Trust Region Methods For Nonconvex Stochastic Optimization Beyond Lipschitz Smoothness*

Chenghan Xie† Chenxi Li‡ Chuwen Zhang § Qi Deng ¶ Dongdong Ge ‖ Yinyu Ye **

February 7, 2025

Abstract

In many important machine learning applications, the standard assumption of having a globally Lipschitz continuous gradient may fail to hold. This paper delves into a more general (L_0, L_1)-smoothness setting, which gains particular significance within the realms of deep neural networks and distributionally robust optimization (DRO). We demonstrate the significant advantage of trust region methods for stochastic nonconvex optimization under such generalized smoothness assumption. We show that first-order trust region methods can recover the normalized and clipped stochastic gradient as special cases and then provide a unified analysis to show their convergence to first-order stationary conditions. Motivated by the important application of DRO, we propose a generalized high-order smoothness condition, under which second-order trust region methods can achieve a complexity of $\mathcal{O}(\epsilon^{-3.5})$ for convergence to second-order stationary points. By incorporating variance reduction, the second-order trust region method obtains an even better complexity of $\mathcal{O}(\epsilon^{-3})$, matching the optimal bound for standard smooth optimization. To our best knowledge, this is the first work to show convergence beyond the first-order stationary condition for generalized smooth optimization. Preliminary experiments show that our proposed algorithms perform favorably compared with existing methods.

1 Introduction

We study the problem of minimizing a nonconvex function $F : \mathbb{R}^n \to \mathbb{R}$ which is expressed as the expectation of a stochastic function, i.e.,

$$\min_{x \in \mathbb{R}^n} \quad F(x) = \mathbb{E}_\xi[f(x;\xi)], \tag{1}$$

where the random variable ξ is realized according to a distribution $\mathcal{P}$. Over the years, substantial progress has been made in studying functions that possess Lipschitzian gradients, commonly referred to as L-smoothness functions. Notable contributions in this area can be found in (Ghadimi and Lan, 2013; Johnson and Zhang, 2013; Fang et al., 2018; Carmon et al., 2019), among others.

However, the assumption of Lipschitz smoothness may not hold in many important applications. For instance, in language models such as LSTM (Zhang et al., 2019) and transformers (Crawshaw et al., 2022), the function smoothness parameter can exhibit a strong correlation with the gradient norm along the training trajectory. Beyond these challenges in the standard Empirical Risk Minimization (ERM) framework, the L-smoothness condition could also easily fail in distributionally robust optimization (DRO) (Delage and Ye, 2010; Duchi and Namkoong, 2021; Levy et al., 2020a). DRO is particularly significant as it serves as a

*Chenghan Xie and Chenxi Li contribute equally

†Fudan University. Email: 20307130043@fudan.edu.cn
‡Shanghai University of Finance and Economics. Email: chenxili@stu.sufe.edu.cn
§Shanghai University of Finance and Economics. Email: chuwzhang@gmail.com
¶Shanghai University of Finance and Economics. Email: qideng@sufe.edu.cn
‖Shanghai University of Finance and Economics. Email: ge.dongdong@mail.shufe.edu.cn
**Stanford University. Email: yyye@stanford.edu

foundational element for ethical algorithms (Kearns and Roth, 2020) arising from accountability and fairness issues in machine learning (Fuster et al., 2022; Tang et al., 2023; Berk et al., 2021).

Motivated by this challenge, Zhang et al. (2019) introduced first-order generalized smoothness, also known as (L_0, L_1)-smoothness, where the Hessian norm is unbounded but allowed to grow linearly with the gradient norm. This condition can be further relaxed without the need of twice differentiability. Specifically, the (L_0, L_1)-smoothness condition (Zhang et al., 2020; Reisizadeh et al., 2023) is defined as

$$\|\nabla F(x) - \nabla F(x')\| \leq (L_0 + L_1\|\nabla F(x)\|)\, \|x - x'\| \tag{2}$$

holds for any $x, x' \in \mathbb{R}^n$ such that $\|x - x'\| \leq 1/L_1$, for constants $L_0 > 0, L_1 \geq 0$. Jin et al. (2021) showed that (L_0, L_1)-smoothness (2) holds for a broad class of DRO objectives when expressed in the dual form. Due to the difficulty in handling unbounded Lipschitz parameters, significant effort has been devoted to developing efficient algorithms under (L_0, L_1)-smoothness (Crawshaw et al., 2022; Reisizadeh et al., 2023; Wang et al., 2022). Typically, these works focus on developing more stable stepsizes for stochastic gradient descent through techniques like gradient clipping and step size normalization.

Despite these recent progresses, existing research remains limited to identifying approximate first-order stationary points (FOSP), which may be suboptimal in nonconvex settings. This drawback prompts the central question addressed in this paper: *Is it possible to develop an effective method capable of achieving approximate second-order stationary points under the conditions of generalized smoothness?*

In this paper, we firmly answer this question by proposing an algorithmic framework based on classical trust region methods (Sorensen, 1982; Conn et al., 2000). The crux of our method is to impose a trust region radius, which also coincides with the mutual concept of the aforementioned gradient-based methods. On one hand, this positions our method as a unifying analysis for gradient clipping and normalized gradient (Zhang et al., 2019; Jin et al., 2021) in which combinations of them can be derived. On the other hand, this framework naturally extends to finding second-order solutions if granted second-order derivatives. To our special interest, a second-order theory of generalized smoothness is proposed for DRO, which further empowers the complexity analysis of our framework. Our developments consist of four major steps:

- Firstly, we propose a unified trust region framework, under which the first-order variant, FOTRGS, unifies NSGD and clipped gradient methods with a weaker requirement of variance condition.
- Secondly, we propose a second-order theory of generalized smoothness and variance condition. We show that many divergence-based DRO problems with ψ-divergence satisfy our proposed assumptions.
- Thirdly, under the unified framework, we propose SOTRGS, namely, second-order trust region methods for generalized smoothness, and prove that it can achieve a second-order stationary point with $\mathcal{O}(\epsilon^{-3.5})$ sample complexity, which is better than first-order methods without variance-reduction techniques.
- Finally, we employ variance reduction techniques and propose SOTRGS-VR, demonstrating that identifying a second-order stationary point can be achieved in an optimal complexity of $\mathcal{O}(\epsilon^{-3})$.

A brief comparison of our methods and existing proposals are presented in Table 1. To our best knowledge, both the second-order generalized smoothness and convergence to SOSP are novel. In addition to the theoretical contribution, we conduct extensive experiments on DRO problems with imbalanced datasets, which justify the empirical advantage of our proposed methods.

2 Related Works

(L_0, L_1)-smoothness The concept of (L_0, L_1)-smoothness was first introduced by Zhang et al. (2019) to understand the superior performance of clipped algorithms over traditional non-adaptive gradient methods in natural language processing. Under the (L_0, L_1)-smoothness setting, Zhang et al. (2019) shows that normalized and clipped gradient methods converge to an ϵ-stationary point of the nonconvex objective function with at most $\mathcal{O}(\epsilon^{-4})$ gradient samples. This initiative sparked a series of follow-up studies, including Zhang et al. (2020); Qian et al. (2021); Zhao et al. (2021). Zhang et al. (2020) proposes a general framework which combines momentum acceleration with the clipped method. More recently, Reisizadeh et al. (2023)applies the variance reduced techniques to the clipped gradient method and improves the gradient complexity to $\mathcal{O}(\epsilon^{-3})$.

A parallel line of research has focused on analyzing algorithms that go beyond the normalized and clipping gradient methods in the (L_0, L_1)-smoothness setting. These include studies by Wang et al. (2022); Li et al. (2023a) on Adam, Crawshaw et al. (2022) on unclipped gradient methods, and more recently Sun et al. (2023) for (L_0, L_1)-smoothness in the variational inference problems. Another vein of research has sought to relaxed a heavy reliance on bounded variance assumptions; see Faw et al. (2023); Wang et al. (2023) and the references therein.

We are also aware of the works on even more general smoothness conditions based on (L_0, L_1)-smoothness. Chen et al. (2023) proposes a new notion of α-symmetric generalized smoothness, which is roughly as general as (L_0, L_1)-smoothness. Crawshaw et al. (2022) and Pan and Li (2023) provide a coordinate-wise type of (L_0, L_1)-smoothness. Li et al. (2023b) showed that classic first-order methods such as stochastic gradient and accelerated methods still have convergence guarantee under a mild ℓ-smoothness condition, which allows the Hessian norm to be bounded by a more general non-decreasing function $\ell(\|\nabla F(x)\|)$. Despite these advances, no previous work has contributed to the second-order generalization of (L_0, L_1)-smoothness for second-order stationary points.

Algorithm	**Smoothness**	**Complexity**	**Property**
SGD (Ghadimi and Lan, 2013)	Lipschitz	$\mathcal{O}(\epsilon^{-4})$	FOSP
SPIDER (Fang et al., 2018)	Lipschitz	$\mathcal{O}(\epsilon^{-3})$	FOSP
STR (Shen et al., 2019)	Lipschitz	$\mathcal{O}(\epsilon^{-3.5})$	SOSP
SCR (Tripuraneni et al., 2018)	Lipschitz	$\mathcal{O}(\epsilon^{-3.5})$	SOSP
ClippedSGD (Zhang et al., 2019)	FO-Generalized Smooth	$\mathcal{O}(\epsilon^{-4})$	FOSP
Clipped+ (Zhang et al., 2020)	FO-Generalized Smooth	$\mathcal{O}(\epsilon^{-4})$	FOSP
NSGD (Jin et al., 2021)	FO-Generalized Smooth	$\mathcal{O}(\epsilon^{-4})$	FOSP
(L_0, L_1)-SPIDER (Reisizadeh et al., 2023)	FO-Generalized Smooth	$\mathcal{O}(\epsilon^{-3})$	FOSP
FOTRGS	FO-Generalized Smooth	$\mathcal{O}(\epsilon^{-4})$	FOSP
FOTRGS-VR	FO-Generalized Smooth	$\mathcal{O}(\epsilon^{-3})$	FOSP
SOTRGS	**SO-Generalized Smooth**	$\mathcal{O}(\epsilon^{-3.5})$	SOSP
SOTRGS-VR	**SO-Generalized Smooth**	$\mathcal{O}(\epsilon^{-3})$	SOSP
Lower bound (Arjevani et al., 2020)	Lipschitz	$\Omega(\epsilon^{-3})$	SOSP

Table 1: Comparison of related algorithms. FOSP: First-order stationary point; SOSP: Second-order stationary point

Distributionally robust optimization Distributionally robust optimization (DRO) (Delage and Ye, 2010), originally designed for a middle ground between stochastic programming (Shapiro et al., 2014) and robust optimization (Ben-Tal et al., 2009), has attracted great interest in machine learning research communities in recent years for the purposes of distribution shifts and algorithmic fairness (Levy et al., 2020b; Duchi and Namkoong, 2021). For ϕ-divergence penalized DRO, Levy et al. (2020b) prove that it can be transformed into a stochastic optimization problem after duality arguments. Jin et al. (2021) later proves that it fits the settings (L_0, L_1)-smoothness that opens the possibility of a better understanding of first-order methods.

Trust region methods Trust region methods are renowned for their ability to reliably find second-order stationary points (Conn et al., 2000). For stochastic optimization, Shen et al. (2019) proposed a sample-efficient stochastic trust region (STR) algorithm for finite-sum minimization problems and achieved $\mathcal{O}(\sqrt{n}/\epsilon^{1.5})$ complexity to find $(\epsilon, \sqrt{\epsilon})$-SOSP. Other works (Curtis et al., 2019; Curtis and Shi, 2020) tackled the fully stochastic setting and proved they could achieve $\mathcal{O}(\epsilon^{-3.5})$ complexity to find $(\epsilon, \sqrt{\epsilon})$-SOSP. Trust region methods are also widely used in the real of policy optimization (Schulman et al., 2015; Liu et al., 2023). However, despite these advances, none of the previous studies have explored the properties of trust region methods under the generalized smoothness setting.

Variance reduction techniques Variance reduction techniques are first applied to accelerate the convergence speed of SGD for convex finite-sum optimization problems (Johnson and Zhang, 2013; Zhang et al., 2013; Wang et al., 2013). As for the non-convex setting, Stochastic variance-reduced gradient (SVRG) and Stochastically Controlled Stochastic Gradient (SCSG) improves the convergence rate to a first-order stationary point from $\mathcal{O}(\epsilon^{-4})$ to $\mathcal{O}(\epsilon^{-10/3})$ (Allen-Zhu and Hazan, 2016; Reddi et al., 2016; Lei et al., 2017). Recently, several new variance reduction techniques are able to achieve the optimal complexity rate of $\mathcal{O}(\epsilon^{-3})$ (Fang et al., 2018; Cutkosky and Orabona, 2019; Tran-Dinh et al., 2019; Liu et al., 2020; Li et al., 2021). In this paper, we use the techniques in Fang et al. (2018) to construct the variance-reduced trust region methods.

3 Preliminaries

Notations For a square matrix $A \in \mathbb{R}^{n\times n}$, we define norm for matrix as $\|A\| = \sqrt{\sigma_M}$, where σ_M is the eigenvalue of A^TA with largest absolute value. For a vector $v \in \mathbb{R}^n$, we use $\|v\|$ to express the standard Euclidean norm. $\|v\|_A := \sqrt{v^TAv}$ where A is a positive-definite matrix. We assert that objective function F is bounded below throughout the paper and define $F^* := \inf_x F(x) > -\infty$, $\Delta_F := F(x_0) - F^*$.

We review preliminary characteristics of (L_0, L_1)-smooth functions introduced in prior works. In the pioneer work (Zhang et al., 2019), a function F is said to be (L_0, L_1) smooth if there exist constants $L_0 > 0$ and $L_1 \geq 0$ such that for all $x \in \mathbb{R}^n$,

$$\left\|\nabla^2 F(x)\right\| \leq L_0 + L_1\|\nabla F(x)\|. \tag{3}$$

Note that the twice-differentiability assumption in this definition could be relaxed. Specifically, we adopt the (L_0, L_1)-smoothness assumption as follows:

Assumption 1. *((L_0, L_1)-smoothness). A differentiable function F is said to be (L_0, L_1)-smooth if there exist constants $L_0 > 0$, $L_1 \geq 0$ such that if $\|x - x'\| \leq 1/L_1$, then*

$$\|\nabla F(x) - \nabla F(x')\| \leq (L_0 + L_1\|\nabla F(x)\|)\,\|x - x'\|.$$

If F is twice differentiable, Assumption 1 implies condition (3). Moreover, condition (3) implies Assumption 1 with constants $(2L_0, 2L_1)$ (see (Reisizadeh et al., 2023)). We then state the required condition on the noise of the stochastic gradient.

Assumption 2. *((G_0, G_1)-bounded gradient variance) The stochastic gradient $\nabla f(\cdot;\xi)$ is unbiased and (G_0, G_1)-variance-bounded, that is,*

$$\mathbb{E}_\xi[\nabla f(x;\xi)] = \nabla F(x),$$
$$\mathbb{E}_\xi\|\nabla f(x;\xi) - \nabla F(x)\|^2 \leq G_0^2 + G_1^2\|\nabla F(x)\|^2.$$

Note that (G_0, G_1)-bounded variance is more general than the standard bounded variance assumption $\mathbb{E}_\xi\|\nabla f(x;\xi) - \nabla F(x)\|^2 \leq \sigma^2$. We extend standard assumptions to Assumption 2 following Faw et al. (2023). In addition, one can verify that DRO satisfies Assumption 1 and 2; for details, see Section 4.3.

Let $\mathcal{S}$ be the batch of samples. We define the batch stochastic component function by

$$f(x;\mathcal{S}) := \frac{1}{|\mathcal{S}|}\sum_{\xi\in\mathcal{S}} f(x;\xi).$$

Our goal is to find first-order and second-order stationary points defined as follows.

Definition 1. *We say that x is a first-order approximate stationary point (ϵ-FOSP) of $F(\cdot)$ if*

$$\|\nabla F(x)\| \leq c_1 \cdot \epsilon.$$

We say that x is a second-order approximate stationary point ($(\epsilon, \sqrt{\epsilon})$-SOSP) of $F(\cdot)$ if

$$\|\nabla F(x)\| \leq c_1 \cdot \epsilon,\ \lambda_{\min}(\nabla^2 F(x)) \geq -c_2 \cdot \sqrt{\epsilon}$$

for some positive constants $c_1, c_2 > 0$.

DRO Instead of assuming a known underlying probability distribution, DRO minimizes the worst-case loss over a set of distributions Q around the original distribution P. This can be formally stated as the following problem (Delage and Ye, 2010; Rahimian and Mehrotra, 2019; Shapiro, 2017):

$$\min_{x\in\mathbb{R}^n} \quad \Psi(x) := \sup_{Q\in\mathcal{U}(P)} \mathbb{E}_{\xi\sim Q}[\ell(x;\xi)],$$

Here, ξ is some random sample and $\ell(x,\xi)$ stands for the stochastic loss function. The uncertainty set $\mathcal{U}(P)$ with respect to certain distance measure d is defined as $\mathcal{U}(P) := \{Q : d(Q,P) \leq r\}$.

Another popular and equivalent formulation of DRO is to add a regularization term rather than imposing the uncertainty set constraints, which leads to the penalized DRO form:

$$\min_{x\in\mathbb{R}^n} \quad \Psi(x) := \sup_{Q} \left\{\mathbb{E}_{\xi\sim Q}[\ell(x;\xi)] - \lambda d(Q,P)\right\}, \tag{4}$$

where $\lambda > 0$ is the prespecified regularization weight. In this paper, we adopt the widely used ψ-divergence (Shapiro, 2017). The ψ-divergence between Q and P is defined as $d_\psi(Q,P) := \int \psi\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)\mathrm{d}P$, where ψ is a valid divergence function, namely, ψ is non-negative, and it satisfies $\psi(1) = 0$ and $\psi(t) = +\infty$ for all $t < 0$. The conjugate function ψ^* is defined as $\psi^*(t) := \sup_{s\in\mathbb{R}}(st - \psi(s))$.

4 Methodology

In this section, we first propose a unified trust region framework for generalized smoothness. Then, by specifying a scaling matrix, we give a general first-order trust region algorithm that covers normalized gradient and clipped gradient methods. Moreover, we devote ourselves to a second-order theory of smoothness based on which a second-order trust region method is introduced. We also extend our framework to include variance-reduced versions for both first-order and second-order trust region methods. Lastly, we discuss inexact second-order variants to facilitate scalable implementations. For the sake of brevity, we have relegated all proofs of the theoretical results to the appendix.

4.1 A Unified Trust Region Framework for Generalized Smoothness

We now introduce our unified trust region framework for generalized smoothness, as described in Algorithm 1. In each iteration, the framework involves solving the following constrained quadratic subproblem

$$\begin{aligned} \min_{d\in\mathbb{R}^n} \quad & m_t(d) := F(x_t) + g_t^T d + \frac{1}{2} d^T B_t d \\ \text{s.t.} \quad & \|d\| \leq \Delta_t, \end{aligned} \tag{5}$$

It is important to note that the square matrix B_t is not predetermined in this abstract framework. By making different choices for B_t, we can develop more specific first- and second-order methods under this unified framework. For example, both normalized gradient and clipped gradient can be viewed as a special case with a certain choice of B_t. As our analysis will demonstrate, when only first-order information is available, the trust-region algorithm guarantees convergence as long as B_t has a bounded norm. Furthermore, leveraging second-order information can enhance our convergence towards high-order optimality conditions.

Algorithm 1 The trust region framework

```
Given T, error ϵ
for t = 0, 1, ..., T − 1 do
    Draw samples S_1 and compute g_t = ∇f(x_t; S_1)
    (if needed) Draw samples S_2 and compute H_t = ∇²f(x_t; S_2)
    Compute step d_{t+1} by solving the subproblem (5)
    Update: x_{t+1} ← x_t + d_{t+1}
end for
```

For generality, we first provide some important properties about the solution of subproblem (5). By the optimality condition of subproblem(Conn et al. (2000), the vector d_{t+1} is the global solution to problem 5 if and only if there exists a Lagrange multiplier λ_t such that (d_{t+1}, λ_t) is the solution to the following equations:

$$(B_t + \lambda I)d + g_t = 0, \lambda(\Delta_t - \|d\|) = 0, (B_t + \lambda I) \succeq 0 \tag{6}$$

Lemma 1 (Model reduction)**.** *For any matrix variable B_t, at the t-th iteration, let d_{t+1} and λ_t be the optimal primal and dual solution of* (6)*. We have the following amount of decrease on m_t*

$$m_t(d_{t+1}) - m_t(0) \leq -\frac{1}{2}\lambda_t \|d_{t+1}\|^2.$$

4.2 First-Order Trust Region Methods

We first consider the first-order trust region method for generalized smoothness, FOTRGS, where only gradient information is used in Algorithm 1. We show that as long as $\|B_t\|$ is uniformly bounded by a constant, by setting proper parameters, Algorithm 1 is able to return an ϵ-FOSP.

Theorem 1 (Sample complexity of FOTRGS)**.** *Suppose Assumption 1 - 2 hold. Let B_t be a matrix with bounded norm i.e. there exists a constant β such that $\|B_t\| \leq \beta$. By setting $\epsilon \leq \min\left\{\frac{4L_0G_0+16\beta G_0}{L_1G_0+2L_0G_1+8\beta G_1}, \frac{4L_0+16\beta}{L_1}\right\}$, $\Delta_t = \Delta = (4L_0+16\beta)^{-1}\epsilon$, $|\mathcal{S}_1| = 64G_0^2\epsilon^{-2}$, $T = 32\Delta_F(L_0+4\beta)\epsilon^{-2}$ in Algorithm 1, we have $\mathbb{E}\|\nabla F(x_{\bar{t}})\| \leq \epsilon$, where $\bar{t}$ is sampled from $\{0, 1, \ldots, T-1\}$ uniformly at random. Moreover, the sample complexity of finding an ϵ-FOSP is bounded by*

$$\mathcal{O}\left(\frac{\Delta_F(L_0+\beta)G_0^2}{\epsilon^4}\right).$$

When fixing B_t as specific constants, we are able to represent the normalized and clipped gradient method in this framework. To be specific, if we set $B_t = 0$, then we are able to cover the normalized gradient descent method in trust region framework.

Corollary 1 (Equivalence to the normalized method)**.** *Suppose Assumption 1 - 2 hold. Let $B_t = 0$ in Algorithm 1, then the solution of the subproblem* (5) *is*

$$d_{t+1} = \frac{\Delta_t}{\|g_t\|} \cdot (-g_t).$$

By setting $\epsilon \leq \min\left\{\frac{4L_0G_0}{L_1G_0+2L_0G_1}, \frac{4L_0}{L_1}\right\}$, $\Delta_t = \Delta = (4L_0)^{-1}\epsilon$, $|\mathcal{S}_1| = 64G_0^2\epsilon^{-2}$, $T = 32\Delta_F L_0\epsilon^{-2}$, we have $\mathbb{E}\|\nabla F(x_{\bar{t}})\| \leq \epsilon$, where $\bar{t}$ is sampled from $\{0, 1, ..., T-1\}$ uniformly at random. Moreover, the sample complexity of finding an ϵ-FOSP is bounded by $\mathcal{O}\left(\frac{\Delta_F L_0 G_0^2}{\epsilon^4}\right)$.

By setting $B_t = \rho I$, we are also able to represent the clipped method in this unified framework.

Corollary 2 (Equivalence to the clipped method)**.** *Suppose Assumption 1 - 2 hold. Let $B_t = \rho I$ in Algorithm 1, then the solution of the subproblem* (5) *is*

$$d_{t+1} = \min\left\{\frac{\Delta_t}{\|g_t\|}, \frac{1}{\rho}\right\} \cdot (-g_t).$$

By setting $\epsilon \leq \min\left\{\frac{4L_0G_0+16\rho G_0}{L_1G_0+2L_0G_1+8\rho G_1}, 4L_0 + 16\rho L_1^{-1}\right\}$, $\Delta_t = \Delta = (4L_0 + 16\rho)^{-1}\epsilon$, $|\mathcal{S}_1| = 64G_0^2\epsilon^{-2}$, $T = 32\Delta_F(L_0+4\rho)\epsilon^{-2}$ in Algorithm 1, we have $\mathbb{E}\|\nabla F(x_{\bar{t}})\| \leq \epsilon$, where $\bar{t}$ is sampled from $\{0, 1, \ldots, T-1\}$ uniformly at random. Moreover, the sample complexity of finding an ϵ-FOSP is bounded by $\mathcal{O}\left(\frac{\Delta_F(L_0+\rho)G_0^2}{\epsilon^4}\right)$.

A few remarks are in order. First, it's worth noting that our proposed first-order trust-region method offers greater flexibility in step size compared to normalized and clipped gradient methods, as we can choose different B_t values in each iteration. Exploring more choices for B_t remains an interesting direction for future research. Second, our complexity results closely align with some recent work. For instance, the prior work (Reisizadeh et al., 2023) has analyzed the convergence rate of the clipped method. Under similar assumptions, an ϵ-FOSP can be found by the clipped method with $\mathcal{O}(\epsilon^{-4})$ gradient samples. A key distinction between our analysis and prior work lies in the variance bound requirements on stochastic gradients. Specifically, while Reisizadeh et al. (2023) requires a uniform variance bound $\mathbb{E}_\xi\|\nabla f(x;\xi) - \nabla F(x)\|^2 \leq \sigma^2$, we allow for a variance bound related to the gradient norm of the current point, as stated in Assumption 2. This makes our analysis more general and extends its applicability to the DRO setting.

4.3 A Second-Order Theory of Generalized Smoothness

This subsection introduces a generalized second-order smoothness condition, drawing inspiration from the (L_0, L_1)-smoothness concept. Subsequently, we demonstrate that DRO is a significant application that aligns with this newly proposed second-order condition.

Assumption 3 (Second-order generalized smoothness and variance condition)**.** *F is twice-differentiable and satisfies that there exist constants $\delta > 0$, $M_0 > 0$ and $M_1 \geq 0$ such that if $\|x - x'\| \leq \delta$, then*

$$\|\nabla^2 F(x) - \nabla^2 F(x')\| \leq (M_0 + M_1\|\nabla F(x)\|)\|x - x'\|.$$

Moreover, the stochastic Hessian is unbiased and (K_0, K_1) variance-bounded, that is,

$$\mathbb{E}_\xi[\nabla^2 f(x;\xi)] = \nabla^2 F(x),$$
$$\mathbb{E}_\xi\|\nabla^2 f(x;\xi) - \nabla^2 F(x)\|^2 \leq K_0^2 + K_1^2\|\nabla F(x)\|^2.$$

Similar to the (L_0, L_1)-smoothness, we can interpret the proposed second-order generalized smoothness from the perspective of the boundness of higher-order derivatives. Further discussion of this condition can be found in the appendix. We claim that Penalized DRO (4) satisfies this assumption. The original formulation involves a max operation over distributions, which makes optimization challenging. By duality arguments (see details in Levy et al. (2020b, Section A.1.2)), we can write (4) equivalently as

$$\Psi(x) = \min_{\eta \in \mathbb{R}} \mathcal{L}(x, \eta) := \lambda \mathbb{E}_{\xi \sim P} \psi^* \left(\frac{\ell(x;\xi) - \eta}{\lambda} \right) + \eta. \tag{7}$$

This suggests that to minimize the DRO objective, one can perform a joint minimization of $\mathcal{L}(x, \eta)$ over $(x, \eta) \in \mathbb{R}^{n+1}$. Crucially, it is sufficient to find an $(\epsilon, \sqrt{\epsilon})$-SOSP of $\Psi(x)$ by optimizing $\mathcal{L}(x, \eta)$ instead. To establish this relationship more formally, we build the connection between the gradient and Hessian of $\Psi(x)$ and those of $\mathcal{L}(x, \eta)$ as follows.

Theorem 2. *Under mild assumptons for ℓ and ψ^*, if some (x, η) is a $(\epsilon, \sqrt{\epsilon})$-SOSP for $\mathcal{L}(x, \eta)$, then x is also a $(\epsilon, \sqrt{\epsilon})$-SOSP for $\Psi(x)$.*

The following theorem analyzes the smoothness and variance properties of $\mathcal{L}(x, \eta)$, which motivates us to propose our second-order generalized smoothness and variance conditions.

Theorem 3. *Under mild assumptons for ℓ and ψ^*, the objective $\mathcal{L}(x, \eta)$, serving as F, satisfies Assumption 1, 2 and 3.*

4.4 Second-Order Trust Region Methods

We propose SOTRGS by setting $B_t = H_t$ in Algorithm 1. We first present a result on bounding the variance of Hessian.

Lemma 2 (Variance bounds on Hessian estimators)**.** *Suppose that Assumption 3 holds in Algorithm 1, if we set $|\mathcal{S}_2| = 22\log(n)\epsilon^{-1}$, then*

$$\mathbb{E}_t\left[\|H_t - \nabla^2 F(x_t)\|^2\right] \leq (K_0^2 + K_1^2\|\nabla F(x_t)\|^2)\epsilon,$$

where $\mathbb{E}_t$ denotes the expectation conditioned on all the randomness before the t-th iteration.

Next, we provide the convergence result of the second-order trust region method in the generalized smoothness setting.

Theorem 4 (Sample complexity of SOTRGS)**.** *Suppose Assumptions 1, 2 and 3 hold. Let* $\Delta_t = \Delta = \sqrt{\epsilon}$*, by setting* $B_t = H_t$*,* $\epsilon < \min\left\{\frac{3}{5M_1+18G_1+12K_1}, \frac{1}{L_1^2}\right\}$*,* $|\mathcal{S}_1| = \epsilon^{-2}$*,* $|\mathcal{S}_2| = 22\log(n)\epsilon^{-1}$*,* $T = \mathcal{O}(\epsilon^{-3/2})$ *in Algorithm 1, we have*

$$\mathbb{E}[\|\nabla F\left(x_{\bar{t}+1}\right)\|] \leq \mathcal{O}(\epsilon), \mathbb{E}[\lambda_{\min}(\nabla^2 F(x_{\bar{t}+1}))] \geq -\mathcal{O}(\sqrt{\epsilon}),$$

where $\bar{t}$ *is sampled from* $\{0, 1, \dots, T-1\}$ *uniformly at random. Moreover, the sample complexity of finding an* $(\epsilon, \sqrt{\epsilon})$*-SOSP is bounded by*

$$\mathcal{O}\left(\frac{\Delta_F}{\epsilon^{7/2}} + \frac{\Delta_F}{\epsilon^{5/2}}\right).$$

To our best knowledge, this is the first work to show convergence achieving the second-order stationary points for generalized smooth optimization, and its sample complexity is better than first-order methods without variance reduction techniques.

4.5 Variance Reduction

We now turn our attention to the variance-reduced variants of the trust-region method. Arjevani et al. (2020) shows that for any $L, \sigma > 0$, there exists a function F of the form (1) satisfying σ^2 bounded variance and expected Lipschitz smoothness with stochastic gradients $\nabla f(\cdot; \xi)$ such that

$$\mathbb{E}_\xi[\nabla f(x;\xi)] = \nabla F(x), \quad \mathbb{E}_\xi\|\nabla f(x;\xi) - \nabla F(x)\|^2 \leq \sigma^2,$$

and

$$\mathbb{E}_\xi\left[\|\nabla f(x;\xi) - \nabla f(x';\xi)\|^2\right]^{1/2} \leq L\|x - x'\|,$$

for which finding an ϵ-stationary solution requires $\Omega\left(\sigma\epsilon^{-3} + \sigma^2\epsilon^{-2}\right)$ stochastic gradient queries. Since our generalized smoothness is more general than its requirements, the lower bound can be directly applied to our settings. To close the optimality gap, we employ a variance reduction technique (Fang et al., 2018) to construct an improved gradient estimator g_t. Specifically, if $\mathrm{mod}(t, q) = 0$, then we take

$$g_t = \nabla f(x_t; \mathcal{S}_1);$$

otherwise, we compute g_t based on the value of g_{t-1}

$$g_t = \nabla f(x_t; \mathcal{S}_3) - \nabla f(x_{t-1}; \mathcal{S}_3) + g_{t-1},$$

where $\mathcal{S}_1, \mathcal{S}_3$ and q are parameters to be determined. The abstract variance-reduced trust region framework is presented in Algorithm 2. Next, we develop the sample complexity of both first-order and second-order variance reduced methods.

First-order methods We apply the above gradient estimator and propose a variance-reduced first-order trust region method, FOTRGS-VR. We give the upper bound of sample complexity for finding an ϵ-FOSP in the following theorem.

Theorem 5. *Suppose Assumption 1, 2 and 8 hold. Let* B_t *be a positive semi-definite matrix with bounded norm i.e. there exists a constant* β *such that* $\|B_t\| \leq \beta$*. By setting* $\epsilon \leq \min\left\{\frac{G_1^2}{2L_1^2}, \frac{1}{L_1}\right\}$*,* $\Delta_t = \Delta = \epsilon$*,* $|\mathcal{S}_1| = \epsilon^{-2}$*,* $|\mathcal{S}_3| = \epsilon^{-1}$*,* $q = (8G_1\epsilon)^{-1}$*,* $T = \mathcal{O}(\epsilon^{-2})$*, then we have* $\mathbb{E}\|\nabla F(x_{\bar{t}})\| \leq \mathcal{O}(\epsilon)$*, where* $\bar{t}$ *is sampled from* $\{0, 1, \dots, T-1\}$ *uniformly at random. Moreover, the total complexity of finding an* ϵ*-FOSP is bounded by*

$$\mathcal{O}\left(\frac{\Delta_F}{\epsilon^3}\right).$$

Second-order methods To reduce the second-order oracle complexity, we apply the same idea to both the gradient and Hessian estimator in the second-order trust region method. Similar to the analysis of the first-order variance-reduced trust region method, the following theorem gives the upper bound of sample complexity for finding an $(\epsilon, \sqrt{\epsilon})$-SOSP.

Theorem 6 (Sample complexity of SOTRGS-VR)**.** *Suppose Assumption 1, 2, 3 and 8 hold. Let B_t be the Hessian estimator as shown in Algorithm 2. By setting $\epsilon \leq \min\left\{\frac{G_1^4}{4L_1^4}, \frac{1}{36G_1^2}, \frac{1}{L_1^2}\right\}$, $\Delta_t = \Delta = \sqrt{\epsilon}$, $|\mathcal{S}_1| = \epsilon^{-2}$, $|\mathcal{S}_2| = 22\log(n)\epsilon^{-1}$, $|\mathcal{S}_3| = \epsilon^{-3/2}$, $T = \mathcal{O}(\epsilon^{-3/2})$, then we have $\mathbb{E}\|\nabla F(x_{\bar{t}+1})\| \leq \epsilon$, $\mathbb{E}[\lambda_{\min}(\nabla^2 F(x_{\bar{t}+1}))] \geq -\mathcal{O}(\sqrt{\epsilon})$, where $\bar{t}$ is sampled from $\{0, 1, \ldots, T-1\}$ uniformly at random. Moreover, the total complexity of finding an $(\epsilon, \sqrt{\epsilon})$-SOSP is bounded by*

$$\mathcal{O}\left(\frac{\Delta_F}{\epsilon^3} + \frac{\Delta_F}{\epsilon^{5/2}}\right).$$

Algorithm 2 Variance-reduced trust region method

Given T, error ϵ
for $t = 0, 1, \ldots, T-1$ **do**
  if $\mathrm{mod}(t, q) = 0$ **then**
    Draw samples $\mathcal{S}_1$ and compute $g_t = \nabla f(x_t; \mathcal{S}_1)$
  else
    Draw samples $\mathcal{S}_3$ and compute $g_t = g_{t-1} + \nabla f(x_t; \mathcal{S}_3) - \nabla f(x_{t-1}; \mathcal{S}_3)$
  end if
  (if needed) Draw samples $\mathcal{S}_2$ and compute $H_t = \nabla^2 f(x_t; \mathcal{S}_2)$
  Compute step d_{t+1} by solving the subproblem (5)
  Update: $x_{t+1} \leftarrow x_t + d_{t+1}$
end for

4.6 Inexactness and Scalability

For large-scale machine learning problems, exactly solving the second-order trust region subproblem (5) can be computationally prohibitive. To mitigate this, we can relax the need for exact Hessian calculations and subproblem solutions by allowing for inexact approximations. In the sequel, we assume $\tilde{\nabla}^2 F(x)$ is the approximation of $\nabla^2 F(x)$. At each x_t we adopt a low-dimensional subspace with orthonormal basis $V_t \in \mathbb{R}^{n\times k}$ for $k \ll n$, and compute second-order derivatives in the subspace. Inspired by the work of Cartis et al. (2011); Zhang et al. (2022), we propose the following regularity assumption on inexactness in Hessian approximation.

Assumption 4. *For certain constants $C_0, C_1 > 0$, there exists a V_t whose columns form an orthonormal basis such that*

$$\|(\nabla^2 F(x_t) - \tilde{\nabla}^2 F(x_t))d_{t+1}\| \leq (C_0 + C_1\|\nabla F(x_t)\|)\|d_{t+1}\|^2,$$

where $\tilde{\nabla}^2 F(x) := V_t V_t^T \nabla^2 F(x) V_t V_t^T$ is the projected Hessian in the column space of V_t.

Setting $B_t = \tilde{H}_t := V_t V_t^{\mathrm{T}} H_t V_t V_t^{\mathrm{T}}$ and then using the auxiliary variable $y = V_t^T x^t$, Algorithm 1 only needs to solve an approximate trust-region subproblem with a much lower dimension. Theoretically, our new assumption can be satisfied in various ways (Xu et al., 2020; Cartis et al., 2022). We leave the details in Appendix G.

The following theorem provides the performance bound of the inexact version of the second-order trust region method.

Theorem 7. *Suppose Assumptions 1, 2, 3 and 4 hold. Let $B_t = V_t {V_t}^T H_t V_t V_t^T$. In Algorithm 1, let $\epsilon < \min\left\{\frac{3}{5M_1+18G_1+24K_1+6C_1}, \frac{1}{L_1^2}\right\}$, $\Delta_t = \Delta = \sqrt{\epsilon}$, $|\mathcal{S}_1| = \epsilon^{-2}$, $|\mathcal{S}_2| = 22\log(n)\epsilon^{-1}$, $T = \mathcal{O}(\epsilon^{-3/2})$, then we have*

$$\mathbb{E}[\|\nabla F(x_{\bar{t}+1})\|] \leq \mathcal{O}(\epsilon),\ \mathbb{E}[\lambda_{\min}(\nabla^2 \tilde{F}(x_{\bar{t}+1}))] \geq -\mathcal{O}(\sqrt{\epsilon}),$$

where $\bar{t}$ is sampled from $\{0, 1, \ldots, T-1\}$ uniformly at random.

5 Experiments

We perform three sets of experiments in machine learning with a focus on DRO to justify our analysis. Due to space limitation, we only present a brief description and the tuned methods with the best performance for SGD, FOTRGS, and SOTRGS; complete details are left in Appendix G.

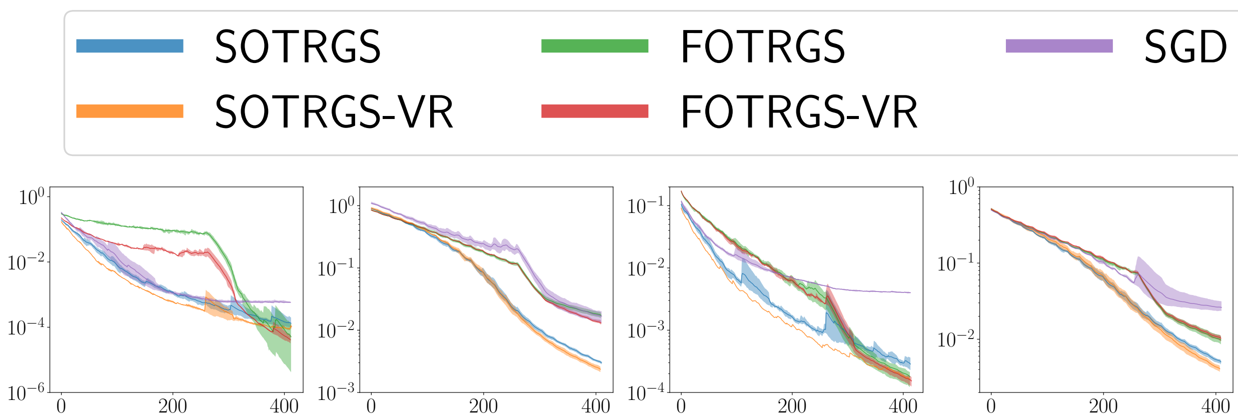
5.1 Basic Settings

We focus on classification tasks with **imbalanced** distributions arising from applications with heterogeneous (but often latent) subpopulations. Since in standard datasets like MNIST, Fashion MNIST and CIFAR-10, the population ratios (number of images per class) are the same, we create a perturbed dataset that inherits a disparity (Hashimoto et al., 2018) by choosing only a subset of training samples for each one of the categories. Since all these datasets consist of 10 categories, we fix them at a uniform set of levels without loss of generality. In all the tests, the worst class only takes a proportion of 0.254 from the samples; after preprocessing, we only use $33,260$ out of the original $50,000$ training samples.

We adopt penalized DRO for classification tasks with two specific divergence functions satisfying Assumption 3: the smoothed χ^2 and smoothed CVaR. To fairly compare the algorithms, we perform a grid search over the parameters. The complete description is left in Appendix H.

5.2 Experiment Results

The results show the trust region methods are efficient in DRO with second-order generalized smoothness in training efficiency and test accuracy, especially for minority classes. In all our experiments, we do not differentiate between smoothed and the original divergence functions and may use them interchangeably. Figure



(a) MNIST with χ^2 (b) F-M with χ^2 (c) MNIST with CVaR (d) F-M with CVaR

Figure 1: Training curves with different smoothed DRO loss on imbalanced MNIST and Fashion-MNIST datasets. We report the per-step losses by aggregating every 20 iteration. Shaded areas indicate the range of variability across 5 repetitions.

(1a) and (1b) present the training curves of SGD, first-order (FOTRGS) and second-order (SOTRGS) trust region methods on MNIST and Fashion MNIST datasets, respectively. The normalized SGD outperforms standard SGD as a representative of the FOTRGS family. Furthermore, it is clear that SOTRGS accelerates the rate of convergence in all our tests. We leave test results in Appendix H.

Table (2a) and (2b) presents the test accuracy of different methods. It is clear that the trust region methods have an advantage in preserving fairness for minority classes while also achieving the best overall average performance.

	Worst Category	Overall Accuracy
SOTRGS	0.681	0.889
FOTRGS	0.705	0.898
SGD	0.629	0.894

(a) Imbalanced CIFAR10 with χ^2 loss.

	Worst Category	Overall Accuracy
SOTRGS	0.616	0.896
FOTRGS	0.615	0.899
SGD	0.607	0.888

(b) Imbalanced CIFAR10 with CVaR loss

Table 2: Test accuracy on imbalanced CIFAR10. Besides overall test accuracy, we also present the worst-performing class indicated as the "worst category".

6 Discussion

This work opens up several intriguing avenues for future exploration. One primary question that emerges is whether machine learning problems, beyond DRO, exhibit properties of second-order generalized smoothness. Additionally, it would be interesting to see whether our framework can be extended to more complex scenarios, which involve either a constrained domain or an additional non-smooth proximal term in the objective.

References

Z. Allen-Zhu and E. Hazan. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pages 699–707. PMLR, 2016.

Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, A. Sekhari, and K. Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299. PMLR, 2020.

A. S. Bandeira, K. Scheinberg, and L. N. Vicente. Convergence of Trust-Region Methods Based on Probabilistic Models. *SIAM Journal on Optimization*, 24(3):1238–1264, Jan. 2014. ISSN 1052-6234. doi: 10.1137/130915984. URL https://epubs.siam.org/doi/abs/10.1137/130915984. Publisher: Society for Industrial and Applied Mathematics.

A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust optimization*. 2009. ISBN 978-0-691-14368-2. doi: 10.1016/s1474-6670(17)42591-2. Publication Title: Robust Optimization.

A. S. Berahas, R. Bollapragada, and J. Nocedal. An investigation of Newton-Sketch and subsampled Newton methods. *Optimization Methods and Software*, 35(4):661–680, July 2020. ISSN 1055-6788. doi: 10.1080/10556788.2020.1725751. URL https://doi.org/10.1080/10556788.2020.1725751. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10556788.2020.1725751.

R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50(1):3–44, Feb. 2021. ISSN 0049-1241. doi: 10.1177/0049124118782533. URL https://doi.org/10.1177/0049124118782533. Publisher: SAGE Publications Inc.

R. Bollapragada, R. H. Byrd, and J. Nocedal. Exact and inexact subsampled Newton methods for optimization. *IMA Journal of Numerical Analysis*, 39(2):545–578, 2019. Publisher: Oxford University Press.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower Bounds for Finding Stationary Points I, Aug. 2019. URL http://arxiv.org/abs/1710.11606. arXiv:1710.11606 [math].

C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, 127 (2):245–295, Apr. 2011. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-009-0286-5. URL http://link.springer.com/10.1007/s10107-009-0286-5.

C. Cartis, N. I. M. Gould, and P. L. Toint. *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives*. Society for Industrial and Applied Mathematics, Philadelphia, PA, Jan. 2022. ISBN 978-1-61197-698-4 978-1-61197-699-1. doi: 10.1137/1.9781611976991. URL https://epubs.siam.org/doi/book/10.1137/1.9781611976991.

X. Chen. Superlinear convergence of smoothing quasi-Newton methods for nonsmooth equations. *Journal of Computational and Applied Mathematics*, 80(1):105–126, Apr. 1997. ISSN 0377-0427. doi: 10.1016/S0377-0427(97)80133-1. URL https://www.sciencedirect.com/science/article/pii/S0377042797801331.

X. Chen. Smoothing methods for nonsmooth, nonconvex minimization. *Mathematical Programming*, 134(1): 71–99, Aug. 2012. ISSN 1436-4646. doi: 10.1007/s10107-012-0569-0. URL https://doi.org/10.1007/s10107-012-0569-0.

Z. Chen, Y. Zhou, Y. Liang, and Z. Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. *arXiv preprint arXiv:2303.02854*, 2023.

A. R. Conn, N. I. Gould, and P. L. Toint. *Trust region methods*. SIAM, 2000.

M. Crawshaw, M. Liu, F. Orabona, W. Zhang, and Z. Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in Neural Information Processing Systems*, 35:9955–9968, 2022.

F. E. Curtis and R. Shi. A fully stochastic second-order trust region method. *Optimization Methods and Software*, pages 1–34, 2020.

F. E. Curtis, K. Scheinberg, and R. Shi. A stochastic trust region algorithm based on careful step normalization. *Informs Journal on Optimization*, 1(3):200–220, 2019.

A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3), 2010. doi: 10.1287/opre.1090.0741.

J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, June 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS2004. URL https://projecteuclid.org/journals/annals-of-statistics/volume-49/issue-3/Learning-models-with-uniform-performance-via-distributionally-robust-optimization/10.1214/20-AOS2004.full. Publisher: Institute of Mathematical Statistics.

C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

M. Faw, L. Rout, C. Caramanis, and S. Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive sgd. *arXiv preprint arXiv:2302.06570*, 2023.

A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther. Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47, 2022. Publisher: Wiley Online Library.

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

T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness Without Demographics in Repeated Loss Minimization. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1929–1938. PMLR, July 2018. URL https://proceedings.mlr.press/v80/hashimoto18a.html. ISSN: 2640-3498.

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

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

M. Kearns and A. Roth. *The ethical algorithm: the science of socially aware algorithm design.* Oxford University Press, New York, 2020. ISBN 978-0-19-094820-7.

L. Lei, C. Ju, J. Chen, and M. I. Jordan. Non-convex finite-sum optimization via scsg methods. *Advances in Neural Information Processing Systems*, 30, 2017.

D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in neural information processing systems*, volume 33, pages 8847–8860. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/64986d86a17424eeac96b08a6d519059-Paper.pdf.

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

H. Li, A. Jadbabaie, and A. Rakhlin. Convergence of adam under relaxed assumptions. *arXiv preprint arXiv:2304.13972*, 2023a.

H. Li, J. Qian, Y. Tian, A. Rakhlin, and A. Jadbabaie. Convex and non-convex optimization under generalized smoothness. *arXiv preprint arXiv:2306.01264*, 2023b.

Z. Li, H. Bao, X. Zhang, and P. Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, pages 6286–6295. PMLR, 2021.

D. Liu, L. M. Nguyen, and Q. Tran-Dinh. An optimal hybrid variance-reduced algorithm for stochastic composite nonconvex optimization. *arXiv preprint arXiv:2008.09055*, 2020.

J. Liu, C. Xie, Q. Deng, D. Ge, and Y. Ye. Stochastic dimension-reduced second-order methods for policy optimization. *arXiv preprint arXiv:2301.12174*, 2023.

L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. 2014.

J. L. Nazareth. Homotopy techniques in linear programming. *Algorithmica*, 1(1):529–535, Nov. 1986. ISSN 1432-0541. doi: 10.1007/BF01840461. URL https://doi.org/10.1007/BF01840461.

Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Y. Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

J. Nocedal and S. J. Wright. *Numerical optimization.* Springer, 1999.

Y. Pan and Y. Li. Toward understanding why adam converges faster than sgd for transformers. *arXiv preprint arXiv:2306.00204*, 2023.

J. Qian, Y. Wu, B. Zhuang, S. Wang, and J. Xiao. Understanding gradient clipping in incremental gradient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 1504–1512. PMLR, 2021.

H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

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

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

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

A. Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

Z. Shen, P. Zhou, C. Fang, and A. Ribeiro. A stochastic trust region method for non-convex minimization. *arXiv preprint arXiv:1903.01540*, 2019.

D. C. Sorensen. Newton's method with a model trust region modification. *SIAM Journal on Numerical Analysis*, 19(2):409–426, 1982.

L. Sun, A. Karagulyan, and P. Richtarik. Convergence of stein variational gradient descent under a weaker smoothness condition. In *International Conference on Artificial Intelligence and Statistics*, pages 3693–3717. PMLR, 2023.

Z. Tang, J. Zhang, and K. Zhang. What-Is and How-To for Fairness in Machine Learning: A Survey, Reflection, and Perspective. *ACM Computing Surveys*, 55(13s):1–37, 2023. Publisher: ACM New York, NY.

Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *arXiv preprint arXiv:1905.05920*, 2019.

N. Tripuraneni, M. Stern, C. Jin, J. Regier, and M. I. Jordan. Stochastic cubic regularization for fast nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

B. Wang, Y. Zhang, H. Zhang, Q. Meng, Z.-M. Ma, T.-Y. Liu, and W. Chen. Provable adaptivity in adam. *arXiv preprint arXiv:2208.09900*, 2022.

B. Wang, H. Zhang, Z. Ma, and W. Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 161–190. PMLR, 2023.

C. Wang, X. Chen, A. J. Smola, and E. P. Xing. Variance reduction for stochastic gradient optimization. *Advances in neural information processing systems*, 26, 2013.

D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014. Publisher: Now Publishers, Inc.

P. Xu, J. Yang, F. Roosta, C. Ré, and M. W. Mahoney. Sub-sampled Newton methods with non-uniform sampling. *Advances in Neural Information Processing Systems*, 29, 2016.

P. Xu, F. Roosta, and M. W. Mahoney. Newton-type methods for non-convex optimization under inexact Hessian information. *Mathematical Programming*, 184(1-2):35–70, Nov. 2020. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-019-01405-z. URL http://link.springer.com/10.1007/s10107-019-01405-z.

B. Zhang, J. Jin, C. Fang, and L. Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020.

C. Zhang, D. Ge, B. Jiang, and Y. Ye. Drsom: A dimension reduced second-order method and preliminary analyses. *arXiv preprint arXiv:2208.00208*, 2022.

J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

L. Zhang, M. Mahdavi, and R. Jin. Linear convergence with condition number independent access of full gradients. *Advances in Neural Information Processing Systems*, 26, 2013.

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

Appendix

Appendix Structure The appendix sections will proceed as follows. Section A provides proof of the properties of the unified trust region framework. Section B provides proof of first-order algorithms, and Section E provides proof of second-order algorithms. Section C provides further discussion of second-order generalized smoothness. Section D provides applications on divergence-based penalized DRO and proves that it satisfies our second-order generalized smoothness. Section F provides proof for variance-reduced variants. Section G and Section H provides the guidelines about how to implement second-order methods efficiently and the details of the experiments respectively.

A Proof of the Unified Framework

When $L_0 = 0$, Assumption 1 simplifies to the L_1 smoothness employed in standard analysis. Similar to the analysis of L-smooth functions, we give the following descent inequality for (L_0, L_1)-smooth functions.

Lemma A.1. *Under Assumption 1, when $\|x - x'\| \le \frac{1}{L_1}$, we have*

$$F(x') \le F(x) + \nabla F(x)^T(x' - x) + \tfrac{1}{2}\|x' - x\|^2 \big(L_0 + L_1\|\nabla F(x)\|\big).$$

We omit the proof since it is well-known in the literature for (L_0, L_1) smoothness.

A.1 Proof of Lemma 1 (Model reduction)

Proof. According to the optimality condition (6), we have

$$\begin{aligned}
m_t(d_{t+1}) - m_t(0) &= g_t^T d_{t+1} + \frac{1}{2} d_{t+1}^T B_t d_{t+1} \\
&= -\lambda_t d_{t+1}^T d_{t+1} - \frac{1}{2} d_{t+1}^T B_t d_{t+1} \\
&= -\frac{1}{2}\lambda_t \|d_{t+1}\|^2 - \frac{1}{2} d_{t+1}^T \big(B_t + \lambda_t I\big) d_{t+1} \\
&\le -\frac{1}{2}\lambda_t \|d_{t+1}\|^2.
\end{aligned}$$

The last inequality is because $B_t + \lambda_t I \succeq 0$. □

B Proof of First-Order Methods

We first present a result on bounding the gradient variance.

Lemma B.1 (Variance bounds on gradient estimators)**.** *Suppose that Assumption 2 holds in Algorithm 1, then we have*

$$\mathbb{E}_t \|g_t - \nabla F(x_t)\|^2 \le \frac{G_0^2 + G_1^2\|\nabla F(x_t)\|^2}{|\mathcal{S}_1|},$$

where $\mathbb{E}_t$ denotes the expectation conditioned on all the randomness before the t-th iteration.

Proof.

$$\begin{aligned}
\mathbb{E}_t\big[\, \|g_t - \nabla F(x_t)\|^2 \,\big] &= \mathbb{E}_t\Big[\Big\| \frac{1}{|\mathcal{S}_1|} \sum_{\xi \in \mathcal{S}_1} \nabla f(x_t; \xi) - \nabla F(x_t) \Big\|^2\Big] \\
&= \frac{\sum_{\xi \in \mathcal{S}_1} \mathbb{E}_t\big[\|\nabla f(x_t;\xi) - \nabla F(x_t)\|^2\big]}{|\mathcal{S}_1|^2} \\
&\le \frac{G_0^2 + G_1^2\|\nabla F(x_t)\|^2}{|\mathcal{S}_1|}.
\end{aligned}$$

The last inequality is because Assumption 2. □

B.1 Proof of Theorem 1

Firstly, we give the following lemma.

Lemma B.2. *For λ_t satisfing the optimality condition (6) of the subproblem (5), we have*

$$\lambda_t \geq \frac{\|g_t\|}{\Delta} - \|B_t\|. \tag{8}$$

Proof. According to the optimality condition (6), we have $-g_t = \big(B_t + \lambda_t I\big)d_{t+1}$ and

$$\|g_t\| = \big\|\big(B_t + \lambda_t I\big)d_{t+1}\big\| \leq \|B_t + \lambda_t I\| \cdot \|d_{t+1}\| \leq \big(\|B_t\| + \lambda_t\big)\Delta.$$

Therefore, we have

$$\lambda_t \geq \frac{\|g_t\|}{\Delta} - \|B_t\|.$$

□

Then we prove Theorem 1 based on Lemma B.2.

Proof. According to Lemma A.1, we have

$$\begin{aligned}
F(x_{t+1}) \leq& F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{1}{2}\big(L_0 + L_1\|\nabla F(x_t)\|\big)\|x_{t+1} - x_t\|^2 \\
\overset{(\Diamond)}{=}& F(x_t) + \langle \nabla F(x_t) - g_t, d_{t+1} \rangle - \lambda_t \|d_{t+1}\|^2 - d_{t+1}^T B_t d_{t+1} + \frac{1}{2}\big(L_0 + L_1\|\nabla F(x_t)\|\big)\|d_{t+1}\|^2 \\
\overset{(\triangle)}{\leq}& F(x_t) + \|\nabla F(x_t) - g_t\|\Delta - \lambda_t \Delta^2 + \|B_t\|\Delta^2 + \frac{1}{2}\big(L_0 + L_1\|\nabla F(x_t)\|\big)\Delta^2 \\
\overset{(\vee)}{\leq}& F(x_t) + \|\nabla F(x_t) - g_t\|\Delta - \|g_t\|\Delta + 2\|B_t\|\Delta^2 + \frac{1}{2}\big(L_0 + L_1\|\nabla F(x_t)\|\big)\Delta^2 \\
\leq& F(x_t) + 2\|\nabla F(x_t) - g_t\|\Delta - \|\nabla F(x_t)\|\Delta + 2\|B_t\|\Delta^2 + \frac{1}{2}\big(L_0 + L_1\|\nabla F(x_t)\|\big)\Delta^2.
\end{aligned} \tag{9}$$

where in ($\Diamond$) we use the optimality condition (6), in ($\triangle$) we follow the fact that $\lambda_t\|d_{t+1}\|^2 = \lambda_t\Delta^2$ according to the observation that either $\lambda_t = 0$ or $\|d_{t+1}\| = \Delta$ because of the optimality condition, in ($\vee$) we use Lemma B.2. Therefore, we are able to bound $\mathbb{E}\|\nabla F(x_t)\|$ based on (9) and Lemma B.1

$$\begin{aligned}
\mathbb{E}\|\nabla F(x_t)\| \leq& \frac{1}{\Delta}\mathbb{E}\big[F(x_t) - F(x_{t+1})\big] + 2\mathbb{E}\|\nabla F(x_t) - g_t\| + 2\|B_t\|\Delta + \frac{1}{2}\big(L_0 + L_1\mathbb{E}\|\nabla F(x_t)\|\big)\Delta \\
\leq& \frac{1}{\Delta}\mathbb{E}\big[F(x_t) - F(x_{t+1})\big] + 2\frac{1}{\sqrt{|\mathcal{S}_1|}}\big(G_0 + G_1\mathbb{E}\|\nabla F(x_t)\|\big) + 2\beta\Delta + \frac{1}{2}\big(L_0 + L_1\mathbb{E}\|\nabla F(x_t)\|\big)\Delta.
\end{aligned} \tag{10}$$

Because $\Delta = \frac{1}{4L_0+16\beta}\epsilon$, $|\mathcal{S}_1| = \frac{64G_0^2}{\epsilon^2}$ and $\epsilon \leq \frac{4L_0G_0+16\beta G_0}{L_1G_0+2L_0G_1+8\beta G_1}$, we have $1 - \frac{L_1}{8L_0+32\beta}\epsilon - \frac{G_1}{4G_0}\epsilon \geq \frac{1}{2}$ and

$$\begin{aligned}
\left(1 - \frac{L_1}{8L_0+32\beta}\epsilon - \frac{G_1}{4G_0}\epsilon\right)\mathbb{E}\|\nabla F(x_t)\| \leq& \frac{4L_0+16\beta}{\epsilon}\mathbb{E}\big[F(x_t) - F(x_{t+1})\big] + \big(\frac{1}{4}\epsilon + \frac{1}{8}\epsilon\big) \\
\mathbb{E}\|\nabla F(x_t)\| \leq& \frac{8L_0+32\beta}{\epsilon}\mathbb{E}\big[F(x_t) - F(x_{t+1})\big] + \frac{3}{4}\epsilon.
\end{aligned} \tag{11}$$

By adding inequality (11) from 0 to $T-1$, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_t)\| \leq& \frac{8L_0+32\beta}{T\epsilon}\mathbb{E}\big[F(x_0) - F(x_T)\big] + \frac{3}{4}\epsilon \\
\leq& \frac{8L_0+32\beta}{T\epsilon}\Delta_F + \frac{3}{4}\epsilon.
\end{aligned} \tag{12}$$

When taking $T = \frac{32\Delta_F(L_0+4\beta)}{\epsilon^2}$, we have $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_t)\| \leq \epsilon$. Then we finish the proof. □

B.2 Proof of Corollary 1

According to the optimality condition (6), we have

$$\lambda_t d_{t+1} = -g_t, \lambda_t \|d_{t+1}\| = \lambda_t \cdot \Delta = \|g_t\|.$$

Therefore, we have

$$\lambda_t = \frac{\|g_t\|}{\Delta}, d_{t+1} = \frac{\Delta}{\|g_t\|} \cdot (-g_t).$$

Then we finish the proof for the equivalence between the normalized gradient descent and the first-order trust region method with $B_t = 0$.

B.3 Proof of Corollary 2

Similar to the proof of Corollary 1, according to the optimality condition (6), we have

$$d_{t+1} = \frac{1}{\rho + \lambda_t} \cdot (-g_t), \ \|d_{t+1}\| = \frac{1}{\rho + \lambda_t} \|g_t\|.$$

If $\rho < \frac{\|g_t\|}{\Delta}$, then we have $\lambda_t = \frac{\|g_t\|}{\|d_{t+1}\|} - \rho \geq \frac{\|g_t\|}{\Delta} - \rho > 0$, then we have

$$\|d_{t+1}\| = \Delta = \frac{1}{\rho + \lambda_t} \|g_t\|, \ \rho + \lambda_t = \frac{\|g_t\|}{\Delta},$$
$$d_{t+1} = \frac{\Delta}{\|g_t\|} \cdot (-g_t).$$

If $\rho \geq \frac{\|g_t\|}{\Delta}$, then we have $\|d_{t+1}\| = \frac{1}{\rho+\lambda_t} \|g_t\| \leq \frac{1}{\rho} \|g_t\| \leq \Delta$, then we have $\lambda_t = 0$ and

$$d_{t+1} = \frac{1}{\rho} \cdot (-g_t).$$

Putting the two cases together gives

$$d_{t+1} = \min \left\{ \frac{\|g_t\|}{\Delta}, \frac{1}{\rho} \right\} \cdot (-g_t).$$

Thus, we complete the proof.

C Further Discussion of Second-Order Generalized Smoothness

As mentioned before, we can interpret the second-order generalized smoothness from the perspective of the boundness of higher-order derivatives. According to (Nesterov et al., 2018), we first give the following definition.

Definition C.1. *For $F \in C^3(\mathbb{R}^n)$, define*

$$F'''(x)[u] = \lim_{\alpha \to 0} \frac{1}{\alpha} \left[\nabla^2 F(x + \alpha u) - \nabla^2 F(x) \right].$$

For any $F \in C^3(\mathbb{R}^n)$, the Lipschitz continuity of hessian is equivalent to the following condition

$$\|F'''(x)[u]\| \leq M \|u\|, \ \forall u \in \mathbb{R}^n.$$

For second-order generalized smoothness, we can give a similar equivalent condition

$$\|F'''(x)[u]\| \leq (M_0 + M_1 \|\nabla F(x)\|) \|u\|. \tag{13}$$

Next, we will show that condition (13) and Assumption 3 is equivalent in a sense. Let Assumption 3 hold, for any $u \in \mathbb{R}^n$, by the definition of $F'''(x)[u]$, we have

$$\begin{aligned}
\|F'''(x)[u]\| &= \| \lim_{\alpha \to 0} \frac{1}{\alpha} \left[\nabla^2 F(x + \alpha u) - \nabla^2 F(x) \right] \| \\
&= \lim_{\alpha \to 0} \frac{1}{\alpha} \|\nabla^2 F(x + \alpha u) - \nabla^2 F(x)\| \\
&\leq (M_0 + M_1 \|\nabla F(x)\|)\|u\|.
\end{aligned}$$

For the other direction, if $\|F'''(x)[u]\| \leq (M_0 + M_1\|\nabla F(x)\|)\|u\|$ holds for any $u \in \mathbb{R}^n$ and F is (L_0, L_1)-smooth, then Assumption 3 holds locally.

Lemma C.1. *Assume that F satisfies condition* (13) *and F is (L_0, L_1)-smooth. For any $c > 0$, if $\|x - y\| \leq \frac{2c}{\max\{L_0, L_1\}}$, then*

$$\|\nabla^2 F(x) - \nabla^2 F(y)\| \leq (A + B\|\nabla F(x)\|)\|x - y\|,$$

where $A = M_0 + cM_1$, $B = cM_1$.

Proof. Let $g(t)$ be defined as $g(t) = \nabla^2 F(x + t(y - x)), t \in [0, 1]$, $g'(t) = F'''(x + t(y - x))[y - x]$. Then we have

$$\begin{aligned}
\|\nabla^2 F(y) - \nabla^2 F(x)\| &= \|g(1) - g(0)\| \\
&= \left\| \int_0^1 F'''(\gamma(x + t(y - x)))[y - x] dt \right\| \\
&\leq \int_0^1 \|F'''(\gamma(x + t(y - x)))[y - x]\| dt \\
&\overset{(\triangle)}{\leq} \int_0^1 (M_0 + M_1\|\nabla F(x + t(y - x))\|)\|y - x\| dt \\
&\overset{(\diamondsuit)}{\leq} M_0\|y - x\| + M_1\|y - x\| \int_0^1 (L_0 + L_1\|\nabla F(x)\|)t\|y - x\| dt \\
&= M_0\|y - x\| + \frac{1}{2} M_1\|y - x\|(L_0 + L_1\|\nabla F(x)\|) \\
&\leq M_0\|y - x\| + \tfrac{c}{\max\{L_0, L_1\}} M_1\|y - x\|(L_0 + L_1\|\nabla F(x)\|) \\
&\leq (M_0 + cM_1)\|y - x\| + cM_1\|\nabla F(x)\|,
\end{aligned}$$

where $(\triangle)$ uses the condition (13), $(\diamondsuit)$ uses the (L_0, L_1) smoothness assumption. ☐

By taking $c = \min\{2, \frac{M_0}{M_1}\}$, we have $A \leq 2M_0, B \leq 2M_1$. Therefore, Assumption 3 holds with $2M_0$ and $2M_1$ locally.

D Further Discussion of DRO

Firstly, we make some standard assumptions for divergence-based DRO.

Assumption 5. *We give the following assumption for DRO settings:*

- *Given any ξ in its support set, the loss function $\ell(x, \xi)$ is twice differentiable, and it satisfies*

$$\begin{aligned}
\|\ell(x, \xi) - \ell(x', \xi)\| &\leq G\|x - x'\|, \\
\|\nabla \ell(x, \xi) - \nabla \ell(x', \xi)\| &\leq L\|x - x'\|, \\
\|\nabla^2 \ell(x, \xi) - \nabla^2 \ell(x', \xi)\| &\leq M\|x - x'\|.
\end{aligned}$$

- *The conjugate ψ^* is strictly convex, twice continuously differentiable and it satisfies*

$$|(\psi^*)'(x) - (\psi^*)'(x')| \leq N_1|x - x'|,$$
$$|(\psi^*)''(x) - (\psi^*)''(x')| \leq N_2|x - x'|.$$

- *For all $x \in \mathbb{R}^n$, the stochastic loss has a bounded variance, namely, $\mathbb{E}_{\xi\sim P}\left[\left(\ell(x,\xi) - \ell(x)\right)^2\right] \leq \sigma^2$ where $\ell(x) = \mathbb{E}_{\xi\sim P}[\ell(x,\xi)]$. Moreover, there exist constants m, M such that $m \leq \ell(x,\xi) \leq M$ for any x and $\xi \sim P$.*

- *There exists $\delta > 0$ such that* $\inf\{t|(\psi^*)'(t) \geq 1-\delta\} > -\infty$, $\sup\{t|(\psi^*)'(t) \leq 1+\delta\} < +\infty$

In many applications, the DRO problem may not exhibit Lipschitz smoothness due to its composition with a nonsmooth divergence function. A representative example illustrating this approach is provided in Jin et al. (2021, Example 3.1). To address this, the following subsection shows how we employ smooth approximation techniques to satisfy Assumption 5.

D.1 Smoothing approximation of valid divergence functions

Here we give some commonly used divergence functions:

Table 3: Some commonly used divergences and the corresponding conjugates.

Divergence	$\psi(t)$	$\psi^*(t)$
χ^2	$(t-1)^2$	$-1 + \frac{1}{4}(t+2)_+^2$
K-L	$t\log t - t + 1$	$e^t - 1$
CVaR	$\mathbb{I}_{[0,\alpha^{-1})}(t), \alpha \in (0,1)$	$\alpha^{-1}(t)_+$

In the case of CVaR, ψ^* is not differentiable as shown in Table 2, which is undesirable from an optimization viewpoint. Following Jin et al. (2021), we introduce a smoothed version of CVaR. The conjugate function of the smoothed CVaR is also smooth so that our results can be directly applied in smoothed CVAR.

For standard CVaR at level $\alpha, \psi_\alpha(t)$ takes zero when $t \in [0, 1/\alpha)$ and takes infinity otherwise. Instead, we consider the following smoothed version of CVaR:

$$\psi_\alpha^{\text{smo}}(t) = \begin{cases} t\log t + \frac{1-\alpha t}{\alpha}\log\frac{1-\alpha t}{1-\alpha} & t \in [0, 1/\alpha) \\ +\infty & \text{otherwise} \end{cases}$$

It is easy to see that ψ_α^{smo} is a valid divergence. The corresponding conjugate function is

$$\psi_\alpha^{\text{smo},*}(t) = \frac{1}{\alpha}\log(1 - \alpha + \alpha\exp(t))$$

In the case of χ^2-divergence, ψ^* is not second-order continuous, which also violates our Assumption 5. Therefore, we introduce a smoothed version of χ^2-divergence:

$$\psi_{\chi^2}^{\text{smo}}(t) = \begin{cases} (t-1)^2 & t \geq 1 \\ 2(t\log t - t + 1) & t \in [0,1) \\ +\infty & \text{otherwise} \end{cases}$$

It is easy to see that ψ_α^{smo} is a valid divergence. The corresponding conjugate function is

$$\psi_{\xi^2}^{\text{smo},*}(t) = \begin{cases} -1 + \frac{1}{4}(t+2)^2 & t \geq 0 \\ 2(\exp(\frac{t}{2}) - 1) & t < 0 \end{cases}$$

For a broader context on smooth approximation in nonconvex optimization, we refer the reader to (Nazareth, 1986; Chen, 1997, 2012)

D.2 Proof of Theorem 2

Note that the objective is $\Psi(x) = \min_{\eta\in\mathbb{R}} \mathcal{L}(x,\eta) := \lambda\mathbb{E}_{\xi\sim P}\psi^*\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right) + \eta$. Now we want to build the connections between the properties of $\Psi(x)$ and $\mathcal{L}(x,\eta)$, and then prove Theorem 3.

Lemma D.1. *Under the Assumption 5,* $\arg\min_{\eta'} \mathcal{L}(x,\eta')$ *is a singleton set.* $\Psi(x)$ *is twice-differentiable,* $\nabla\Psi(x) = \nabla_x\mathcal{L}(x,\eta_x^*)$, *and* $\nabla^2\Psi(x) = \nabla_x^2\mathcal{L}(x,\eta_x^*) + \frac{\lambda}{\mathbb{E}_\xi\left[(\psi^*)''(\frac{\ell(x,\xi)-\eta_x^*}{\lambda})\right]}\nabla_{x\eta}\mathcal{L}(x,\eta_x^*)\nabla_{x\eta}\mathcal{L}(x,\eta_x^*)^\top$ *for* $\eta_x^* = \arg\min_{\eta'} \mathcal{L}(x,\eta')$.

Proof. Firstly, since ψ^* is strictly convex, we have $\nabla_\eta^2\mathcal{L}(x,\eta) > 0$, which means the $\arg\min_{\eta'}\mathcal{L}(x,\eta')$ is a singleton set. Then from Lemma 2.6 in Jin et al. (2021), $\nabla\Psi(x) = \nabla_x\mathcal{L}(x,\eta_x^*)$. We first use the implicit function theorem to prove that η_x^* is actually a continuously differentiable function of x. Define $f(a,b) := \mathbb{E}_\xi\left[(\psi^*)'(\frac{l(a,\xi)-b}{\lambda})\right] - 1$, $a\in\mathbb{R}^n, b\in\mathbb{R}$. By the optimality condition, we have $\nabla_\eta\mathcal{L}(x,\eta_x^*) = 0$. In other words, we have

$$f(x,\eta_x^*) = 0.$$

Note that by the strict convexity of ψ^*, we have $\nabla_b f(x,\eta_x^*) = -\frac{1}{\lambda}\mathbb{E}_\xi\left[(\psi^*)''(\frac{\ell(x,\xi)-\eta_x^*}{\lambda})\right] \neq 0$. Hence, by the implicit function theorem, for any x, there exists an open set U_x containing x such that there exists a unique continuously differentiable function $g: U\to\mathbb{R}$ such that $g(x) = \eta_x^*$. Moreover, we have

$$\begin{aligned}
\nabla g(x) &= (\nabla_b f(x,\eta_x^*))^{-1}\cdot\nabla_a f(x,\eta_x^*)\\
&= -\frac{\mathbb{E}_\xi\left[\nabla\ell(x,\xi)(\psi^*)''(\frac{\ell(x,\xi)-\eta_x^*}{\lambda})\right]}{\mathbb{E}_\xi\left[(\psi^*)''(\frac{\ell(x,\xi)-\eta_x^*}{\lambda})\right]}\\
&= \frac{\lambda\nabla_{x\eta}\mathcal{L}(x,\eta_x^*)}{\mathbb{E}_\xi\left[(\psi^*)''(\frac{\ell(x,\xi)-\eta_x^*}{\lambda})\right]},
\end{aligned}$$

according to the chain rule, we have

$$\begin{aligned}
\nabla^2\Psi(x) &= \nabla_x^2\mathcal{L}(x,\eta_x^*) + \nabla_{x\eta}\mathcal{L}(x,\eta_x^*)\nabla g(x)^\top\\
&= \nabla_x^2\mathcal{L}(x,\eta_x^*) + \frac{\lambda}{\mathbb{E}_\xi\left[(\psi^*)''(\frac{\ell(x,\xi)-\eta_x^*}{\lambda})\right]}\nabla_{x\eta}\mathcal{L}(x,\eta_x^*)\nabla_{x\eta}\mathcal{L}(x,\eta_x^*)^\top\\
&\succeq \nabla_x^2\mathcal{L}(x,\eta_x^*).
\end{aligned}$$

$\square$

We first give some additional assumptions on the loss function ψ^*.

Assumption 6. *There exist constants m, M such that for any ϵ-FOSP (x,η), $m \le \ell(x,\xi) \le M$ a.e..*

Assumption 7. *There exists $\delta > 0$ such that* $\inf\{t|(\psi^*)'(t) \ge 1-\delta\} > -\infty$, $\sup\{t|(\psi^*)'(t) \le 1+\delta\} < +\infty$

Lemma D.2. *Given ϵ small enough, consider any (x,η) such that $\|\nabla\mathcal{L}(x,\eta)\| \le \epsilon$. Let $\underline{t} = \inf\{t|(\psi^*)'(t) \ge 1-\epsilon\}$, $\bar{t} = \sup\{t|(\psi^*)'(t) \le 1+\epsilon\}$. Under Assumptions 5, 6 and 7, we have $\eta \in [m-\lambda\bar{t}, M-\lambda\underline{t}]$.*

Proof. Because $\|\nabla\mathcal{L}(x,\eta)\| \le \epsilon$, we have

$$1-\epsilon \le \mathbb{E}_\xi\left[(\psi^*)'\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)\right] \le 1+\epsilon.$$

Let event $A = \{\xi|\ell(x,\xi) < m\}\cup\{\xi|\ell(x,\xi) > M\}$. According to Assumption 6, we have $P(A) = 0$. Then we have

$$1-\epsilon \le \mathbb{E}_\xi\left[(\psi^*)'\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)|A\right] = E_\xi\left[(\psi^*)'\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)\right] \le 1+\epsilon.$$

There exists $\xi_1, \xi_2 \in \overline{A}$ such that

$$(\psi^*)' \left(\frac{\ell(x;\xi_1) - \eta}{\lambda} \right) \geq 1 - \epsilon,$$
$$(\psi^*)' \left(\frac{\ell(x;\xi_2) - \eta}{\lambda} \right) \leq 1 + \epsilon.$$

Note that $(\psi^*)'$ is monotonically increasing because of the convexity of ψ^*, we have

$$\frac{\ell(x;\xi_1) - \eta}{\lambda} \geq \underline{t}, \quad \frac{\ell(x;\xi_2) - \eta}{\lambda} \leq \bar{t}.$$

Therefore, we can bound η by the following inequality

$$\begin{aligned} \eta &\leq \ell(x;\xi_1) - \lambda \underline{t} \leq M - \lambda \underline{t}, \\ \eta &\geq \ell(x;\xi_2) - \lambda \bar{t} \geq m - \lambda \bar{t}. \end{aligned}$$

□

Now we are ready to prove of Theorem 2.

Proof of Theorem 2. Suppose that we have obtained an $(\epsilon, \sqrt{\epsilon})$-SOSP pair (x, η) such that

$$\|\nabla \mathcal{L}(x,\eta)\| \leq c_1 \epsilon, \ \lambda_{\min}(\nabla^2 \mathcal{L}(x,\eta)) \geq -c_2 \sqrt{\epsilon}.$$

Let x be fixed and $\eta^* \in \arg\min_\eta \mathcal{L}(x,\eta)$. Then we have

$$\begin{aligned} &\|\nabla_x \mathcal{L}(x,\eta) - \nabla_x \mathcal{L}\left(x,\eta^*\right)\| \\ &= \left\| \mathbb{E}_\xi \left[\left((\psi^*)' \left(\frac{\ell(x;\xi) - \eta}{\lambda} \right) - (\psi^*)' \left(\frac{\ell(x;\xi) - \eta^*}{\lambda} \right) \right) \nabla \ell(x;\xi) \right] \right\| \\ &\leq G \cdot \mathbb{E}_\xi \left| (\psi^*)' \left(\frac{\ell(x;\xi) - \eta}{\lambda} \right) - (\psi^*)' \left(\frac{\ell(x;\xi) - \eta^*}{\lambda} \right) \right| \\ &= G \cdot \left| \mathbb{E}_\xi \left[(\psi^*)' \left(\frac{\ell(x;\xi) - \eta}{\lambda} \right) - (\psi^*)' \left(\frac{\ell(x;\xi) - \eta^*}{\lambda} \right) \right] \right| \\ &= G \left| \nabla_\eta \mathcal{L}(x,\eta) - \nabla_\eta \mathcal{L}\left(x,\eta^*\right) \right| = G \left| \nabla_\eta \mathcal{L}(x,\eta) \right|, \end{aligned} \tag{14}$$

where we use the fact that $(\psi^*)'$ is monotonely increasing (due to the convexity of ψ^*). Hence, using Lemma D.1 we obtain

$$\|\nabla \Psi(x)\| = \|\nabla_x \mathcal{L}\left(x,\eta^*\right)\| \leq \|\nabla_x \mathcal{L}(x,\eta)\| + G \left| \nabla_\eta \mathcal{L}(x,\eta) \right| \leq \sqrt{G^2 + 1} c_1 \epsilon \leq c_1' \epsilon.$$

Because of the boundedness of η given in lemma D.2, $(\psi^*)''$ is lower bounded by some positive constant and therefore ψ^* is strongly convex on this internal. So there exists $\mu > 0$ such that

$$\psi^*(y) - \psi^*(x) \geq (\psi^*)'(x) \cdot (y - x) + \frac{\mu}{2} |y - x|^2.$$

Therefore, we have

$$
\begin{aligned}
&\|\nabla_x^2\mathcal{L}\left(x,\eta_x^*\right)-\nabla_x^2\mathcal{L}\left(x,\eta\right)\| \\
&=\left\|\mathbb{E}_\xi\left[\left((\psi^*)''\left(\frac{\ell(x;\xi)-\eta_x^*}{\lambda}\right)-(\psi^*)''\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)\right)\frac{\nabla\ell(x;\xi)\nabla\ell(x;\xi)^T}{\lambda}\right]\right. \\
&\quad\left.+\mathbb{E}_\xi\left[\left((\psi^*)'\left(\frac{\ell(x;\xi)-\eta_x^*}{\lambda}\right)-(\psi^*)'\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)\right)\nabla^2\ell(x;\xi)\right]\right\| \\
&\le\frac{G^2}{\lambda}\mathbb{E}_\xi\left|\left((\psi^*)''\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)-(\psi^*)''\left(\frac{\ell(x;\xi)-\eta^*}{\lambda}\right)\right)\right|+L|\nabla_\eta\mathcal{L}(x,\eta)| \\
&\overset{(\Diamond)}{\le}\frac{G^2}{\lambda^2}N_2\mathbb{E}_\xi[|\eta-\eta_x^*|]+Lc_1\epsilon \\
&\overset{(\triangle)}{\le}\frac{G^2}{\lambda^2}N_2\mathbb{E}_\xi\left[\frac{\lambda}{\mu}\left|(\psi^*)'\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)-(\psi^*)'\left(\frac{\ell(x;\xi)-\eta^*}{\lambda}\right)\right|\right]+Lc_1\epsilon \\
&\overset{(\vee)}{\le}\frac{G^2}{\lambda\mu}N_2|\nabla_\eta\mathcal{L}(x,\eta)|+Lc_1\epsilon \\
&\le(\frac{G^2}{\lambda\mu}N_2+L)c_1\epsilon,
\end{aligned}
$$

where in ($\Diamond$) we use Assumption 5, in ($\triangle$) we use the property of strong convexity and in ($\vee$) we follow the derivation in (14). Hence, using Lemma D.1 we obtain

$$
\begin{aligned}
\lambda_{\min}(\nabla^2\Psi(x))&=\lambda_{\min}\left(\nabla_x^2\mathcal{L}(x,\eta_x^*)+\frac{\lambda}{\mathbb{E}_\xi\left[(\psi^*)''(\frac{\ell(x,\xi)-\eta_x^*}{\lambda})\right]}\nabla_{x\eta}\mathcal{L}(x,\eta_x^*)\nabla_{x\eta}\mathcal{L}(x,\eta_x^*)^\top\right) \\
&\ge\lambda_{\min}\left(\nabla_x^2\mathcal{L}(x,\eta_x^*)\right) \\
&\ge\lambda_{\min}(\nabla_x^2\mathcal{L}(x,\eta))-\|\nabla_x^2\mathcal{L}\left(x,\eta_x^*\right)-\nabla_x^2\mathcal{L}\left(x,\eta\right)\| \\
&\ge\lambda_{\min}(\nabla_x^2\mathcal{L}(x,\eta))-\left(\frac{G^2}{\lambda\mu}N_2+L\right)c_1\epsilon \\
&\ge-c_2'\sqrt{\epsilon}.
\end{aligned}
$$

□

D.3 Proof of Theorem 3

Finally, we want to prove that the penalized version of DRO satisfies our first-order generalized smoothness and second-order generalized smoothness.

Proof. From Lemma 3.3 and 3.4 of (Jin et al., 2021), Assumptions 1 and 2 naturally hold for $\mathcal{L}(x,\eta)$. Therefore, We only prove Assumption 3. Let $(\nabla\ell(x;\xi))^2$ denote $\nabla\ell(x,\xi)\nabla\ell(x,\xi)^T$. We know that

$$\nabla^2\mathcal{L}(x,\eta)=\begin{bmatrix}A_1 & A_2\\ A_3 & A_4,\end{bmatrix},$$

where $A_1=\mathbb{E}_\xi\left[\frac{1}{\lambda}(\psi^*)''\left(\frac{\ell(x,\xi)-\eta}{\lambda}\right)(\nabla\ell(x;\xi))^2+(\psi^*)'\left(\frac{\ell(x,\xi)-\eta}{\lambda}\right)\nabla^2\ell(x,\xi)\right]$, $A_4=\mathbb{E}_\xi\left[\frac{1}{\lambda}(\psi^*)''\left(\frac{\ell(x,\xi)-\eta}{\lambda}\right)\right]$,

$A_2 = A_3 = \mathbb{E}_\xi\left[-\frac{1}{\lambda}(\psi^*)''\left(\frac{\ell(x,\xi)-\eta}{\lambda}\right)\nabla\ell(x,\xi)\right]$. Hence, we have

$$\begin{aligned}
\|A_1(x,\eta) - A_1(x',\eta')\| \leq\, & \mathbb{E}\Big[\|\frac{1}{\lambda}(\psi^*)''\left(\frac{\ell(x,\xi)-\eta}{\lambda}\right)\left((\nabla\ell(x,\xi))^2 - (\nabla\ell(x',\xi))^2\right)\| \\
& + \|\frac{1}{\lambda}(\nabla\ell(x',\xi))^2[(\psi^*)''\left(\frac{\ell(x,\xi)-\eta}{\lambda}\right) - (\psi^*)''\left(\frac{\ell(x',\xi)-\eta'}{\lambda}\right)]\| \\
& + \|[(\psi^*)'\left(\frac{\ell(x,\xi)-\eta}{\lambda}\right) - (\psi^*)'\left(\frac{\ell(x',\xi)-\eta'}{\lambda}\right)]\nabla^2\ell(x,\xi)\| \\
& + \|(\psi^*)'\left(\frac{\ell(x',\xi)-\eta'}{\lambda}\right)(\nabla^2\ell(x,\xi) - \nabla^2\ell(x',\xi))\|\Big] \\
\leq\, & \mathbb{E}[\frac{N_1}{\lambda}\cdot 2GL\|x-x'\| + \frac{G^2N_2}{\lambda^2}\|\ell(x,\xi) - \ell(x',\xi) - (\eta-\eta')\| \\
& + \frac{1}{\lambda}LN_1\|\ell(x,\xi) - \ell(x',\xi) - (\eta-\eta')\| + M\cdot|(\psi^*)'\left(\frac{\ell(x,\xi)-\eta}{\lambda}\right)|\cdot\|x-x'\|] \\
\leq\, & \left(\frac{2}{\lambda}N_1GL + \frac{G^2N_2}{\lambda^2}\sqrt{G^2+1} + \frac{1}{\lambda}LN_1\sqrt{G^2+1} + M + M\|\nabla\mathcal{L}(x,\eta)\|\right)\|(x,\eta)-(x',\eta')\|,
\end{aligned}$$

where in the last inequality we use the fact that $\nabla_\eta\mathcal{L}(x,\eta) = \mathbb{E}_\xi\left[1 - (\psi^*)'\left(\frac{\ell(x,\xi)-\eta}{\lambda}\right)\right]$ and $|\nabla_\eta\mathcal{L}(x,\eta)| \leq \|\nabla\mathcal{L}(x,\eta)\|$. Similarly, we can get

$$\|A_2(x,\eta) - A_2(x',\eta')\| = \|A_3(x,\eta) - A_3(x',\eta)\| \leq \left(\frac{1}{\lambda}LN_1 + \frac{1}{\lambda^2}GN_2\sqrt{G^2+1}\right)\|(x,\eta)-(x',\eta')\|$$
$$\|A_4(x,\eta) - A_4(x',\eta')\| \leq \frac{1}{\lambda^2}N_2\sqrt{G^2+1}\|(x,\eta)-(x',\eta')\|.$$

Since $\|\nabla^2\mathcal{L}(x,\eta)\| \leq \|A_1\| + \|A_2\| + \|A_3\| + \|A_4\|$, we have

$$\|\nabla^2\mathcal{L}(x,\eta)\| \leq \left[\frac{1}{\lambda^2}N_2(G+1)^2\sqrt{G^2+1} + \frac{2}{\lambda}LN_1(G+1) + \frac{1}{\lambda}LN_1\sqrt{G^2+1} + M\big(1 + \|\nabla\mathcal{L}(x,\eta)\|\big)\right]\|(x,\eta)-(x',\eta')\|.$$

Hence, the first part of Assumption 3 is satisfied with $M_0 = \frac{1}{\lambda^2}N_2(G+1)^2\sqrt{G^2+1} + \frac{2}{\lambda}LN_1(G+1) + \frac{1}{\lambda}LN_1\sqrt{G^2+1} + M$, $M_1 = M$.

For the second part of Assumption 3, let A be a random matrix. Then Jensen's inequality and convexity imply that

$$\mathbb{V}[A] = \mathbb{E}[\|A - \mathbb{E}[A]\|^2] \leq \mathbb{E}_{A,A'}[\|A - A^{'}\|^2] = 2\mathbb{V}[A],$$

where A' is an i.i.d. copy of A. Denote $A_i(\xi_j) = A_i(x, \eta, \xi_j)$. It follows that

$$
\begin{aligned}
\mathbb{V}[A_1(\xi)] &\leq \mathbb{E}[\|A_1(\xi_1) - A_1(\xi_2)\|^2] \\
&\leq 4\mathbb{E}\Big[\Big\|\frac{1}{\lambda}(\psi^*)''\left(\frac{\ell(x,\xi_1)-\eta}{\lambda}\right)[(\nabla\ell(x,\xi_1))^2 - (\nabla\ell(x,\xi_2))^2]\Big\|^2 \\
&+ 4\Big\|\frac{1}{\lambda}(\nabla\ell(x,\xi_2))^2\left[(\psi^*)''\left(\frac{\ell(x,\xi_1)-\eta}{\lambda}\right) - (\psi^*)''\left(\frac{\ell(x,\xi_2)-\eta}{\lambda}\right)\right]\Big\|^2 \\
&+ 4\Big\|[(\psi^*)'\left(\frac{\ell(x,\xi_1)-\eta}{\lambda}\right) - (\psi^*)'\left(\frac{\ell(x,\xi_2)-\eta}{\lambda}\right)]\nabla^2\ell(x,\xi_1)\Big\|^2 \\
&+ 4\Big\|(\psi^*)'\left(\frac{\ell(x,\xi_2)-\eta}{\lambda}\right)(\nabla^2\ell(x,\xi_1) - \nabla^2\ell(x,\xi_2))\Big\|^2\Big] \\
&\leq \frac{16}{\lambda^2}G^4N_1^2 + \frac{4}{\lambda^4}G^4N_2^2\mathbb{E}[(\ell(x,\xi_1)-\ell(x,\xi_2))^2] \\
&+ \frac{4}{\lambda^2}L^2N_1\mathbb{E}[\ell(x,\xi_1)-\ell(x,\xi_2)^2] + 16L^2\mathbb{E}\left[\left\|(\psi^*)'\left(\frac{\ell(x,\xi_1)-\eta}{\lambda}\right)\right\|^2\right] \\
&\leq \frac{16}{\lambda^2}G^4N_1^2 + \frac{8}{\lambda^4}G^4N_2^2\sigma^2 + \frac{8}{\lambda^2}L^2N_1\sigma^2 + 16L^2\mathbb{E}\left[\left\|(\psi^*)'\left(\frac{\ell(x,\xi_1)-\eta}{\lambda}\right)\right\|^2\right]
\end{aligned}
$$

Now, we deal with the first term. Using $2(a-1)^2 + 2 \geq a^2$ for any a, we have

$$
\begin{aligned}
\mathbb{E}_\xi\left[\left((\psi^*)'\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)\right)^2\right] &\leq 2 + 2\mathbb{E}_\xi\left[\left(1-(\psi^*)'\left(\frac{\ell(x;\xi)-\eta}{\lambda}\right)\right)^2\right] \\
&= 2\left(1 + \|\nabla_\eta\mathcal{L}(x,\eta)\|^2 + \mathbb{V}\left[\nabla_\eta\mathcal{L}(x,\eta;\xi)\right]\right).
\end{aligned}
$$

For $\mathbb{V}[\nabla_\eta\mathcal{L}(x,\eta;\xi)]$, we have

$$
\begin{aligned}
\mathbb{V}[\nabla_\eta\mathcal{L}(x,\eta;\xi)] &\leq \mathbb{E}[\|\nabla_\eta\mathcal{L}(x,\eta;\xi_1) - \nabla_\eta\mathcal{L}(x,\eta;\xi_2)\|^2] \\
&= \mathbb{E}\left[\left|(\psi^*)'\left(\frac{\ell(x;\xi_1)-\eta}{\lambda}\right) - (\psi^*)'\left(\frac{\ell(x;\xi_2)-\eta}{\lambda}\right)\right|^2\right] \\
&\leq \frac{N_1^2}{\lambda^2}\mathbb{E}\|\ell(x;\xi_1)-\ell(x;\xi_2)\|^2 \\
&\leq \frac{2}{\lambda^2}N_1^2\sigma^2.
\end{aligned}
$$

Therefore, $\mathbb{V}[A_1(\xi)]$ can be bounded by constants and $\|\nabla\mathcal{L}(x,\eta)\|$. It follows that

$$
\begin{aligned}
\mathbb{V}[A_1(\xi)] &\leq \frac{16}{\lambda^2}G^4N_1^2 + \frac{8}{\lambda^4}G^4N_2^2\sigma^2 + \frac{8}{\lambda^2}L^2N_1\sigma^2 + 16L^2\mathbb{E}\left[\left\|(\psi^*)'\left(\frac{\ell(x,\xi_1)-\eta}{\lambda}\right)\right\|^2\right] \\
&\leq \frac{16}{\lambda^2}G^4N_1^2 + \frac{8}{\lambda^4}G^4N_2^2\sigma^2 + \frac{8}{\lambda^2}L^2N_1\sigma^2 + 32L^2\big(1 + \|\nabla_\eta\mathcal{L}(x,\eta)\|^2 + \mathbb{V}\left[\nabla_\eta\mathcal{L}(x,\eta;\xi)\right]\big) \\
&\leq \frac{16}{\lambda^2}G^4N_1^2 + \frac{8}{\lambda^4}G^4N_2^2\sigma^2 + \frac{8}{\lambda^2}L^2N_1\sigma^2 + 32L^2 + \frac{64}{\lambda^2}L^2N_1^2\sigma^2 + 32L^2\left\|\nabla\mathcal{L}(x,\eta)\right\|^2.
\end{aligned}
$$

Using a similar argument, we can show

$$
\begin{aligned}
\mathbb{V}[A_2(\xi)] = \mathbb{V}[A_3(\xi)] &\leq \frac{8}{\lambda^2}G^2N_1^2 + \frac{4}{\lambda^4}G^2N_2^2\sigma^2, \\
\mathbb{V}[A_4(\xi)] &\leq \frac{2}{\lambda^4}N_2^2\sigma^2.
\end{aligned}
$$

Furthermore, according to Cauchy-Schwarz inequality, we have

$$\mathbb{V}[\nabla^2 \mathcal{L}(x,\eta)\| \leq \mathbb{E}[(\|A_1 - \mathbb{E}[A_1]\| + \|A_2 - \mathbb{E}[A_2]\| + \|A_3 - \mathbb{E}[A_3]\| + \|A_4 - \mathbb{E}[A_4]\|)^2] \leq 4\sum_{i=1}^{4} \mathbb{V}[A_i(\xi)].$$

Combining the above properties, we obtain the desired conclusions. □

E Proof of Second-Order Methods

Analogous to standard results of second-order Lipschitz smooth functions (Nesterov, 2018), we develop some useful properties for the second-order generalized smoothness.

Lemma E.1. *Under Assumption 3, we have*

$$\|\nabla F(x') - \nabla F(x) - \nabla^2 F(x)(x'-x)\| \leq \frac{M_0 + M_1\|\nabla F(x)\|}{2}\|x'-x\|^2.$$

Lemma E.2. *Under Assumption 3, we have*

$$F(x') - F(x) \leq \nabla F(x)^T(x'-x) + \frac{1}{2}(x'-x)^T\nabla^2 F(x)(x'-x) + \frac{1}{6}\big(M_0 + M_1\|\nabla F(x)\|\big)\|x'-x\|^3.$$

E.1 Proof of Lemma E.1 and Lemma E.2

Proof. We have

$$\begin{aligned}
y^T\big(\nabla F(x') - \nabla F(x) - \nabla^2 F(x)(x'-x)\big) =& y^T\int_0^1 \big(\nabla^2 F(x + k(x'-x)) - \nabla^2 F(x)\big)(x'-x)\mathrm{d}k \\
\leq& \|y\|\int_0^1 \|\nabla^2 F(x+k(x'-x)) - \nabla^2 F(x)\|\|x'-x\|\mathrm{d}k \\
\leq& \|y\|\frac{M_0 + M_1\|\nabla F(x)\|}{2}\|x'-x\|^2
\end{aligned}$$

and the result follows by Cauchy-Schwarz inequality.

We prove the second equation analogously

$$\begin{aligned}
&F(x') - F(x) \\
&= \int_0^1 \nabla F(x + t(x'-x))^T(x'-x)\mathrm{d}t \\
&= \nabla F(x)^T(x'-x) + \int_0^1 \big(\nabla F(x+t(x'-x))^T(x'-x) - \nabla F(x)^T(x'-x)\big)\mathrm{d}t \\
&= \nabla F(x)^T(x'-x) + \int_0^1\int_0^1 t(x'-x)^T\nabla^2 F(x+tk(x'-x))^T(x'-x)dkdt \\
&= \nabla F(x)^T(x'-x) + \frac{1}{2}(x'-x)^T\nabla^2 F(x)(x'-x) \\
&\quad + \int_0^1\int_0^1 t(x'-x)^T\big(\nabla^2 F(x+tk(x'-x)) - \nabla^2 F(x)\big)^T(x'-x)\mathrm{d}k\mathrm{d}t \\
&\leq \nabla F(x)^T(x'-x) + \frac{1}{2}(x'-x)^T\nabla^2 F(x)(x'-x) + \int_0^1\int_0^1 t\|x'-x\|^2\|\nabla^2 F(x+tk(x'-x)) - \nabla^2 F(x)\|\mathrm{d}k\mathrm{d}t \\
&\leq \nabla F(x)^T(x'-x) + \frac{1}{2}(x'-x)^T\nabla^2 F(x)(x'-x) + \int_0^1\int_0^1 t^2k\|x'-x\|^3\big(M_0 + M_1\|\nabla F(x)\|\big)\mathrm{d}k\mathrm{d}t \\
&= \nabla F(x)^T(x'-x) + \frac{1}{2}(x'-x)^T\nabla^2 F(x)(x'-x) + \frac{1}{6}\big(M_0 + M_1\|\nabla F(x)\|\big)\|x'-x\|^3.
\end{aligned}$$

□

E.2 Proof of Lemma 2

Proof. As for the variance bound of Hessian, it is a direct result of the following auxiliary lemma (whose proof can be found in Arjevani et al. (2020). We derive it here just for completeness) by setting $A_i = \nabla^2 f(x_t;\xi_i), B = \nabla^2 F(x_t), \sigma^2 = K_0^2 + K_1^2\|\nabla F(x_t)\|^2, m = |\mathcal{S}_2|$.

Lemma E.3. *Let $(A_i)_{i=1}^m$ be a collection of i.i.d. sampled matrices in $\mathbb{S}^n$, with $\mathbb{E}[A_i] = B$ and $\mathbb{E}\|A_i - B\|^2 \leq \sigma^2$. Then it holds that*

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^m A_i - B\right\|^2 \leq \frac{22\sigma^2\log n}{m}.$$

Proof: We drop the normalization by m throughout this proof. We first symmetrize. Observe that by Jensen's inequality we have

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^m A_i - mB\right\|^2 &\leq \mathbb{E}_A\mathbb{E}_{A'}\left\|\sum_{i=1}^m A_i - A_i'\right\|^2 \\
&= \mathbb{E}_A\mathbb{E}_{A'}\left\|\sum_{i=1}^m (A_i - B) - (A_i' - B)\right\|^2 \\
&= \mathbb{E}_A\mathbb{E}_A\mathbb{E}_\epsilon\left\|\sum_{i=1}^m \epsilon_i\left((A_i - B) - (A_i' - B)\right)\right\|^2 \leq 4\mathbb{E}_A\mathbb{E}_\epsilon\left\|\sum_{i=1}^m \epsilon_i\left(A_i - B\right)\right\|^2,
\end{aligned}$$

where $(A')_{i=1}^m$ is a sequence of independent copies of $(A_i)_{i=1}^m$ and $(\epsilon_i)_{i=1}^m$ are Rademacher random variables. Henceforth we condition on A. Let $p = \log n$, and let $\|\cdot\|_{S_p}$ denote the Schatten p-norm. In what follows, we will use that for any matrix X, $\|X\| \leq \|X\|_{S_{2p}} \leq e^{1/2}\|X\|$. To begin, we have

$$\mathbb{E}_\epsilon\left\|\sum_{i=1}^m \epsilon_i\left(A_i - B\right)\right\|^2 \leq \mathbb{E}_\epsilon\left\|\sum_{i=1}^m \epsilon_i\left(A_i - B\right)\right\|_{S_{2p}}^2 \leq \left(\mathbb{E}_\epsilon\left\|\sum_{i=1}^m \epsilon_i\left(A_i - B\right)\right\|_{S_{2p}}^{2p}\right)^{1/p},$$

where the second inequality follows from Jensen's inequality. We apply the matrix Khintchine inequality ((Mackey et al., 2014), Corollary 7.4), which implies that

$$\begin{aligned}
\left(\mathbb{E}_\epsilon\left\|\sum_{i=1}^m \epsilon_i\left(A_i - B\right)\right\|_{S_{2p}}^{2p}\right)^{1/p} \leq (2p-1)\left\|\sum_{i=1}^m (A_i - B)^2\right\|_{S_{2p}} &\leq (2p-1)\sum_{i=1}^m \|(A_i - B)\|_{S_{2p}}^2 \\
&\leq e(2p-1)\sum_{i=1}^m \|(A_i - B)\|^2.
\end{aligned}$$

Putting all the developments so far together and taking expectations with respect to A, we have

$$\mathbb{E}\left\|\sum_{i=1}^m A_i - B\right\|^2 \leq 4e(2p-1)\sum_{i=1}^m \mathbb{E}_{A_i}\|(A_i - B)\|^2 \leq 4e(2p-1)m\sigma^2.$$

To obtain the final result, we divide both ends of the above inequality by m^2. $\square$

E.3 Proof of Theorem 4

Proof. According to Lemma E.2, we have

$$
\begin{aligned}
&F(x_{t+1}) - F(x_t) \\
&\le \nabla F(x_t)^T d_{t+1} + \frac{1}{2} d_{t+1}^T \nabla^2 F(x_t) d_{t+1} + \frac{M_0 + M_1\|\nabla F(x_t)\|}{6}\|d_{t+1}\|^3 \\
&= g_t^T d_{t+1} + \frac{1}{2}(d_{t+1})^T H_t d_{t+1} + (\nabla F(x_t) - g_t)^T d_{t+1} + \frac{1}{2}(d_{t+1})^T(\nabla^2 F(x_t) - H_t) d_{t+1} \\
&\quad + \frac{M_0 + M_1\|\nabla F(x_t)\|}{6}\|d_{t+1}\|^3 \\
&\overset{(\diamondsuit)}{\le} -\frac{1}{2}\lambda_t \|d_{t+1}\|^2 + \|\nabla F(x_t) - g_t\|\Delta + \frac{1}{2}\|\nabla^2 F(x_t) - H_t\|\Delta^2 + \frac{M_0 + M_1\|\nabla F(x_t)\|}{6}\Delta^3,
\end{aligned}
\tag{15}
$$

where $(\diamondsuit)$ is due to Cauchy-Schwarz inequality and Lemma 1. At t-th iteration, if $\|d_{t+1}\| = \Delta$, then we can bound $\|\lambda_t d_{t+1}\|$ by

$$
\|\lambda_t d_{t+1}\| \le \frac{2}{\Delta}(F(x_t) - F(x_{t+1})) + 2\|\nabla F(x_t) - g_t\| + \|\nabla^2 F(x_t) - H_t\|\Delta + \frac{M_0 + M_1\|\nabla F(x_t)\|}{3}\Delta^2. \tag{16}
$$

Otherwise, if $d_{t+1} < \Delta$, then by the optimal condition of (6) we have $\lambda_t = 0$. Hence, the upper bound (16) still holds.

Considering $\|\nabla F(x_{t+1})\|$, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla F(x_{t+1})\| \\
&\le \mathbb{E}\left[\|\nabla F(x_{t+1}) - \nabla F(x_t) - \nabla^2 F(x_t) d_{t+1}\| + \|\nabla F(x_t) - g_t\| + \|(\nabla^2 F(x_t) - H_t) d_{t+1}\| + \|g_t + H_k d_{t+1}\|\right] \\
&\overset{(\natural)}{\le} \mathbb{E}\left[\frac{M_0 + M_1\|\nabla F(x_t)\|}{2}\Delta^2 + \|\nabla F(x_t) - g_t\| + \Delta\|\nabla^2 F(x_t) - H_t\| + \|\lambda_t d_{t+1}\|\right] \\
&\overset{(\triangle)}{\le} \mathbb{E}\left[\frac{5M_0 + 5M_1\|\nabla F(x_t)\|}{6}\Delta^2 + 3\|\nabla F(x_t) - g_t\| + 2\Delta\|\nabla^2 F(x_t) - H_t\| + \frac{2}{\Delta}(F(x_t) - F(x_{t+1}))\right],
\end{aligned}
\tag{17}
$$

where $(\natural)$ is due to Lemma E.2 and the optimality condition (6), and $(\triangle)$ is due to (16).

By setting $\Delta = \sqrt{\epsilon}$, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla F(x_{t+1})\| \\
&\le \mathbb{E}\Big[\frac{5M_0 + 5M_1\|\nabla F(x_t)\|}{6}\epsilon + 3(G_0 + G_1\|\nabla F(x_t)\|)\epsilon + 2\epsilon(K_0 + K_1\|\nabla F(x_t)\|) + \frac{2}{\sqrt{\epsilon}}\big(F(x_t) - F(x_{t+1})\big)\Big] \\
&\le \left(\frac{5M_0}{6} + 3G_0 + 2K_0\right)\epsilon + \left(\frac{5M_1\epsilon}{6} + 3G_1\epsilon + 2K_1\epsilon\right)\mathbb{E}\left[\|\nabla F(x_t)\|\right] + \mathbb{E}\left[\frac{2}{\sqrt{\epsilon}}\big(F(x_t) - F(x_{t+1})\big)\right] \\
&\le \left(\frac{5M_0}{6} + 3G_0 + 2K_0\right)\epsilon + \left(\frac{5M_1\epsilon}{6} + 3G_1\epsilon + 2K_1\epsilon\right)\mathbb{E}\left[\|\nabla F(x_{t+1})\| + \|\nabla F(x_{t+1}) - F(x_t)\|\right] \\
&\quad + \mathbb{E}\left[\frac{2}{\sqrt{\epsilon}}\big(F(x_t) - F(x_{t+1})\big)\right] \\
&\le \left(\frac{5M_0}{6} + 3G_0 + 2K_0\right)\epsilon + \left(\frac{5M_1\epsilon}{6} + 3G_1\epsilon + 2K_1\epsilon\right)\mathbb{E}[\|\nabla F(x_{t+1})\|] \\
&\quad + \mathbb{E}\left[\left(\frac{5M_1\epsilon}{6} + 3G_1\epsilon + 2K_1\epsilon\right)(L_0 + L1\|\nabla F(x_{t+1})\|)\sqrt{\epsilon}\right] + \mathbb{E}\left[\frac{2}{\sqrt{\epsilon}}\big(F(x_t) - F(x_{t+1})\big)\right].
\end{aligned}
$$

Therefore, we have

$$\begin{aligned}
&\left(1 - \frac{5M_1}{6}\epsilon - 3G_1\epsilon - 2K_1\epsilon - \mathcal{O}(\epsilon)\right)\mathbb{E}[\|\nabla F(x_{t+1})\|] \\
&\leq \left(\frac{5M_0}{6} + 3G_0 + 2K_0\right)\epsilon + \left(\frac{5M_1\epsilon}{6} + 3G_1\epsilon + 2K_1\epsilon\right)L_0\sqrt{\epsilon} \\
&\quad + \mathbb{E}\left[\frac{2}{\sqrt{\epsilon}}(F(x_t) - F(x_{t+1})\right] \\
&\leq \mathcal{O}(\epsilon) + \mathbb{E}\left[\frac{2}{\sqrt{\epsilon}}(F(x_t) - F(x_{t+1})\right].
\end{aligned}$$

Because $\epsilon < \frac{3}{5M_1+18G_1+12K_1}$ and $T = \mathcal{O}(\epsilon^{-3/2})$, we have

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla F(x_{t+1})\|\right] \leq \mathcal{O}(\epsilon) + \frac{4\Delta_F}{T\sqrt{\epsilon}} = \mathcal{O}(\epsilon). \tag{18}$$

For the second-order condition,

$$\begin{aligned}
&\mathbb{E}[F(x_{t+1}) - F(x_t)] \\
\leq&\mathbb{E}\left[-\frac{1}{2}\lambda_t\|d_{t+1}\|^2\right] + \mathbb{E}\left[\|\nabla F(x_t) - g_t\|\Delta_t + \frac{1}{2}\|\nabla^2 F(x_t) - H_t\|\Delta_t^2 + \frac{M_0 + M_1\|\nabla F(x_t)\|}{6}\|d_{t+1}\|^3\right] \\
\leq&\mathbb{E}[\lambda_t]\cdot(-\frac{1}{2}\epsilon) + \left(G_0 + \frac{1}{2}K_0 + \frac{M_0}{6}\right)\epsilon^{3/2} + \left(G_1 + \frac{1}{2}K_1 + \frac{M_1}{6}\right)\|\nabla F(x_t)\|\epsilon^{3/2}
\end{aligned} \tag{19}$$

Summing from $t = 0$ to $T-1$, and plugging T into the equation, we have

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}(F(x_t) - F(x_{t+1}))\right] \leq \frac{1}{T}\cdot(F(x_0) - F^*) \leq \mathcal{O}(\epsilon^{\frac{3}{2}}). \tag{20}$$

Combining (18)-(20) , we have

$$\begin{aligned}
\mathbb{E}\left[\lambda_{\bar{t}}\right] &\leq \frac{2}{\epsilon}\big[\mathcal{O}(\epsilon^{\frac{3}{2}}) + \left(G_0 + \frac{1}{2}K_0 + \frac{M_0}{6}\right)\epsilon^{3/2} + \left(G_1 + \frac{1}{2}K_1 + \frac{M_1}{6}\right)\mathbb{E}\left[\frac{\sum_{t=0}^{T-1}\|\nabla F(x_t)\|}{T}\right]\epsilon^{3/2}\big] \\
&\leq \mathcal{O}(\sqrt{\epsilon}) + 2\left(G_1 + \frac{1}{2}K_1 + \frac{M_1}{6}\right)\mathbb{E}\left[\frac{\sum_{t=0}^{T-1}\|\nabla F(x_{t+1}) - \nabla F(x_t)\|}{T}]\epsilon^{3/2}\right] \\
&\leq \mathcal{O}(\sqrt{\epsilon}) + 2\left(G_1 + \frac{1}{2}K_1 + \frac{M_1}{6}\right)\mathbb{E}\left[\frac{\sum_{t=0}^{T-1}(L_0 + L_1\|\nabla F(x_{t+1})\|)}{T}\epsilon^{3/2}\right] \\
&\leq \mathcal{O}(\sqrt{\epsilon}),
\end{aligned} \tag{21}$$

where $\bar{t}$ is uniformly sampled from $0, \ldots, T-1$. By the optimality condition (6), we have

$$\begin{aligned}
-\mathbb{E}[\lambda_{\bar{t}}]I \preceq \mathbb{E}[H_{\bar{t}}] &\preceq \mathbb{E}[\nabla^2 F(x_{\bar{t}+1})] + \mathbb{E}[\|H_{\bar{t}} - \nabla^2 F(x_{\bar{t}+1})\|I] \\
&\preceq \mathbb{E}[\nabla^2 F(x_{\bar{t}+1})] + \mathbb{E}[\|H_{\bar{t}} - \nabla^2 F(x_{\bar{t}})\|I] + \mathbb{E}[\|\nabla^2 F(x_{\bar{t}}) - \nabla^2 F(x_{\bar{t}+1})\|I] \\
&\overset{(\Diamond)}{\preceq} \mathbb{E}[\nabla^2 F(x_{\bar{t}+1})] + \mathcal{O}(\sqrt{\epsilon})I + \mathcal{O}(\sqrt{\epsilon})\mathbb{E}[\|\nabla F(x_{\bar{t}})\|I] \\
&\preceq \mathbb{E}[\nabla^2 F(x_{\bar{t}+1})] + \mathcal{O}(\sqrt{\epsilon}),
\end{aligned}$$

where $(\Diamond)$ is due to Assumption 3 and Lemma 2. Hence we conclude

$$\mathbb{E}[\lambda_{\min}(\nabla^2 F(x_{\bar{t}+1}))] \geq -\mathcal{O}(\sqrt{\epsilon}).$$

□

F Variance Reduction

Note that the smoothness condition in Assumption 1 is solely imposed on the main objective F irrespective of its components. As it is the standard assumption in variance-reduced optimization (Fang et al., 2018), we impose the following averaged smoothness condition of F and its components.

Assumption 8. *In the stochastic setting, it holds that*

$$\mathbb{E}\|\nabla f(x,\xi) - \nabla f(x',\xi)\| \leq (L_0 + L_1\|\nabla F(x)\|)\|x - x'\|.$$

We first give an important lemma, which helps to bound the variance of the SPIDER gradient estimator.

Lemma F.1. *Let $\hat{\delta}_t = \nabla f(x_t;\mathcal{S}_3) - \nabla f(x_{t-1};\mathcal{S}_3) - \big(\nabla F(x_t) - \nabla F(x_{t-1})\big)$. Given q, $\Delta_t = \Delta$, we have*

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t}\hat{\delta}_\tau\right\|\right] \leq \sqrt{2}L_1\frac{\Delta}{\sqrt{|\mathcal{S}_3|}}\sum_{\tau=1}^{t}\mathbb{E}\big[\|\nabla F(x_\tau)\|\big] + \sqrt{2}L_0\frac{\Delta}{\sqrt{|\mathcal{S}_3|}}\sqrt{q}$$

.

F.1 Proof of Theorem 5

Proof. Consider t such that $0 < t < q$. Let $\delta_t = g_t - \nabla F(x_t)$, $\hat{\delta}_t = \nabla f(x_t;\mathcal{S}_3) - \nabla f(x_{t-1};\mathcal{S}_3) - \big(\nabla F(x_t) - \nabla F(x_{t-1})\big)$, then we have

$$\begin{aligned}\delta_t &= g_t - \nabla F(x_t)\\ &= g_{t-1} - \nabla F(x_{t-1}) + \nabla f(x_t;\mathcal{S}_3) - \nabla f(x_{t-1};\mathcal{S}_3) - \big(\nabla F(x_t) - \nabla F(x_{t-1})\big)\\ &= \delta_{t-1} + \hat{\delta}_t\\ &= \delta_0 + \hat{\delta}_1 + \cdots + \hat{\delta}_t.\end{aligned} \tag{22}$$

Then we can bound $\|\delta_t\|$ by the following inequality

$$\|\delta_t\| \leq \|\delta_0\| + \|\hat{\delta}_1 + \cdots + \hat{\delta}_t\| \tag{23}$$

Let $\{\xi_j^{(\tau)}\}$ be the random samples used in the τ-th iteration. Let $\mathcal{F}_t = \sigma\{x_1, ..., x_t\}$. We use $\mathbb{E}_{\mathcal{F}_t}$ to denote the expectation on $\mathcal{F}_t$ and $\mathbb{E}_t$ to denote the conditional expectation on $\mathcal{F}_t$. Note that x_t is $\mathcal{F}_t$-measurable. We first bound $\|\hat{\delta}_t\|$

$$\begin{aligned}\mathbb{E}_t\left[\|\hat{\delta}_t\|^2\right] =& \mathbb{E}_t\left[\left\|\frac{1}{|\mathcal{S}_3|}\sum_{j=1}^{|\mathcal{S}_3|}\big(\nabla f(x_t;\xi_j) - \nabla f(x_{t-1};\xi_j)\big) - \big(\nabla F(x_t) - \nabla F(x_{t-1})\big)\right\|^2\right]\\ =& \frac{1}{|\mathcal{S}_3|^2}\sum_{j=1}^{|\mathcal{S}_3|}\mathbb{E}_t\big[\|\nabla f(x_t;\xi_j) - \nabla f(x_{t-1};\xi_j)\|^2 + \|\nabla F(x_t) - \nabla F(x_{t-1})\|^2\\ &- 2\big(\nabla f(x_t;\xi_j) - \nabla f(x_{t-1};\xi_j)\big)^T\big(\nabla F(x_t) - \nabla F(x_{t-1})\big)\big]\\ \overset{\diamondsuit}{=}& \frac{1}{|\mathcal{S}_3|^2}\sum_{j=1}^{|\mathcal{S}_3|}\mathbb{E}_t\big[\|\nabla f(x_t;\xi_j) - \nabla f(x_{t-1};\xi_j)\|^2 - \|\nabla F(x_t) - \nabla F(x_{t-1})\|^2\big]\\ \leq& \frac{1}{|\mathcal{S}_3|^2}\sum_{j=1}^{|\mathcal{S}_3|}\mathbb{E}_t\big[\|\nabla f(x_t;\xi_j) - \nabla f(x_{t-1};\xi_j)\|^2\big]\\ \overset{\triangle}{\leq}& \frac{1}{|\mathcal{S}_3|^2}\sum_{j=1}^{|\mathcal{S}_3|}\big(L_0 + L_1\|\nabla F(x_t)\|\big)^2\|d_t\|^2\\ \leq& 2\frac{\Delta^2}{|\mathcal{S}_3|}\big(L_0^2 + L_1^2\|\nabla F(x_t)\|^2\big),\end{aligned}$$

where $\Diamond$ is due to $\mathbb{E}\big[\nabla f(x_t;\xi_j) - \nabla f(x_{t-1};\xi_j)\big] = \nabla F(x_t) - \nabla F(x_{t-1})$ and $\triangle$ is due to Assumption 8.

Applying Lemma F.1, we have

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t} \hat{\delta}_\tau\right\|\right] \leq \sqrt{2}L_1 \frac{\Delta}{\sqrt{|\mathcal{S}_3|}} \sum_{\tau=1}^{t} \mathbb{E}\big[\|\nabla F(x_\tau)\|\big] + \sqrt{2}L_0 \frac{\Delta}{\sqrt{|\mathcal{S}_3|}} \sqrt{q}.$$

According to Lemma B.1, we have $\mathbb{E}\big[\|\delta_0\|\big] = \mathbb{E}\big[\|\nabla F(x_0) - g_0\|\big] \leq \frac{1}{\sqrt{|\mathcal{S}_1|}} \big(G_0 + G_1\mathbb{E}\big[\|\nabla F(x_0)\|\big]\big)$

$$\begin{aligned} \mathbb{E}\big[\|\delta_t\|\big] &\leq \mathbb{E}\big[\|\delta_0\|\big] + \mathbb{E}\big[\|\hat{\delta}_1 + \cdots + \hat{\delta}_t\|\big] \\ &\leq \frac{1}{\sqrt{|\mathcal{S}_1|}} \big(G_0 + G_1\mathbb{E}\big[\|\nabla F(x_0)\|\big]\big) + \sqrt{2}L_1 \frac{\Delta}{\sqrt{|\mathcal{S}_3|}} \sum_{\tau=1}^{t} \mathbb{E}\big[\|\nabla F(x_\tau)\|\big] + \sqrt{2}L_0 \frac{\Delta}{\sqrt{|\mathcal{S}_3|}} \sqrt{q}. \end{aligned} \tag{24}$$

Because $\Delta = \epsilon$, $|\mathcal{S}_1| = \frac{1}{\epsilon^2}$, $|\mathcal{S}_3| = \frac{1}{\epsilon}$, $q = \frac{1}{8G_1\epsilon}$ and $\epsilon \leq \frac{G_1^2}{2L_1^2}$, we have $\sqrt{2}L_1\epsilon^{3/2} \leq G_1\epsilon$, and

$$\begin{aligned} \mathbb{E}\big[\|\delta_t\|\big] &\leq \left(G_0 + \frac{L_0}{2\sqrt{G_1}}\right)\epsilon + G_1\epsilon\mathbb{E}\big[\|\nabla F(x_0)\|\big] + \sqrt{2}L_1\epsilon^{3/2} \sum_{\tau=1}^{t} \mathbb{E}\big[\|\nabla F(x_\tau)\|\big] \\ &\leq \left(G_0 + \frac{L_0}{2\sqrt{G_1}}\right)\epsilon + G_1\epsilon \sum_{\tau=0}^{t} \mathbb{E}\|\nabla F(x_\tau)\|, \\ \sum_{\tau=0}^{t} \mathbb{E}\big[\|\delta_\tau\|\big] &\leq \left(G_0 + \frac{L_0}{2\sqrt{G_1}}\right)\epsilon \cdot t + G_1\epsilon \cdot t \sum_{\tau=0}^{t} \mathbb{E}\big[\|\nabla F(x_\tau)\|\big] \\ &\leq \left(G_0 + \frac{L_0}{2\sqrt{G_1}}\right)\epsilon \cdot t + G_1\epsilon \cdot q \sum_{\tau=0}^{t} \mathbb{E}\big[\|\nabla F(x_\tau)\|\big] \\ &\leq \left(G_0 + \frac{L_0}{2\sqrt{G_1}}\right)\epsilon \cdot t + \frac{1}{8} \sum_{\tau=0}^{t} \mathbb{E}\big[\|\nabla F(x_\tau)\|\big]. \end{aligned} \tag{25}$$

It is easy to verify that the above inequality also holds for any $t \geq q$. According to the equation (10) of Theorem 1, we are able to bound $\|\nabla F(x_t)\|$

$$\begin{aligned} \mathbb{E}\big[\|\nabla F(x_t)\|\big] \leq& \frac{1}{\Delta}\mathbb{E}\big[F(x_t) - F(x_{t+1})\big] + 2\mathbb{E}\big[\|\nabla F(x_t) - g_t\|\big] + 2\|B_t\|\Delta + \frac{1}{2}\big(L_0 + L_1\mathbb{E}\big[\|\nabla F(x_t)\|\big]\big)\Delta \\ =& \frac{1}{\Delta}\mathbb{E}\big[F(x_t) - F(x_{t+1})\big] + 2\mathbb{E}\big[\|\delta_t\|\big] + 2\beta\Delta + \frac{1}{2}\big(L_0 + L_1\mathbb{E}\big[\|\nabla F(x_t)\|\big]\big)\Delta. \end{aligned} \tag{26}$$

Summing (26) from 0 to $T-1$ and taking expectation on both sides, we have

$$\begin{aligned} &\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla F(x_t)\|\big] \\ &\leq \frac{1}{\Delta}\mathbb{E}\Big[\sum_{t=0}^{T-1} \big(F(x_t) - F(x_{t+1})\big)\Big] + 2\sum_{t=0}^{T-1} \mathbb{E}\big[\|\delta_t\|\big] + 2\beta\Delta T + \frac{1}{2}\sum_{t=0}^{T-1} \big(L_0 + L_1\mathbb{E}\big[\|\nabla F(x_t)\|\big]\big)\Delta \\ &\leq \frac{1}{\epsilon}\mathbb{E}\big[F(x_0) - F(x_T)\big] + \Big(2G_0 + \frac{L_0}{\sqrt{G_1}} + 2\beta + \frac{1}{2}L_0\Big)T\epsilon + \big(\frac{1}{4} + \frac{1}{2}L_1\epsilon\big) \sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla F(x_\tau)\|\big], \end{aligned} \tag{27}$$

and

$$\Big(1 - \frac{1}{4} - \frac{1}{2}L_1\epsilon\Big) \sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla F(x_t)\|\big] \leq \frac{1}{\epsilon}\Delta_F + \Big(2G_0 + \frac{L_0}{\sqrt{G_1}} + 2\beta + \frac{1}{2}L_0\Big)T\epsilon. \tag{28}$$

Since $\epsilon \leq \frac{1}{2L_1}$ and $T = \mathcal{O}(\epsilon^{-2})$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla F(x_t)\|\big] \leq \frac{\Delta_F}{\epsilon T} + \mathcal{O}(\epsilon) = \mathcal{O}(\epsilon). \tag{29}$$

□

F.2 Proof of Theorem 6

Proof. Let $\epsilon_t = \nabla^2 F(x_t) - H_t$. According to Lemma 2, we have

$$\mathbb{E}\,\|\epsilon_t\| \leq (K_0 + K_1\mathbb{E}\|\nabla F(x_t)\|)\sqrt{\epsilon}.$$

According to equation (24) in Theorem 5, we are able to bound $\mathbb{E}\big[\|\delta_t\|\big]$

$$\begin{aligned}\mathbb{E}\big[\|\delta_t\|\big] &\leq \mathbb{E}\big[\|\delta_0\|\big] + \mathbb{E}\big[\|\hat{\delta}_1 + \cdots + \hat{\delta}_t\|\big] \\ &\leq \frac{1}{\sqrt{|\mathcal{S}_1|}}\big(G_0 + G_1\mathbb{E}\big[\|\nabla F(x_0)\|\big]\big) + \frac{\Delta}{\sqrt{|\mathcal{S}_3|}}\sqrt{2}L_1\sum_{\tau=1}^{t}\mathbb{E}\big[\|\nabla F(x_\tau)\|\big] + \frac{\Delta}{\sqrt{|\mathcal{S}_3|}}\sqrt{2q}L_0\end{aligned}$$

Because $\Delta = \sqrt{\epsilon}$, $|\mathcal{S}_1| = \frac{1}{\epsilon^2}$, $\mathcal{S}_2 = \frac{22\log(n)}{\epsilon}$, $|\mathcal{S}_3| = \frac{1}{\epsilon^{3/2}}$, $q = \frac{1}{\epsilon^{1/2}}$ and $\epsilon \leq \frac{G_1^4}{4L_1^4}$, we have

$$\begin{aligned}\mathbb{E}\big[\|\delta_t\|\big] &\leq \mathbb{E}\big[\|\delta_0\|\big] + \mathbb{E}\big[\|\hat{\delta}_1 + \cdots + \hat{\delta}_t\|\big] \\ &\leq \big(G_0 + G_1\mathbb{E}\big[\|\nabla F(x_0)\|\big]\big)\,\epsilon + \sqrt{2}L_1\epsilon^{5/4}\sum_{\tau=1}^{t}\mathbb{E}\big[\|\nabla F(x_\tau)\|\big] + \sqrt{2}L_0\epsilon \\ &\leq \Big(G_0 + \sqrt{2}L_0\Big)\,\epsilon + G_1\epsilon\sum_{\tau=0}^{t}\mathbb{E}\big[\|\nabla F(x_\tau)\|\big].\end{aligned} \tag{30}$$

It follows that

$$\begin{aligned}\sum_{\tau=0}^{t-1}\mathbb{E}\|\delta_\tau\| &\leq \Big(G_0 + \sqrt{2}L_0\Big)\,t\epsilon + G_1\epsilon\cdot t\sum_{\tau=0}^{t-1}\mathbb{E}\|\nabla F(x_\tau)\| \\ &\leq \Big(G_0 + \sqrt{2}L_0\Big)\,t\epsilon + G_1\epsilon\cdot q\sum_{\tau=0}^{t-1}\mathbb{E}\|\nabla F(x_\tau)\| \\ &= \Big(G_0 + \sqrt{2}L_0\Big)\,t\epsilon + G_1\epsilon^{1/2}\sum_{\tau=0}^{t-1}\mathbb{E}\|\nabla F(x_\tau)\|.\end{aligned} \tag{31}$$

According to the equation (17) in the analysis of Theorem 4, we are able to bound $\|\nabla F(x_t)\|$

$$\mathbb{E}[\|\nabla F(x_{t+1})\|] \leq \mathbb{E}\Big[\frac{5M_0 + 5M_1\|\nabla F(x_t)\|}{6}\epsilon + 3\|\nabla F(x_t) - g_t\| + 2\|\nabla^2 F(x_t) - H_t\|\sqrt{\epsilon} + \frac{2}{\sqrt{\epsilon}}(F(x_t) - F(x_{t+1}))\Big]$$

and

$$
\begin{aligned}
&\frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}[\|\nabla F(x_{t+1})\|] \\
&\leq \left(\frac{5M_0}{6}+3G_0+3\sqrt{2}L_0+2K_0\right)\epsilon+\left(\frac{5M_1}{6}\epsilon+3G_1\sqrt{\epsilon}+2K_1\epsilon\right)\frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}[\|\nabla F(x_t)\|] \\
&\quad+\frac{2}{T\sqrt{\epsilon}}\mathbb{E}[F(x_0)-F(x_T)] \\
&\leq \mathcal{O}(\epsilon)+\big(\mathcal{O}(\epsilon)+3G_1\sqrt{\epsilon}\big)\frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}[\|\nabla F(x_{t+1})\|] \\
&\quad+\big(\mathcal{O}(\epsilon)+3G_1\sqrt{\epsilon}\big)\frac{1}{T}\sum_{\tau=0}^{T-1}\big(L_0+L_1\mathbb{E}\|\nabla F(x_{t+1})\|\big)\sqrt{\epsilon}+\frac{2}{T\sqrt{\epsilon}}\Delta_F \\
&\leq \mathcal{O}(\epsilon)+\big(\mathcal{O}(\epsilon)+3G_1\sqrt{\epsilon}\big)\frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}[\|\nabla F(x_{t+1})\|]+\frac{2}{T\sqrt{\epsilon}}\Delta_F.
\end{aligned}
\tag{32}
$$

Since $\epsilon \leq \frac{1}{36G_1^2}$ and $T=\mathcal{O}(\epsilon^{-3/2})$, it follows that

$$
\frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}[\|\nabla F(x_{t+1})\|] \leq \mathcal{O}(\epsilon)+\frac{4}{T\sqrt{\epsilon}}\Delta_J \leq \mathcal{O}(\epsilon). \tag{33}
$$

For the second-order condition, we have

$$
\begin{aligned}
&F(x_{t+1})-F(x_t) \\
&\leq -\frac{1}{2}\lambda_t\|d_{t+1}\|^2+\|\nabla F(x_t)-g_t\|\Delta+\frac{1}{2}\|\nabla^2F(x_t)-H_t\|\Delta^2+\frac{M_0+M_1\|\nabla F(x_t)\|}{6}\Delta^3 \\
&= -\frac{1}{2}\lambda_t\Delta^2+\|\delta_t\|\Delta+\frac{1}{2}\|\epsilon_t\|\Delta^2+\frac{M_0+M_1\|\nabla F(x_t)\|}{6}\Delta^3.
\end{aligned}
\tag{34}
$$

Therefore, we have

$$
\lambda_t \leq \frac{2}{\Delta}\|\delta_t\|+\|\epsilon_t\|+\frac{M_0+M_1\|\nabla F(x_t)\|}{3}\Delta+\frac{2}{\Delta^2}\big(F(x_t)-F(x_{t+1})\big),
$$

and hence

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\lambda_t] \\
&\leq \frac{2}{T\sqrt{\epsilon}}\sum_{t=0}^{T-1}\mathbb{E}\|\delta_t\| + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\epsilon_t\| + \frac{1}{T}\sum_{t=0}^{T-1}\frac{M_0 + M_1\mathbb{E}\|\nabla F(x_t)\|}{3}\sqrt{\epsilon} + \frac{2}{T\epsilon}\mathbb{E}\big[F(x_0) - F(x_T)\big] \\
&\leq \left(2G_0 + 2\sqrt{2}L_0 + K_0 + \frac{M_0}{3}\right)\sqrt{\epsilon} + \left(2G_1 + K_1\sqrt{\epsilon} + \frac{M_1}{3}\sqrt{\epsilon}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_t)\| + \frac{2}{T\epsilon}\Delta_F \\
&\leq \left(2G_0 + 2\sqrt{2}L_0 + K_0 + \frac{M_0}{3}\right)\sqrt{\epsilon} + \left(2G_1 + K_1\sqrt{\epsilon} + \frac{M_1}{3}\sqrt{\epsilon}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_{t+1})\| \qquad (35)\\
&+ \left(2G_1 + K_1\sqrt{\epsilon} + \frac{M_1}{3}\sqrt{\epsilon}\right)\frac{1}{T}\sum_{t=0}^{T-1}\big(L_0 + L_1\mathbb{E}\|\nabla F(x_{t+1})\|\big)\sqrt{\epsilon} + \frac{2}{T\epsilon}\Delta_F \\
&= \mathcal{O}(\sqrt{\epsilon}) + \big(2G_1 + 2L_1G_1 + \mathcal{O}(\sqrt{\epsilon})\big)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_{t+1})\| + \frac{2}{T\epsilon}\Delta_F \\
&\overset{\Diamond}{\leq} \mathcal{O}(\sqrt{\epsilon}) + \mathcal{O}(\epsilon) + \mathcal{O}(\sqrt{\epsilon}) \\
&= \mathcal{O}(\sqrt{\epsilon}),
\end{aligned}
$$

where $\Diamond$ is due to $\mathbb{E}\|\nabla F(x_{\bar{t}+1})\| \leq \mathcal{O}(\epsilon)$ and $T = \mathcal{O}(\epsilon^{-3/2})$.

By the optimality condition (6) and the (M_0, M_1)-smoothness of Hessian, we have

$$
\begin{aligned}
-\mathbb{E}[\lambda_t]I \preceq \mathbb{E}[H_t] &\preceq \mathbb{E}[\|H_t - \nabla^2 F(x_t)\|]I + \mathbb{E}[\nabla^2 F(x_t)] \\
&\preceq \mathbb{E}[\|H_t - \nabla^2 F(x_t)\|]I + \mathbb{E}[\nabla^2 F(x_{t+1})] + \mathbb{E}[\|\nabla^2 F(x_t) - \nabla^2 F(x_{t+1})\|]I \\
&\preceq \mathbb{E}[\|\epsilon_t\|]I + \mathbb{E}[\nabla^2 F(x_{t+1})] + \mathbb{E}\big[\big(M_0 + M_1\|\nabla F(x_{t+1})\|\big)\Delta\big]I.
\end{aligned}
$$

According to Lemma 2, we have

$$
\begin{aligned}
\mathbb{E}\|\epsilon_{\bar{t}}\| = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\epsilon_t\| &\leq K_0\sqrt{\epsilon} + \frac{1}{T}K_1\sqrt{\epsilon}\sum_{\tau=0}^{T-1}\mathbb{E}\|\nabla F(x_t)\| \\
&\leq K_0\sqrt{\epsilon} + K_1\sqrt{\epsilon}\,\mathbb{E}\|\nabla F(x_{\bar{t}+1})\| + \frac{1}{T}K_1\sqrt{\epsilon}\sum_{\tau=0}^{T-1}\mathbb{E}\big[\big(L_0 + L_1\|\nabla F(x_{t+1})\|\big)\Delta\big] \qquad (36)\\
&= \big(K_0\sqrt{\epsilon} + L_0K_1\epsilon\big) + \big(K_1\sqrt{\epsilon} + L_1K_1\epsilon\big)\mathbb{E}\|\nabla F(x_{\bar{t}+1})\| \\
&\leq \mathcal{O}(\sqrt{\epsilon}).
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
-\mathbb{E}[\lambda_{\bar{t}}]I \preceq \mathbb{E}[H_{\bar{t}}] &\preceq \mathbb{E}[\|\epsilon_{\bar{t}}\|]I + \mathbb{E}[\nabla^2 F(x_{\bar{t}+1})] + \mathbb{E}\big[\big(M_0 + M_1\|\nabla F(x_{\bar{t}+1})\|\big)\Delta\big]I \\
&\preceq \mathcal{O}(\sqrt{\epsilon})I + \mathbb{E}[\nabla^2 F(x_{\bar{t}+1})] + \mathbb{E}\big[\big(M_0 + M_1\|\nabla F(x_{\bar{t}+1})\|\big)\sqrt{\epsilon}\big]I \qquad (37)\\
&\preceq \mathcal{O}(\sqrt{\epsilon})I + \mathbb{E}[\nabla^2 F(x_{\bar{t}+1})].
\end{aligned}
$$

Hence, we have

$$\mathbb{E}[\lambda_{\min}(\nabla^2 F(x_{\bar{t}+1}))] \geq -\mathcal{O}(\sqrt{\epsilon}).$$

$\square$

F.3 Proof of Lemma F.1

Proof. We prove that for any $i \in \{0, 1, \cdots, t\}$

$$\mathbb{E}\Big[\Big\|\sum_{\tau=1}^{t}\hat{\delta}_\tau\Big\|\Big] \le \sqrt{2}L_1\frac{\Delta}{\sqrt{|\mathcal{S}_3|}}\sum_{\tau=t-i+1}^{t}\mathbb{E}\big[\|\nabla F(x_\tau)\|\big] + \mathbb{E}\left[\sqrt{2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot i + \Big\|\sum_{\tau=1}^{t-i}\hat{\delta}_\tau\Big\|^2}\right].$$

We prove it by induction. The above inequality holds for $i = 0$. Suppose that it holds for i, we then prove that it also holds for $i + 1$.

$$\begin{aligned}
\mathbb{E}\left[\sqrt{2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot i + \Big\|\sum_{\tau=1}^{t-i}\hat{\delta}_\tau\Big\|^2}\right] &= \mathbb{E}_{\mathcal{F}_{t-i}}\left[\mathbb{E}_{t-i}\sqrt{2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot i + \Big\|\sum_{\tau=1}^{t-i}\hat{\delta}_\tau\Big\|^2}\right] \\
&\le \mathbb{E}_{\mathcal{F}_{t-i}}\left[\sqrt{\mathbb{E}_{t-i}\left[2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot i + \|\sum_{\tau=1}^{t-i}\hat{\delta}_\tau\|^2\right]}\right].
\end{aligned}$$

Because $\hat{\delta}_\tau = \nabla f(x_\tau; \mathcal{S}_3) - \nabla f(x_{\tau-1}; \mathcal{S}_3) - \big(\nabla F(x_\tau) - \nabla F(x_{\tau-1})\big)$ is $\mathcal{F}_{t-i}$-measurable, we have any $1 \le \tau \le t - i - 1$,

$$\begin{aligned}
\mathbb{E}_{t-i}\left[\Big\|\sum_{\tau=1}^{t-i}\hat{\delta}_\tau\Big\|^2\right] &= \mathbb{E}_{t-i}\left[\|\hat{\delta}_{t-i}\|^2 + \Big\|\sum_{\tau=1}^{t-i-1}\hat{\delta}_\tau\Big\|^2 + 2\sum_{\tau=1}^{t-i-1}\hat{\delta}_\tau^T\hat{\delta}_{t-i}\right] \\
&= \mathbb{E}_{t-i}\left[\|\hat{\delta}_{t-i}\|^2\right] + \Big\|\sum_{\tau=1}^{t-i-1}\hat{\delta}_\tau\Big\|^2 + 2\sum_{\tau=1}^{t-i-1}\hat{\delta}_\tau^T\mathbb{E}_{t-i}\left[\hat{\delta}_{t-i}\right].
\end{aligned}$$

In view of $\mathbb{E}_{t-i}\big[\hat{\delta}_{t-i}\big] = 0$, we have

$$\begin{aligned}
\mathbb{E}_{t-i}\left[\Big\|\sum_{\tau=1}^{t-i}\hat{\delta}_\tau\Big\|^2\right] &= \mathbb{E}_{t-i}\left[\|\hat{\delta}_{t-i}\|^2\right] + \Big\|\sum_{\tau=1}^{t-i-1}\hat{\delta}_\tau\Big\|^2 \\
&\le 2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|} + 2L_1^2\frac{\Delta^2}{|\mathcal{S}_3|}\|\nabla F(x_{t-i})\|^2 + \Big\|\sum_{\tau=1}^{t-i-1}\hat{\delta}_\tau\Big\|^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sqrt{2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot i + \Big\|\sum_{\tau=1}^{t-i}\hat{\delta}_\tau\Big\|^2}\right] &\le \mathbb{E}_{\mathcal{F}_{t-i}}\left[\sqrt{\mathbb{E}_{t-i}\left[2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot i + \|\sum_{\tau=1}^{t-i}\hat{\delta}_\tau\|^2\right]}\right] \\
&\le \mathbb{E}_{\mathcal{F}_{t-i}}\left[\sqrt{2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot(i+1) + 2L_1^2\frac{\Delta^2}{|\mathcal{S}_3|}\|\nabla F(x_{t-i})\|^2 + \Big\|\sum_{\tau=1}^{t-i-1}\hat{\delta}_\tau\Big\|^2}\right] \\
&\le \mathbb{E}_{\mathcal{F}_{t-i}}\left[\sqrt{2}L_1\frac{\Delta}{\sqrt{|\mathcal{S}_3|}}\|\nabla F(x_{t-i})\| + \sqrt{2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot(i+1) + \Big\|\sum_{\tau=1}^{t-i-1}\hat{\delta}_\tau\Big\|^2}\right] \\
&= \sqrt{2}L_1\frac{\Delta}{\sqrt{|\mathcal{S}_3|}}\mathbb{E}\big[\|\nabla F(x_{t-i})\|\big] + \mathbb{E}\left[\sqrt{2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot(i+1) + \Big\|\sum_{\tau=1}^{t-i-1}\hat{\delta}_\tau\Big\|^2}\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\Big[\Big\|\sum_{\tau=1}^{t}\hat{\delta}_\tau\Big\|\Big] &\leq \sqrt{2}L_1\frac{\Delta}{\sqrt{|\mathcal{S}_3|}}\sum_{\tau=t-i+1}^{t}\mathbb{E}\big[\|\nabla F(x_\tau)\|\big] + \mathbb{E}\left[\sqrt{2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot i + \Big\|\sum_{\tau=1}^{t-i}\hat{\delta}_\tau\Big\|^2}\right] \\ &\leq \sqrt{2}L_1\frac{\Delta}{\sqrt{|\mathcal{S}_3|}}\sum_{\tau=t-i}^{t}\mathbb{E}\big[\|\nabla F(x_\tau)\|\big] + \mathbb{E}\left[\sqrt{2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot (i+1) + \Big\|\sum_{\tau=1}^{t-i-1}\hat{\delta}_\tau\Big\|^2}\right].\end{aligned}$$

Let $i = t$ and note that $t \leq q = \epsilon^{-1/2}$, we deduce

$$\begin{aligned}\mathbb{E}\Big[\Big\|\sum_{\tau=1}^{t}\hat{\delta}_\tau\Big\|\Big] &\leq \sqrt{2}L_1\frac{\Delta}{\sqrt{|\mathcal{S}_3|}}\sum_{\tau=1}^{t}\mathbb{E}\big[\|\nabla F(x_\tau)\|\big] + \sqrt{2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot t} \\ &\leq \sqrt{2}L_1\frac{\Delta}{\sqrt{|\mathcal{S}_3|}}\sum_{\tau=1}^{t}\mathbb{E}\big[\|\nabla F(x_\tau)\|\big] + \sqrt{2L_0^2\frac{\Delta^2}{|\mathcal{S}_3|}\cdot q}.\end{aligned}$$

□

G Inexactness of Second-Order Trust Region Method

To remedy computational costs in second-order methods, generally, one can apply two different strategies:

- Hessian approximation using probabilistic models (Bandeira et al., 2014; Bollapragada et al., 2019; Xu et al., 2016) and subspace (sketching) techniques (Woodruff, 2014; Berahas et al., 2020; Zhang et al., 2022; Liu et al., 2023).
- Inexact minimization of the subproblems that are frequently imposed in Krylov subspace methods; see Nocedal and Wright (1999); Cartis et al. (2011) for example.

Assumption 4 assembles the common results for second-order Lipschitzian functions[1] that naturally extends for generalized smoothness. Note that it can be satisfied with a large enough Krylov subspace to construct V_t, however, the choice can be flexible (Woodruff, 2014; Cartis et al., 2022). Then we have the following results.

Lemma 3. *Under Assumption 4, then we have*

$$\mathbb{E}_t\left[\|(H_t - \tilde{H}_t)d_{t+1}\|\right] \leq \tilde{C}\Delta^2,$$

where $\tilde{C} = C_0 + 2K_0 + (C_1 + 2K_1)\|\nabla F(x_t)\|$ *and* $d_t \leq \Delta = \sqrt{\epsilon}$.

Proof. Recalling $\tilde{\nabla}^2 F(x_t) = V_t {V_t}^T \nabla^2 F(x_t) V_t {V_t}^T$, we have

$$\begin{aligned}
&\mathbb{E}_t\|(H_t - \tilde{H}_t)d_{t+1}\| \\
&\leq \mathbb{E}_t\left[\|(\nabla^2 F(x_t) - \tilde{\nabla}^2 F(x_t))d_{t+1}\| + \|(H_t - \nabla^2 F(x_t))d_{t+1}\| + \|V_t {V_t}^T(\nabla^2 F(x_t) - H_t)V_t {V_t}^T d_{t+1}\|\right] \\
&\overset{(\Diamond)}{\leq} (C_0 + C_1\|\nabla F(x_t)\|)\Delta^2 + \mathbb{E}_t[\|(H_t - \nabla^2 F(x_t))d_{t+1}\|] + \mathbb{E}_t[\|V_t {V_t}^T\| \cdot \|(\nabla^2 F(x_t) - H_t)d_{t+1}\|] \\
&\overset{(\sharp)}{\leq} (C_0 + C_1\|\nabla F(x_t)\|)\Delta^2 + 2\mathbb{E}_t[\|(\nabla^2 F(x_t) - H_t)d_{t+1}\|] \\
&\overset{(\natural)}{\leq} (C_0 + C_1\|\nabla F(x_t)\|)\Delta^2 + 2 \cdot (K_0 + K_1\|\nabla F(x_t)\|)\sqrt{\epsilon}\Delta = (C_0 + 2K_0 + (C_1 + 2K_1)\|\nabla F(x_t)\|)\Delta^2,
\end{aligned}$$

where $(\Diamond)$ is due to Assumption 4 and the fact that $V_t {V_t}^T d_{t+1} = d_{t+1}$; $(\sharp)$ is due to the fact that ${V_t}^T V_t = I$ and $V_t {V_t}^T$ has the same non-zero eigenvalue with ${V_t}^T V_t$; $(\natural)$ is due to Lemma B.1 and the fact that $d_{t+1} \leq \Delta_t = \Delta = \sqrt{\epsilon}$. ☐

Selecting V_t from span$\{g_t, d_t\}$ Inspired by Zhang et al. (2022); Liu et al. (2023), we find that by introducing a specific V_t, subproblem (5) can be solved equivalently by solving a two-dimensional trust region model. This model is devised to ascertain the step size for both the gradient and momentum within the context of the Heavy Ball method. Notably, this approach provides a significantly simpler alternative to the full-dimensional quadratic program traditionally employed in standard trust region methods, offering increased efficiency without sacrificing the effectiveness of the optimization process.

More specifically, let $d_t = x_t - x_{t-1}$, $B_t = \tilde{H}_t$, where $\tilde{H}_t = V_t V_t^T H_t V_t V_t^T$ and V_t is the orthonormal bases for $\mathcal{L}_t := \operatorname{span}\{g_t, d_t\}$. We find in this case, (5) is equivalent to a two-dimension model; see Zhang et al. (2022).

Lemma 4. *When setting* $B_t = \tilde{H}_t$, *the subproblem (5) is equivalent to*

$$\begin{aligned}
\min_{\alpha \in \mathbb{R}^2} \quad & m_t(\alpha) := F(x_t) + c_t^T \alpha + \frac{1}{2}\alpha^T Q_t \alpha \\
\text{s.t.} \quad & \|\alpha\|_{G_t} \leq \Delta_t,
\end{aligned} \tag{38}$$

if we update by

$$x_{t+1} = x_t - \alpha_t^1 g_t + \alpha_t^2 d_t,$$

[1] For example, see (Xu et al., 2020).

with

$$Q_t = \begin{bmatrix} g_t^T H_t g_t & -d_t^T H_t g_t \\ -d_t^T H_t g_t & d_t^T H_t d_t \end{bmatrix} \in \mathcal{S}^2,$$

$$c_t := \begin{pmatrix} -\|g_t\|^2 \\ g_t^T d_t \end{pmatrix}, \; G_t = \begin{bmatrix} g_t^T g_t & -g_t^T d_t \\ -g_t^T d_t & d_t^T d_t \end{bmatrix},$$

and $\|\alpha\|_{G_t} = \sqrt{\alpha^T G_t \alpha}$.

The vector α_t is the global solution to DRTR problem (38) if it is feasible and there exists a Lagrange multiplier $\lambda_t \geq 0$ such that (α_t, λ_t) is the solution to the following equations:

$$(Q_t + \lambda G_t)\,\alpha + c_t = 0, Q_t + \lambda G_t \succeq 0, \lambda\,(\Delta - \|\alpha\|_{G_t}) = 0. \tag{39}$$

Then by the optimal condition (39), we have the closed form solution of α_t:

$$\alpha_t = -(Q_t + \lambda_t G_t)^{-1} c_t.$$

Since α_t only has two dimensions, it can be easily solved numerically. Therefore, by Lemma 4, we can solve subproblem (5) without explicitly computing the Hessian when setting $B_t = H_t$, although it conceptually utilizes the curvature information. In fact, we only require two additional Hessian-vector products to formulate (38). This leads to the following more practical second-order algorithm in Algorithm 3.

Algorithm 3 Dimension-Reduced Trust Region Method

1: Given T, error ϵ
2: **for** $t = 1, \ldots, T$ **do**
3: Draw samples $|\mathcal{S}_1|$ and compute $g_t = \nabla f(x_t; |\mathcal{S}_1|)$
4: Draw samples S_2 and compute $H_t = \nabla^2 f(x_t; S_2)$
5: Compute λ_t by formalizing Q_t and set Δ_t
6: Compute stepsize α_1, α_2 by solving the DRTR problem (38)
7: Update: $x_{t+1} \leftarrow x_t - \alpha_1 g_t + \alpha_2 d_t$
8: **end for**

G.1 Proof of Theorem 7

Proof. By the definition of $\tilde{H}_t$, we know that $d_{t+1}^{\mathrm{T}} H_t d_{t+1} = d_{t+1} t^{\mathrm{T}} \tilde{H}_t d_{t+1}$, so the derivation of (15) still holds, and we have

$$\|\lambda_t d_{t+1}\| \leq \frac{2}{\sqrt{\epsilon}}(F(x_t) - F(x_{t+1})) + 2\|\nabla F(x_t) - g_t\| + \|\nabla^2 F(x_t) - H_t\|\Delta_t + \frac{M_0 + M_1\|\nabla F(x_t)\|}{3}\|d_{t+1}\|^2. \tag{40}$$

So we have

$$
\begin{aligned}
&\mathbb{E}[\|\nabla F(x_{t+1})\|] \\
&\le \mathbb{E}[\|\nabla F(x_{t+1}) - \nabla F(x_t) - \nabla^2 F(x_t) d_{t+1}\| + \|\nabla F(x_t) - g_t\| + \|(\nabla^2 F(x_t) - H_t) d_{t+1}\| \\
&\quad + \|g_t + \tilde{H}_t d_{t+1}\| + \|(H_t - \tilde{H}_t) d_{t+1}\|] \\
&\overset{(\natural)}{\le} \mathbb{E}\left[\frac{M_0 + M_1\|\nabla F(x_t)\|}{2}\Delta^2 + \|\nabla F(x_t) - g_t\| + \Delta\|\nabla^2 F(x_t) - H_t\| + \|\lambda_t d_{t+1}\| + \tilde{C}\Delta^2\right] \\
&\le \mathbb{E}\left[\frac{5M_0 + 5M_1\|\nabla F(x_t)\| + 6\tilde{C}}{6}\Delta^2 + 3\|\nabla F(x_t) - g_t\| + 2\Delta\|\nabla^2 F(x_t) - H_t\| + \frac{2}{\sqrt{\epsilon}}(F(x_t) - F(x_{t+1})\right] \\
&\le \mathbb{E}\left[\frac{5M_0 + 5M_1\|\nabla F(x_t)\| + 6\tilde{C}}{6}\epsilon + 3(G_0 + G_1\|\nabla F(x_t)\|)\epsilon + 2\epsilon(K_0 + K_1\|\nabla F(x_t)\|) + \frac{2}{\sqrt{\epsilon}}(F(x_t) - F(x_{t+1})\right] \\
&\le \left(\frac{5M_0}{6} + 3G_0 + 4K_0 + C_0\right)\epsilon + \left(\frac{5M_1\epsilon}{6} + 3G_1\epsilon + 4K_1\epsilon + C_1\epsilon\right)\mathbb{E}[\|\nabla F(x_t)\|] + \mathbb{E}\left[\frac{2}{\sqrt{\epsilon}}(F(x_t) - F(x_{t+1})\right] \\
&\le \left(\frac{5M_0}{6} + 3G_0 + 4K_0 + C_0\right)\epsilon + \left(\frac{5M_1\epsilon}{6} + 3G_1\epsilon + 4K_1\epsilon + C_1\epsilon\right)(\mathbb{E}[\|\nabla F(x_{t+1})\| + \|\nabla F(x_{t+1}) - F(x_t)\|]) \\
&\quad + \mathbb{E}[\frac{2}{\sqrt{\epsilon}}(F(x_t) - F(x_{t+1})] \\
&\le \left(\frac{5M_0}{6} + 3G_0 + 4K_0 + C_0\right)\epsilon + \left(\frac{5M_1\epsilon}{6} + 3G_1\epsilon + 4K_1\epsilon + C_1\epsilon\right)\mathbb{E}[\|\nabla F(x_{t+1})\|] \\
&\quad + \mathbb{E}[\left(\frac{5M_1\epsilon}{6} + 3G_1\epsilon + 4K_1\epsilon + C_1\epsilon\right)(L_0 + L1\|\nabla F(x_{t+1})\|)\Delta] + \mathbb{E}[\frac{2}{\sqrt{\epsilon}}(F(x_t) - F(x_{t+1})].
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\left(1 - \frac{5M_1}{6}\epsilon - 3G_1\epsilon - 4K_1\epsilon - C_1\epsilon - \mathcal{O}(\epsilon)\right)\mathbb{E}[\|\nabla F(x_{t+1})\|] \\
&\le \left(\frac{5M_0}{6} + 3G_0 + 4K_0 + C_0\right)\epsilon + \left(\frac{5M_1\epsilon}{6} + 3G_1\epsilon + 4K_1\epsilon + C_1\epsilon\right)L_0\Delta + \mathbb{E}\left[\frac{2}{\sqrt{\epsilon}}(F(x_t) - F(x_{t+1})\right] \\
&\le \mathcal{O}(\epsilon) + \mathbb{E}\left[\frac{2}{\sqrt{\epsilon}}(F(x_t) - F(x_{t+1})\right].
\end{aligned}
$$

Choose ϵ such that $(\frac{5M_1}{6} + 3G_1 + 4K_1 + C_1)\epsilon < 1/2$ and T such that $T = \mathcal{O}(\epsilon^{-3/2})$, we can have

$$
\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla F(x_{t+1})\|\right] \le \mathcal{O}(\epsilon) + \frac{4\Delta_F}{T\sqrt{\epsilon}} = \mathcal{O}(\epsilon). \tag{41}
$$

For the second-order conditions, following the same derivation in the proof of Theorem 4, we have

$$
\mathbb{E}[\lambda_{\bar{t}}] \le \mathcal{O}(\sqrt{\epsilon}),
$$

where $\bar{t}$ is uniformly sampled from $0, \dots, T-1$. By the optimality condition (6), we have

$$
\begin{aligned}
-\mathbb{E}[\lambda_{\bar{t}}]I \preceq \mathbb{E}[\tilde{H}_{\bar{t}}] &\preceq \mathbb{E}[\tilde{\nabla}^2 F(x_{\bar{t}+1})] + \mathbb{E}[\|\tilde{H}_{\bar{t}} - \tilde{\nabla}^2 F_{\bar{t}+1}\|I] \\
&\preceq \mathbb{E}[\tilde{\nabla}^2 F(x_{\bar{t}+1})] + \mathbb{E}[\|\tilde{H}_{\bar{t}} - \tilde{\nabla}^2 F(x_{\bar{t}})\|I] + \mathbb{E}[\|\tilde{\nabla}^2 F(x_{\bar{t}}) - \tilde{\nabla}^2 F(x_{\bar{t}+1})\|I] \\
&\overset{(\diamondsuit)}{\preceq} \mathbb{E}[\tilde{\nabla}^2 F(x_{\bar{t}+1})] + \mathcal{O}(\sqrt{\epsilon})I + \mathcal{O}(\sqrt{\epsilon})\mathbb{E}[\|\nabla F(x_{\bar{t}})\|I] \\
&\preceq \mathbb{E}[\tilde{\nabla}^2 F(x_{\bar{t}+1})] + \mathcal{O}(\sqrt{\epsilon}),
\end{aligned}
$$

where $\diamondsuit$ is due to Assumption 3, Lemma 2 and the fact that $\|V_t V_t^{\mathrm{T}}\| = 1$. So we have

$$
\mathbb{E}[\lambda_{\min}(\tilde{\nabla}^2 F(x_{\bar{t}+1}))] \ge -\mathcal{O}(\sqrt{\epsilon}).
$$

□

H Experiments

We implement the algorithms in PyTorch that enables experiments in neural network training so as to provide a comparison of trust region methods to SGD. The SGD optimizer used in our experiments is provided by the official implementation of PyTorch[2]. As we have shown, the normalized SGD is a first-order variant of our trust region framework (algorithm 1); we use it to represent a first-order trust region method denoted as FOTRGS. To develop a practical second-order variant, we only compute the second-order derivatives in the low-rank subspace, which is enabled by several inquiries of the Hessian-vector product, see Algorithm 3. While such a method is merely an inexact or low-rank trust region method with limited second-order information, we use it to represent the SOTRGS and demonstrate the benefits of high-order information.

For each of the datasets, including MNIST, Fashion MNIST and CIFAR10, we fix the category ratios at

$$\{0.738, 0.986, 0.446, 0.254, 0.768, 0.593, 0.918, 0.731, 0.929, 0.284\}$$

so that the datasets become **imbalanced**, involving $33,260$ out of the original $50,000$ training samples. For the first-order methods, we uniformly set a momentum with parameter $\beta = 0.9$ in all of our experiments, following the standard definition of PyTorch. Since SGD typically needs a smaller learning rate, the search intervals are different for normalized SGD and SGD. **There is no learning rate for the second-order trust region method SOTRGS.** Note again we use normalized SGD to represent a first-order trust region method denoted as FOTRGS hereafter.

H.1 Detailed Results of MNIST and Fashion-MNIST

For MNIST and Fashion MNIST, we use the mini-batches of size 64 and a simple convolutional neural network with two convolutional and two fully-connected layers, which has about 16.84 million parameters. We run the optimizers for 25 epochs and repeat each of the combinations – optimizer $\times$ learning rate η – with 5 runs.

We put **the training curves are reported in the main text**. For completeness, we report the overall average statistics of training loss, training accuracy, and finally test accuracy of all categories. Results on MNIST and Fashion-MNIST with smoothed χ^2 are listed in Table 4a and 4b, respectively, and Table 5a and 5b for results with smoothed CVaR.

Method	η	train_loss	train_acc	test_acc
SOTRGS	-	0.000	100.000	99.036
FOTRGS	0.010	0.000	100.000	99.152
	0.050	0.000	100.000	99.352
	0.100	0.005	99.994	99.174
SGD	0.005	0.001	100.000	99.196
	0.010	0.003	99.996	99.112

(a) Imbalanced MNIST

Method	η	train_loss	train_acc	test_acc
SOTRGS	-	0.001	100.000	91.330
FOTRGS	0.010	0.013	100.000	90.965
	0.050	0.000	100.000	91.602
	0.100	0.009	99.980	90.993
SGD	0.005	0.012	99.997	91.177
	0.010	0.025	99.927	90.955

(b) Imbalanced Fashion-MNIST

Table 4: Tuning results of imbalanced datasets with smoothed χ^2 loss. The results are averaged over 5 runs

We note again that the datasets used in the tests are imbalanced and thus incomplete: each category consists of fewer samples.

H.2 Detailed Results of CIFAR10

The original version of CIFAR10 contains 50,000 training images and 10,000 validation images of size 32×32 with 10. For this dataset, we apply the same **imbalanced** sampling as the previous section that similarly

[2]For details, see https://pytorch.org/docs/stable/generated/torch.optim.SGD.html

Method	η	train_loss	train_acc	test_acc
SOTRGS	-	0.000	100.000	99.020
FOTRGS	0.0010	0.028	99.725	98.610
	0.0050	0.000	100.000	98.975
	0.0010	0.000	100.000	98.971
SGD	0.0001	0.027	99.747	98.615
	0.0005	0.004	100.000	98.885

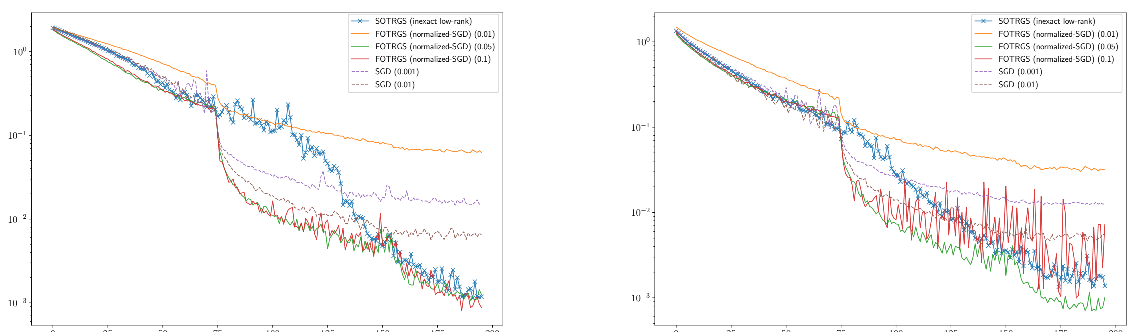
(a) Imbalanced MNIST

Method	η	train_loss	train_acc	test_acc
SOTRGS	-	0.003	100.000	90.880
FOTRGS	0.0010	0.347	92.590	88.760
	0.0050	0.064	99.701	90.590
	0.0100	0.005	100.000	91.150
SGD	0.0001	0.210	95.629	89.610
	0.0005	0.027	100.000	91.150

(b) Imbalanced Fashion-MNIST

Table 5: Tuning results of imbalanced datasets with smoothed CVaR loss.
The results are averaged over 5 runs

gives 33, 260 samples. We train a ResNet-18 model with mini-batches of size 128 and run the optimizers for 200 epochs. For CIFAR10, we decrease the learning rates of SGD and Normalized-SGD $\eta \leftarrow 0.1 \cdot \eta$ at 80, 160 epochs. We provide training results of CIFAR10 with a ResNet18 model. We include tuning curves in Figure 2 after some simple verification. Again SGD needs smaller learning rates so that we choose from $\{0.01, 0.001\}$.



(a) Imbalanced CIFAR10 with χ^2 loss

(b) Imbalanced CIFAR10 with CVaR loss

Figure 2: Training curves of imbalanced CIFAR10.

For testing results, we provide an overall accuracy in Table 6.

More importantly, since we work on imbalanced datasets that inherit disparity amongst categories, we also put emphasis on the test accuracy of the category that takes the least group portion. We present these results in Table 7 to complement the main text.

Method	η	test_acc
SOTRGS	-	88.91
FOTRGS	0.010	83.01
	0.050	89.78
	0.100	89.73
SGD	0.001	88.66
	0.010	89.46

(a) Imbalanced CIFAR10 with smoothed χ^2 loss.

Method	η	test_acc
SOTRGS	-	89.56
FOTRGS	0.010	85.23
	0.050	89.88
	0.100	88.22
SGD	0.001	87.41
	0.010	88.79

(b) Imbalanced CIFAR10 with smoothed CVaR loss.

Table 6: Overall test results of imbalanced CIFAR10.

Y	group ratio in training	SOTRGS	FOTRGS			SGD	
		-	0.005	0.01	0.1	0.001	0.01
0	0.738	0.905	0.867	0.936	0.934	0.913	0.935
1	0.986	0.981	0.908	0.973	0.981	0.975	0.972
2	0.446	0.832	0.804	0.826	0.823	0.821	0.822
3	**0.254**	**0.681**	**0.705**	0.635	0.651	0.608	**0.629**
4	0.768	0.893	0.820	0.939	0.930	0.927	0.932
5	0.593	0.876	0.759	0.898	0.883	0.872	0.877
6	0.918	0.951	0.885	0.961	0.957	0.951	0.963
7	0.731	0.908	0.846	0.944	0.939	0.935	0.942
8	0.929	0.939	0.908	0.957	0.960	0.958	0.950
9	0.284	0.891	0.921	0.902	0.905	0.906	0.904

(a) Tuning results of imbalanced CIFAR10 with χ^2 loss.

Y	group ratio in training	SOTRGS	FOTRGS			SGD	
		-	0.005	0.01	0.1	0.001	0.01
0	0.738	0.901	0.941	0.898	0.917	0.932	0.937
1	0.986	0.981	0.968	0.968	0.969	0.967	0.977
2	0.446	0.855	0.816	0.756	0.787	0.833	0.869
3	**0.254**	**0.616**	**0.615**	0.552	0.562	**0.607**	0.567
4	0.768	0.950	0.955	0.890	0.928	0.925	0.928
5	0.593	0.881	0.893	0.817	0.871	0.887	0.877
6	0.918	0.974	0.963	0.948	0.952	0.948	0.954
7	0.731	0.942	0.945	0.890	0.924	0.934	0.925
8	0.929	0.959	0.962	0.946	0.955	0.950	0.955
9	0.284	0.744	0.929	0.858	0.872	0.896	0.907

(b) Tuning results of imbalanced CIFAR10 with CVaR loss.

Table 7: Per-category test results of imbalanced CIFAR10. The bold text either recognizes the least group or the best test accuracy among the trials of each method.