Trust Region Methods for Nonconvex Stochastic Optimization beyond Lipschitz Smoothness

Chenghan Xie[*1,2], **Chenxi Li**[*1], **Chuwen Zhang**[1], **Qi Deng**[1†], **Dongdong Ge**[1], **Yinyu Ye**[3]

[1]School of Information Management and Engineering, Shanghai University of Finance and Economics
[2]School of Mathematical Sciences, Fudan University
[3]Department of Management Science and Engineering, Stanford University
20307130043@fudan.edu.cn, chenxili@stu.sufe.edu.cn, chuwzhang@gmail.com, qideng@sufe.edu.cn, ge.dongdong@mail.shufe.edu.cn, yyye@stanford.edu

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.

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

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

holds for any $x, x' \in \mathbb{R}^n$ such that $\|x - x'\| \le 1/L_1$, for constants $L_0 > 0, L_1 \ge 0$. 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?*

[*]These authors contributed equally.
[†]Corresponding author

In this paper, we firmly answer this question by proposing an algorithmic framework based on classical trust region methods (Sorensen 1982; Conn, Gould, and Toint 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.

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, Xie, and Li (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, Jadbabaie, and Rakhlin (2023) on Adam, Crawshaw et al. (2022) on unclipped gradient methods, and more recently Sun, Karagulyan, and Richtarik (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. (2023) 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.

Distributionally robust optimization Distributionally robust optimization (DRO) (Delage and Ye 2010), originally designed for a middle ground between stochastic programming (Shapiro, Dentcheva, and Ruszczyński 2014) and robust optimization (Ben-Tal, Ghaoui, and Nemirovski 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, Gould, and Toint 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, Scheinberg, and Shi 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, Mahdavi, and Jin 2013; Wang

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

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, Nguyen, and Tran-Dinh 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.

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 .

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:

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

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 $\psi^*(t) := \sup_{s\in\mathbb{R}}(st - \psi(s))$.

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.

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

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

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.

For generality, we first provide some important properties about the solution of subproblem (5). By the optimality condition of subproblem(Conn, Gould, and Toint (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:

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

Algorithm 1: The trust region framework

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

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*

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

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 $\|B_t\| \le \beta$. By setting $\epsilon \le \min\left\{\frac{4L_0G_0+16\beta G_0}{L_1G_0+2L_0G_1+8\beta G_1}, \frac{4L_0+16\beta}{L_1}\right\}$, $\Delta_t = \Delta = (4L_0+16\beta)^{-1}\epsilon$, $|\mathcal{S}_1| = 64G_0^2\epsilon^{-2}$, $T = 32\Delta_F(L_0+4\beta)\epsilon^{-2}$ in Algorithm 1, we have $\mathbb{E}\|\nabla F(x_{\bar{t}})\| \le \epsilon$, 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 $\epsilon \le \min\left\{\frac{4L_0G_0}{L_1G_0+2L_0G_1}, \frac{4L_0}{L_1}\right\}$, $\Delta_t = \Delta = (4L_0)^{-1}\epsilon$, $|\mathcal{S}_1| = 64G_0^2\epsilon^{-2}$, $T = 32\Delta_F L_0\epsilon^{-2}$, we have $\mathbb{E}\|\nabla F(x_{\bar{t}})\| \le \epsilon$, 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.

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.*

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*

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

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

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

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

and

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

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. 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 4. *In the stochastic setting, it holds that*

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

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 4 hold. Let B_t be a positive semi-definite matrix with bounded norm i.e. there exists a constant β such that $\|B_t\| \le \beta$. By setting $\epsilon \le \min\left\{\frac{G_1^2}{2L_1^2}, \frac{1}{L_1}\right\}$, $\Delta_t = \Delta = \epsilon$, $|\mathcal{S}_1| = \epsilon^{-2}$, $|\mathcal{S}_3| = \epsilon^{-1}$, $q = (8G_1\epsilon)^{-1}$, $T = \mathcal{O}(\epsilon^{-2})$, then we have $\mathbb{E}\|\nabla F(x_{\bar{t}})\| \le \mathcal{O}(\epsilon)$, where $\bar{t}$ is sampled from $\{0, 1, \ldots, T-1\}$ 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 4 hold. Let B_t be the Hessian estimator as shown in Algorithm 2. By setting $\epsilon \le \min\left\{\frac{G_1^4}{4L_1^4}, \frac{1}{36G_1^2}, \frac{1}{L_1^2}\right\}$, $\Delta_t = \Delta = \sqrt{\epsilon}$, $|\mathcal{S}_1| = \epsilon^{-2}$, $|\mathcal{S}_2| = 22\log(n)\epsilon^{-1}$, $|\mathcal{S}_3| = \epsilon^{-3/2}$, $T = \mathcal{O}(\epsilon^{-3/2})$, then we have $\mathbb{E}\|\nabla F(x_{\bar{t}+1})\| \le \epsilon$, $\mathbb{E}[\lambda_{\min}(\nabla^2 F(x_{\bar{t}+1}))] \ge -\mathcal{O}(\sqrt{\epsilon})$, 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, ..., T − 1 do
    if mod(t, q) = 0 then
        Draw samples S_1 and compute g_t = ∇f(x_t; S_1)
    else
        Draw samples S_3 and compute g_t = g_{t−1} + ∇f(x_t; S_3) − ∇f(x_{t−1}; S_3)
    end if
    (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
```

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, Gould, and Toint (2011); Zhang et al. (2022), we propose the following regularity assumption on inexactness in Hessian approximation.

Assumption 5. *For certain constants $C_0, C_1 > 0$, there exists a V_t whose columns form an orthonormal basis such that*

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

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, Roosta, and Mahoney 2020; Cartis, Gould, and Toint 2022). We leave the details in the Appendix.

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 5 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.*

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 the Appendix.

Basic Settings

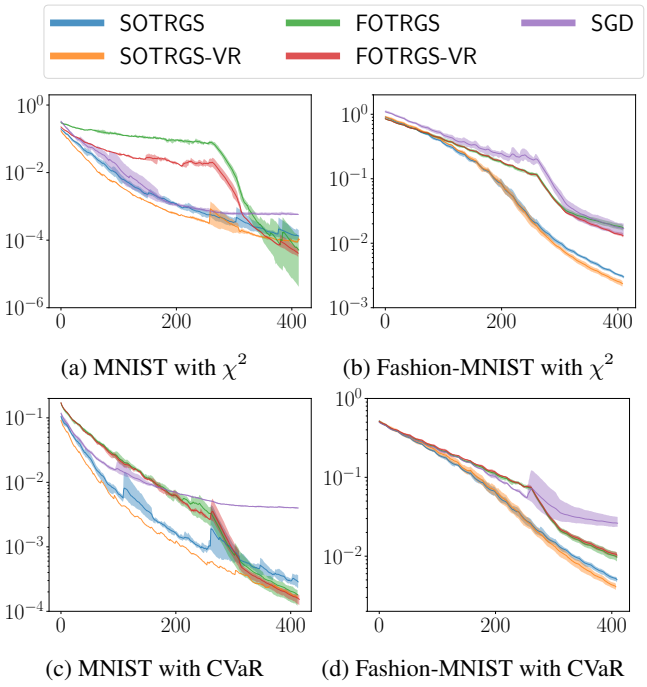
We focus on classification tasks with **imbalanced** distributions arising from applications with heterogeneous (but often latent) subpopulations. Since in standard datasets like MNIST, Fashion MNIST and CIFAR-10, the population ratios (number of images per class) are the same, we create a perturbed 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 the Appendix, and code is available at https://github.com/bzhangcw/pydrsom-dro.

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 (1a) and (1b) present the training curves of SGD, first-order (FOTRGS) and second-order (SOTRGS) trust region methods and also their corresponding variance-reduced variants 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 and can be more robust. We leave test results in the Appendix.

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.



(a) MNIST with χ^2 (b) Fashion-MNIST with χ^2

(c) MNIST with CVaR (d) Fashion-MNIST 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.

	Worst Category	Overall Accuracy
SOTRGS	0.681	0.889
SOTRGS-VR	0.678	0.887
FOTRGS	0.705	0.898
FOTRGS-VR	0.701	0.895
SGD	0.629	0.894

(a) Imbalanced CIFAR10 with χ^2 loss.

	Worst Category	Overall Accuracy
SOTRGS	0.616	0.896
SOTRGS-VR	0.611	0.889
FOTRGS	0.615	0.899
FOTRGS-VR	0.605	0.892
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".

Discussion

This work opens up several intriguing avenues for future exploration, while the question that we are interested in doing next is whether machine learning problems, beyond DRO, exhibit properties of second-order generalized smoothness.

Acknowledgements

This research is partially supported by the National Natural Science Foundation of China (NSFC) [Grant NSFC-72150001, 72225009, 72394360, 72394365]

References

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

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

Ben-Tal, A.; Ghaoui, L. E.; and Nemirovski, A. 2009. *Robust optimization*. Princeton university press. ISBN 978-0-691-14368-2. Publication Title: Robust Optimization.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50(1): 3–44. Publisher: SAGE Publications Inc.

Carmon, Y.; Duchi, J. C.; Hinder, O.; and Sidford, A. 2019. Lower Bounds for Finding Stationary Points I. ArXiv:1710.11606 [math].

Cartis, C.; Gould, N. I. M.; and Toint, P. L. 2011. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, 127(2): 245–295.

Cartis, C.; Gould, N. I. M.; and Toint, P. L. 2022. *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives*. Philadelphia, PA: Society for Industrial and Applied Mathematics. ISBN 978-1-61197-698-4 978-1-61197-699-1.

Chen, Z.; Zhou, Y.; Liang, Y.; and Lu, Z. 2023. Generalized-Smooth Nonconvex Optimization is As Efficient As Smooth Nonconvex Optimization. *arXiv preprint arXiv:2303.02854*.

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

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

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

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

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

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

Duchi, J. C.; and Namkoong, H. 2021. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3): 1378–1406. Publisher: Institute of Mathematical Statistics.

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

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

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

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

Hashimoto, T.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *Proceedings of the 35th International Conference on Machine Learning*, 1929–1938. PMLR. ISSN: 2640-3498.

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

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

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

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

Levy, D.; Carmon, Y.; Duchi, J. C.; and Sidford, A. 2020a. Large-scale methods for distributionally robust optimization. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in neural information processing systems*, volume 33, 8847–8860. Curran Associates, Inc.

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

Li, H.; Jadbabaie, A.; and Rakhlin, A. 2023. Convergence of Adam Under Relaxed Assumptions. *arXiv preprint arXiv:2304.13972*.

Li, H.; Qian, J.; Tian, Y.; Rakhlin, A.; and Jadbabaie, A. 2023. Convex and Non-Convex Optimization under Generalized Smoothness. *arXiv preprint arXiv:2306.01264*.

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

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

Liu, J.; Xie, C.; Deng, Q.; Ge, D.; and Ye, Y. 2023. Stochastic Dimension-reduced Second-order Methods for Policy Optimization. *arXiv preprint arXiv:2301.12174*.

Pan, Y.; and Li, Y. 2023. Toward Understanding Why Adam Converges Faster Than SGD for Transformers. *arXiv preprint arXiv:2306.00204*.

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

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

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

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

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

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

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

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

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

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

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

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

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

Wang, B.; Zhang, H.; Ma, Z.; and Chen, W. 2023. Convergence of AdaGrad for Non-convex Objectives: Simple Proofs and Relaxed Assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, 161–190. PMLR.

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

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

Xu, P.; Roosta, F.; and Mahoney, M. W. 2020. Newton-type methods for non-convex optimization under inexact Hessian information. *Mathematical Programming*, 184(1-2): 35–70.

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

Zhang, C.; Ge, D.; Jiang, B.; and Ye, Y. 2022. DRSOM: A Dimension Reduced Second-Order Method and Preliminary Analyses. *arXiv preprint arXiv:2208.00208*.

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

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

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