) = \delta_t; \forall (t+1) \textit{ mod } q \ne 0 \tag{56}$$

$$\mathbb{E}\big(\|\delta_{t+1}\|^2\big|w_{t+1}\big) \le \frac{1}{B}\big(\Gamma^2\|\nabla f(w_{t+1})\|^2+\Lambda^2\big); \forall (t+1) \textit{ mod } q = 0 \tag{57}$$

$$\mathbb{E}\big(\|\delta_{t+1}\|^2\big|S_{1:t}\big) \le \|\delta_t\|^2 + \frac{\epsilon^2}{B'}\big(1+\|\nabla f(w_t)\|^2\big) \tag{58}$$

Therefore, for any $k \in \mathbb{N}$ and $s = 0, 1, \ldots, q-1$, we have

$$\mathbb{E}\|\delta_{qk+s}\| \le \frac{\Lambda}{\sqrt{B}} + \epsilon\sqrt{\frac{q}{B'}} + \Big(\frac{\epsilon}{\sqrt{B'}} + \frac{\Gamma}{\sqrt{B}}\Big)\sum_{u=0}^{q-1}\mathbb{E}\|\nabla f(w_{qk+u})\|. \tag{59}$$

Proof of Lemma 7. We will first prove eq. (57) when $(t+1) \bmod q = 0$ and then prove eqs. (56) & (58) when $(t+1) \bmod q \ne 0$.

If $(t+1) \bmod q = 0$, then $v_{t+1} = \nabla f_{S_{t+1}}(w_{t+1})$ based on Algorithm 2. Hence, eq. (57) can be proved as follows.

$$\begin{aligned}\mathbb{E}\big(\|\delta_{t+1}\|^2\big|S_{1:t}\big) =& \mathbb{E}\big(\|\nabla f_{S_{t+1}}(w_{t+1}) - \nabla f(w_{t+1})\|^2\big|S_{1:t}\big)\\ =& \frac{1}{|S_t|}\mathbb{E}_{\xi\sim\mathbb{P}}\big(\|\nabla f_\xi(w_{t+1}) - \nabla f(w_{t+1})\|^2\big|S_{1:t}\big)\\ \overset{(i)}{\le}& \frac{1}{B}\big(\Gamma^2\|\nabla f(w_{t+1})\|^2 + \Lambda^2\big),\end{aligned}$$

where (i) uses Assumption 1.

If $(t+1) \bmod q \neq 0$, then $v_{t+1} = v_t + \nabla f_{S_{t+1}}(w_{t+1}) - \nabla f_{S_{t+1}}(w_t)$ based on Algorithm 2. Hence, eq. (56) can be proved as follows.

$$\begin{aligned}\mathbb{E}\big(\delta_{t+1}\big|S_{1:t}\big) &= \mathbb{E}\big(v_{t+1} - \nabla f(w_{t+1})\big|S_{1:t}\big)\\ &= \mathbb{E}\big(v_t + \nabla f_{S_{t+1}}(w_{t+1}) - \nabla f_{S_{t+1}}(w_t) - \nabla f(w_{t+1})\big|S_{1:t}\big)\\ &\overset{(i)}{=} v_t - \nabla f(w_t) = \delta_t,\end{aligned}$$

where (i) uses eq. (30). Then eq. (58) can be proved as follows.

$$\begin{aligned}&\mathbb{E}\big(\|\delta_{t+1}\|^2\big|S_{1:t}\big)\\ &\overset{(i)}{=} \|\delta_t\|^2 + \mathbb{E}\big(\|\delta_{t+1} - \delta_t\|^2\big|S_{1:t}\big)\\ &= \|\delta_t\|^2 + \mathbb{E}\big(\|v_{t+1} - v_t - \nabla f(w_{t+1}) + \nabla f(w_t)\|^2\big|S_{1:t}\big)\\ &= \|\delta_t\|^2 + \mathbb{E}\big(\|\nabla f_{S_{t+1}}(w_{t+1}) - \nabla f_{S_{t+1}}(w_t) - \nabla f(w_{t+1}) + \nabla f(w_t)\|^2\big|S_{1:t}\big)\\ &\overset{(ii)}{=} \|\delta_t\|^2 + \frac{1}{|S_{t+1}|}\mathbb{E}_{\xi\sim\mathbb{P}}\big(\|\nabla f_\xi(w_{t+1}) - \nabla f_\xi(w_t) - \nabla f(w_{t+1}) + \nabla f(w_t)\|^2\big|S_{1:t}\big)\\ &\overset{(iii)}{\le} \|\delta_t\|^2 + \frac{1}{|S_{t+1}|}\mathbb{E}_{\xi\sim\mathbb{P}}\big(\|\nabla f_\xi(w_{t+1}) - \nabla f_\xi(w_t)\|^2\big|S_{1:t}\big)\\ &\overset{(iv)}{\le} \|\delta_t\|^2 + \frac{\epsilon^2}{B'}\big(1 + \|\nabla f(w_t)\|^2\big),\end{aligned}$$

where (i) uses eq. (56), (ii) uses eq. (30) which implies that conditional on $S_{1:t}$, S_{t+1} obtained from i.i.d. sampling is the only source of randomness in $\nabla f_\xi(w_{t+1}) - \nabla f_\xi(w_t) - \nabla f(w_{t+1}) + \nabla f(w_t)$, both (ii) and (iii) use $\mathbb{E}\big(\|\nabla f_\xi(w_{t+1}) - \nabla f_\xi(w_t) - \nabla f(w_{t+1}) + \nabla f(w_t)\|^2\big|S_{1:t}\big) = 0$, and (iv) uses Lemma 6 and $|S_{t+1}| = B'$ (since $t+1 \mod q \neq 0$).

Next, to prove eq. (59), we will first prove the following relation for any $s, s', k \in \mathbb{N}$ such that $s' \le s \le q-1$.

$$\mathbb{E}\|\delta_{qk+s}\| \le \mathbb{E}\sqrt{\|\delta_{qk+s'}\|^2 + \frac{\epsilon^2(s-s')}{B'}} + \frac{\epsilon}{\sqrt{B'}}\sum_{u=s'}^{s-1}\mathbb{E}\|\nabla f(w_{qk+u})\|. \tag{60}$$

We prove eq. (60) via backward induction on $s' = s, s-1, \ldots, 1, 0$. Note that eq. (60) holds trivially for $s' = s$. Then, assume that eq. (60) holds for a certain value of $s' \in [1, s]$ and we prove eq. (60) for $s'-1$ as follows.

$$\begin{aligned}&\mathbb{E}\|\delta_{qk+s}\| - \frac{\epsilon}{\sqrt{B'}}\sum_{u=s'}^{s-1}\mathbb{E}\|\nabla f(w_{qk+u})\|\\ &\overset{(i)}{\le} \mathbb{E}\mathbb{E}\left(\sqrt{\|\delta_{qk+s'}\|^2 + \frac{\epsilon^2(s-s')}{B'}}\middle|S_{1:qk+s'-1}\right)\\ &\overset{(ii)}{\le} \mathbb{E}\sqrt{\mathbb{E}\left(\|\delta_{qk+s'}\|^2 + \frac{\epsilon^2(s-s')}{B'}\middle|S_{1:qk+s'-1}\right)}\\ &\overset{(iii)}{\le} \mathbb{E}\sqrt{\|\delta_{qk+s'-1}\|^2 + \frac{\epsilon^2}{B'}\big(1 + \|\nabla f(w_{qk+s'-1})\|^2\big) + \frac{\epsilon^2(s-s')}{B'}}\end{aligned}$$

$$\overset{(iv)}{\leq} \mathbb{E}\sqrt{\|\delta_{qk+s'-1}\|^2 + \frac{\epsilon^2(s-s'+1)}{B'}} + \frac{\epsilon}{\sqrt{B'}}\mathbb{E}\|\nabla f(w_{qk+s'-1})\|,$$

where (i) uses eq. (60) for s', (ii) applies Jensen's inequality to the concave function $\sqrt{\cdot}$, (iii) uses eq. (58), and (iv) uses the inequality that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$. Substituting $s' = 0$ into eq. (60), we prove eq. (59) as follows.

$$\begin{aligned}
\mathbb{E}\|\delta_{qk+s}\| &\leq \mathbb{E}\sqrt{\|\delta_{qk}\|^2 + \frac{s\epsilon^2}{B'}} + \frac{\epsilon}{\sqrt{B'}}\sum_{u=0}^{s-1}\mathbb{E}\|\nabla f(w_{qk+u})\| \\
&\overset{(i)}{\leq} \sqrt{\mathbb{E}\|\delta_{qk}\|^2} + \epsilon\sqrt{\frac{q}{B'}} + \frac{\epsilon}{\sqrt{B'}}\sum_{u=0}^{s-1}\mathbb{E}\|\nabla f(w_{qk+u})\| \\
&\overset{(ii)}{\leq} \frac{1}{\sqrt{B}}\big(\Gamma\|\nabla f(w_{qk})\| + \Lambda\big) + \epsilon\sqrt{\frac{q}{B'}} + \frac{\epsilon}{\sqrt{B'}}\sum_{u=0}^{s-1}\mathbb{E}\|\nabla f(w_{qk+u})\| \\
&\leq \frac{\Lambda}{\sqrt{B}} + \epsilon\sqrt{\frac{q}{B'}} + \Big(\frac{\epsilon}{\sqrt{B'}} + \frac{\Gamma}{\sqrt{B}}\Big)\sum_{u=0}^{q-1}\mathbb{E}\|\nabla f(w_{qk+u})\|,
\end{aligned}$$

where (i) uses the inequality that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$ and then applies Lyapunov inequality, and (ii) uses eq. (57) and then uses $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$. □

Lemma 8. *Apply SPIDER algorithm (Algorithm 2) to $f \in \mathcal{L}^*_{sym}(\alpha)$ with batchsize $|S_t| = B$ when $t \mod q = 0$ and $|S_t| = B'$ otherwise, and stepsize stepsize $\gamma \leq \frac{\epsilon}{2(\overline{K}_0+\overline{K}_1+2\overline{K}_2)+1}$ (when $\alpha \in (0,1)$) or $\gamma \leq \frac{\epsilon}{5L_1+8L_0}$ (when $\alpha = 1$) ($\epsilon \in (0,1)$ is the target accuracy). Then the decrease of the function f has the following bound.*

$$f(w_{t+1}) - f(w_t) \leq \frac{\gamma\epsilon}{8} - \frac{\gamma}{2}\|v_t\| - \frac{3\gamma}{8}\|\nabla f(w_t)\| + \frac{3\gamma}{2}\|v_t - \nabla f(w_t)\|. \tag{61}$$

Proof of Lemma 8. We consider two cases: $\alpha \in (0,1)$ and $\alpha = 1$.

(Case I) When $\alpha \in (0,1)$, eq. (7) holds for $f \in \mathcal{L}^*_{\text{sym}}(\alpha)$. Hence,

$$\begin{aligned}
&f(w_{t+1}) - f(w_t) \\
&\leq \nabla f(w_t)^\top(w_{t+1} - w_t) + \frac{1}{2}\|w_{t+1} - w_t\|^2\big(\overline{K}_0 + \overline{K}_1\|\nabla f(w_t)\|^\alpha + 2\overline{K}_2\|w_{t+1} - w_t\|^{\frac{\alpha}{1-\alpha}}\big) \\
&= -\frac{\gamma}{\|v_t\|}\nabla f(w_t)^\top v_t + \frac{\gamma^2}{2}\big(\overline{K}_0 + \overline{K}_1\|\nabla f(w_t)\|^\alpha + 2\overline{K}_2\gamma^{\frac{\alpha}{1-\alpha}}\big) \\
&\overset{(i)}{\leq} -\frac{\gamma\big(v_t - \nabla f(w_t)\big)^\top v_t + \gamma\|v_t\|^2}{\|v_t\|} + \frac{\gamma^2}{2}(\overline{K}_0 + \overline{K}_1 + 2\overline{K}_2) + \frac{\overline{K}_1\gamma^2}{2}\|\nabla f(w_t)\| \\
&\overset{(ii)}{\leq} \gamma\|v_t - \nabla f(w_t)\| - \frac{\gamma}{2}\|v_t\| - \frac{\gamma}{2}\big(\|\nabla f(w_t)\| - \|v_t - \nabla f(w_t)\|\big) + \frac{\gamma\epsilon}{8} + \frac{\gamma\epsilon}{8}\|\nabla f(w_t)\| \\
&\overset{(iii)}{\leq} \frac{\gamma\epsilon}{8} - \frac{\gamma}{2}\|v_t\| - \frac{3\gamma}{8}\|\nabla f(w_t)\| + \frac{3\gamma}{2}\|v_t - \nabla f(w_t)\|,
\end{aligned} \tag{62}$$

where (i) uses $\|\nabla f(w_t)\|^\alpha \leq \|\nabla f(w_t)\|^2 + 1$ and $\gamma \leq 1$, (ii) uses Cauchy-Schwartz inequality, $\|v_t\| \geq \|\nabla f(w_t)\| - \|v_t - \nabla f(w_t)\|$ and $\gamma \leq \frac{\epsilon}{2(\overline{K}_0+\overline{K}_1+2\overline{K}_2)}$, (iii) uses $\epsilon \leq 1$.

(Case II) When $\alpha = 1$, we have $\beta = 1$ and eq. (8) holds for $f \in \mathcal{L}^*_{\text{sym}}(1)$. Hence,

$$\begin{aligned}
&f(w_{t+1}) - f(w_t) \\
&\leq \nabla f(w_t)^\top(w_{t+1} - w_t) + \frac{1}{2}\|w_{t+1} - w_t\|^2\big(L_0 + L_1\|\nabla f(w_t)\|\big)\exp\big(L_1\|w_{t+1} - w_t\|\big) \\
&\leq -\frac{\gamma}{\|v_t\|}\nabla f(w_t)^\top v_t + \frac{1}{2}\gamma^2\big(L_0 + L_1\|\nabla f(w_t)\|\big)\exp(L_1\gamma)
\end{aligned}$$

$$\begin{aligned}
&\overset{(i)}{\le} -\frac{\gamma\big(v_t - \nabla f(w_t)\big)^\top v_t + \gamma\|v_t\|^2}{\|v_t\|} + L_0\gamma^2 + \frac{\gamma}{8}\|\nabla f(w_t)\| \\
&\overset{(ii)}{\le} \gamma\|v_t - \nabla f(w_t)\| - \frac{\gamma}{2}\|v_t\| - \frac{\gamma}{2}\big(\|\nabla f(w_t)\| - \|v_t - \nabla f(w_t)\|\big) + \frac{\gamma\epsilon}{8} + \frac{\gamma}{8}\|\nabla f(w_t)\| \\
&\le \frac{\gamma\epsilon}{8} - \frac{\gamma}{2}\|v_t\| - \frac{3\gamma}{8}\|\nabla f(w_t)\| + \frac{3\gamma}{2}\|v_t - \nabla f(w_t)\|,
\end{aligned} \tag{63}$$

where (i) uses $\gamma \le \frac{1}{5L_1}$, (ii) uses Cauchy-Schwartz inequality, $\|v_t\| \ge \|\nabla f(w_t)\| - \|v_t - \nabla f(w_t)\|$ and $\gamma \le \frac{\epsilon}{8L_0}$. □

Now we will prove Theorem 4. First, it can be easily verified that the choice of stepsize γ and batchsize $|S_t|$ satisfies the requirements of Lemmas 7 & 8. Therefore, eq. (61) in Lemma 8 holds. Taking expectation of eq. (61) and telescoping over $t = 0, 1, \ldots, T-1$ where $T = qK$, we obtain that

$$\begin{aligned}
&\mathbb{E}f(w_T) - \mathbb{E}f(w_0) \\
&\le \frac{T\gamma\epsilon}{8} - \frac{\gamma}{2}\sum_{t=0}^{qK-1}\mathbb{E}\|v_t\| - \frac{3\gamma}{8}\sum_{t=0}^{qK-1}\mathbb{E}\|\nabla f(w_t)\| + \frac{3\gamma}{2}\sum_{t=0}^{qK-1}\mathbb{E}\|v_t - \nabla f(w_t)\| \\
&\le \frac{T\gamma\epsilon}{8} - \frac{\gamma}{2}\sum_{k=0}^{K-1}\sum_{s=0}^{q-1}\mathbb{E}\|v_{qk+s}\| - \frac{3\gamma}{8}\sum_{k=0}^{K-1}\sum_{s=0}^{q-1}\mathbb{E}\|\nabla f(w_{qk+s})\| + \frac{3\gamma}{2}\sum_{k=0}^{K-1}\sum_{s=0}^{q-1}\mathbb{E}\|\delta_{qk+s}\| \\
&\overset{(i)}{\le} \frac{T\gamma\epsilon}{8} - \frac{\gamma}{2}\sum_{k=0}^{K-1}\sum_{s=0}^{q-1}\mathbb{E}\|v_{qk+s}\| - \frac{3\gamma}{8}\sum_{k=0}^{K-1}\sum_{s=0}^{q-1}\mathbb{E}\|\nabla f(w_{qk+s})\| \\
&\quad + \frac{3q\gamma}{2}\sum_{k=0}^{K-1}\bigg(\frac{\Lambda}{\sqrt{B}} + \epsilon\sqrt{\frac{q}{B'}} + \Big(\frac{\epsilon}{\sqrt{B'}} + \frac{\Gamma}{\sqrt{B}}\Big)\sum_{u=0}^{q-1}\mathbb{E}\|\nabla f(w_{qk+u})\|\bigg) \\
&\overset{(ii)}{\le} \frac{T\gamma\epsilon}{4} - \frac{5\gamma}{16}\sum_{k=0}^{K-1}\sum_{s=0}^{q-1}\mathbb{E}\|\nabla f(w_{qk+s})\|,
\end{aligned} \tag{64}$$

where (i) uses eq. (59) and (ii) uses the following condition satisfied by the hyperparamter choices $B \ge \max(576\Lambda^2\epsilon^{-2}, 2304\Gamma^2q^2)$, $B' \ge \max(576q, 2304q^2\epsilon^2)$.

$$\frac{\Lambda}{\sqrt{B}} + \epsilon\sqrt{\frac{q}{B'}} \le \frac{\epsilon}{12}, \qquad \frac{\epsilon}{\sqrt{B'}} + \frac{\Gamma}{\sqrt{B}} \le \frac{1}{24q}$$

By rearranging eq. (64) and using $f(w_T) \ge f^* = \min_{w\in\mathbb{R}^d} f(w)$, we can prove eq. (22) as follows

$$\mathbb{E}\|\nabla f(w_{\widetilde{T}})\| = \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(w_t)\| \le \frac{16}{5T\gamma}\big(\mathbb{E}f(w_0) - f^*\big) + \frac{4\epsilon}{5}$$

It can be easily verified that the following hyperparameter choices satisfy the condition that $B \ge \max(576\Lambda^2\epsilon^{-2}, 2304\Gamma^2q^2)$ and that $B' \ge \max(576q, 2304q^2\epsilon^2)$ since $\epsilon \in (0,1)$.

$$q = \epsilon^{-1} = \mathcal{O}(\epsilon^{-1}) \tag{65}$$

$$B = \max(576\Lambda^2, 2304\Gamma^2)\epsilon^{-2} = \mathcal{O}(\epsilon^{-2}) \tag{66}$$

$$B' = 2304\epsilon^{-1} = \mathcal{O}(\epsilon^{-1}) \tag{67}$$

$$\gamma = \frac{\epsilon}{2\overline{K}_0 + 4\overline{K}_2 + 2\overline{K}_1(\Lambda^\alpha + \Gamma^\alpha + 1) + 1} = \mathcal{O}(\epsilon); \quad \text{if } \alpha \in (0,1) \tag{68}$$

$$\gamma = \frac{\epsilon}{5L_1\sqrt{\Gamma^2+1} + 8\sqrt{L_0^2 + 2L_1^2\Lambda^2}} = \mathcal{O}(\epsilon); \quad \text{if } \alpha = 1 \tag{69}$$

$$K = \frac{16\epsilon}{5T\gamma}\big(\mathbb{E}f(w_0) - f^*\big) = \mathcal{O}(\epsilon^{-1}) \tag{70}$$

$$T = qK = \frac{16}{5T\gamma}\big(\mathbb{E}f(w_0) - f^*\big) = \mathcal{O}(\epsilon^{-2}) \tag{71}$$

Substituting the choice of T given by eq. (71) into eq. (61), we obtain that $\mathbb{E}\|\nabla f(w_{\widetilde{T}})\| \le \epsilon$.

Under the above hyperparameter choices, the sample complexity is

$$\sum_{t=0}^{qK-1} |S_t| = K\big((q-1)B' + B\big) = \mathcal{O}(\epsilon^{-3}).$$