Faster Gradient-Free Algorithms for Nonsmooth Nonconvex Stochastic Optimization

Lesi Chen [1] **Jing Xu** [2] **Luo Luo** [1]

Abstract

We consider the optimization problem of the form $\min_{x\in\mathbb{R}^d} f(x) \triangleq \mathbb{E}_\xi[F(x;\xi)]$, where the component $F(x;\xi)$ is L-mean-squared Lipschitz but possibly nonconvex and nonsmooth. The recently proposed gradient-free method requires at most $\mathcal{O}(L^4 d^{3/2}\epsilon^{-4} + \Delta L^3 d^{3/2}\delta^{-1}\epsilon^{-4})$ stochastic zeroth-order oracle complexity to find a (δ,ϵ)-Goldstein stationary point of objective function, where $\Delta = f(x_0) - \inf_{x\in\mathbb{R}^d} f(x)$ and x_0 is the initial point of the algorithm. This paper proposes a more efficient algorithm using stochastic recursive gradient estimators, which improves the complexity to $\mathcal{O}(L^3 d^{3/2}\epsilon^{-3} + \Delta L^2 d^{3/2}\delta^{-1}\epsilon^{-3})$.

1. Introduction

We study the stochastic optimization problem

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

where the stochastic component $F(x;\xi)$, indexed by random variable ξ, is possibly nonconvex and nonsmooth. We focus on tackling the problem with Lipschitz continuous objective, which arises in many popular applications including simulation optimization (Hong et al., 2015; Nelson, 2010), deep neural networks (Nair & Hinton, 2010; Glorot et al., 2011; Chen et al., 2017; Ye et al., 2018), statistical learning (Fan & Li, 2001; Zhang et al., 2006; Zhang, 2010a; Mazumder et al., 2011; Zhang, 2010b), reinforcement learning (Mania et al., 2018; Suh et al., 2022; Choromanski et al., 2018; Jing et al., 2021), financial risk minimization (Stadtler, 2008) and supply chain management (Duffie, 2010).

The Clarke subdifferential (Clarke, 1990) for locally Lipschitz continuous function is a natural extension of gradients for smooth function and subdifferentials for convex functions. Unfortunately, hard instances suggest that finding a (near) ϵ-stationary point in terms of Clarke subdifferential is computationally intractable (Zhang et al., 2020; Kornowski & Shamir, 2021). Zhang et al. (2020) addressed this issue by proposing the notion of (δ,ϵ)-Goldstein stationarity as a valid criterion for non-asymptotic convergence analysis in nonsmooth nonconvex optimization, which considers the convex hull of all Clarke subdifferential at points in the δ-ball neighbourhood (Goldstein, 1977). They also proposed the stochastic interpolated normalized gradient descent method (SINGD) for finding a (δ,ϵ)-Goldstein stationary point of Hadamard directionally differentiable objective, which has a stochastic first-order oracle complexity of $\mathcal{O}(\Delta L^3\delta^{-1}\epsilon^{-4})$, where L is the Lipschitz constant of the objective, $\Delta = f(x_0) - \inf_{x\in\mathbb{R}^d} f(x)$ and x_0 is the initial point of the algorithm. In the case where both exact function value and gradient oracle are available, Davis et al. (2022) proposed the perturbed stochastic interpolated normalized gradient descent (PINGD) to relax the Hadamard directionally differentiable assumption. Later, Tian et al. (2022) proposed the perturbed stochastic interpolated normalized gradient descent (PSINGD) for stochastic settings. Very recently, in a concurrent work Cutkosky et al. (2023) presented an algorithm with a stochastic first-order oracle complexity of $\mathcal{O}(\Delta L^2\delta^{-1}\epsilon^{-3})$ via a novel "Online-to-Nonconvex Conversion" (O2NCC), and they also showed the optimality of their method by constructing a matching lower complexity bound. Besides, there also exist many works focusing on the complexity analysis under the special case when the stochastic first-order oracle is noiseless (Tian & So, 2021; Kornowski & Shamir, 2021; Jordan et al., 2022; Kornowski & Shamir, 2022; Davis et al., 2022; Tian & So, 2022). Remarkably, the hard instance constructed by Tian & So (2022) shows that there is no deterministic algorithm that can compute a (δ,ϵ)-Goldstein stationary point in finite-time within dimension-free complexity. Besides, Tian & So (2022) also proves that any deterministic finite-time algorithm for computing a (δ,ϵ)-Goldstein stationary point cannot be general zero-respecting, which rules out many commonly used first and second-order algorithms.

For many real applications (Hong et al., 2015; Nelson, 2010; Chen et al., 2017; Ye et al., 2018; Stadtler, 2008; Duffie,

[1]School of Data Science, Fudan University, Shanghai, China [2]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. Correspondence to: Luo Luo <luoluo@fudan.edu.cn>.

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

2010; Mania et al., 2018), accessing the first-order oracle may be extremely expensive or even impossible. The randomized smoothing method is a well-known approach to designing zeroth-order algorithms (Nesterov & Spokoiny, 2017; Ghadimi & Lan, 2013). It approximates the first-order oracle using the difference of function values. Most of the existing works on zeroth-order optimization focus on convex problems (Nesterov & Spokoiny, 2017; Shamir, 2017; Duchi et al., 2015; Bubeck et al., 2012; Duchi et al., 2015; Shamir, 2013; Bach & Perchet, 2016) and smooth nonconvex problems (Nesterov & Spokoiny, 2017; Ghadimi & Lan, 2013; Liu et al., 2018; Ji et al., 2019; Fang et al., 2018). Recently, Lin et al. (2022) considered the nonsmooth nonconvex case by establishing the relationship between Goldstein subdifferential (Goldstein, 1977) and randomized smoothing (Nesterov & Spokoiny, 2017; Shamir, 2017). As a consequence, they showed the gradient-free method (GFM) could find a (δ, ϵ)-Goldstein stationary point within at most $\mathcal{O}(L^4 d^{3/2}\epsilon^{-4} + \Delta L^3 d^{3/2}\delta^{-1}\epsilon^{-4})$ stochastic zeroth-order oracle calls.

In this paper, we propose an efficient stochastic gradient-free method named GFM$^+$ for nonsmooth nonconvex stochastic optimization. The algorithm takes advantage of randomized smoothing to construct stochastic recursive gradient estimators (Nguyen et al., 2017; Fang et al., 2018; Li et al., 2021; Li, 2019) for the smoothed surrogate of the objective. It achieves a stochastic zeroth-order oracle complexity of $\mathcal{O}(L^3 d^{3/2}\epsilon^{-3} + \Delta L^2 d^{3/2}\delta^{-1}\epsilon^{-3})$ for finding a (δ, ϵ)-Goldstein stationary point, improving the dependency both on L and ϵ compared with GFM (Lin et al., 2022). In the case of $\delta L \lesssim \Delta$, the dominating term in the zeroth-order oracle complexity of GFM$^+$ becomes $\mathcal{O}(d^{3/2}\Delta L^2 \delta^{-1}\epsilon^{-3})$, which matches the optimal rate of the first-order method O2NCC (Cutkosky et al., 2023) except an unavoidable additional dimensional factor due to we can only access the zeroth order oracles (Duchi et al., 2015). We summarize the results of this paper and related works in Table 1, where the deterministic setting refers to the case that both the exact zeroth-order and first-order oracle are available.

We also extend our results to nonsmooth convex optimization. The optimality of the zeroth-order algorithms to minimize convex function in the measure of function value has been established by Shamir (2017), while the goal of finding approximate stationary points is much more challenging (Allen-Zhu, 2018; Nesterov, 2012; Lee et al., 2021). The lower bound provided by Kornowski & Shamir (2022) suggests finding a point with a small subgradient for Lipschitz convex objectives is intractable. Hence, finding an approximate Goldstein stationary point is also a more reasonable task in convex optimization. We propose the two-phase gradient-free methods to take advantage of the convexity. It shows that GFM$^+$ with warm-start strategy could find a (δ, ϵ)-Goldstein stationary point within $\mathcal{O}\left(L^3 d^{3/2}\epsilon^{-3} + RL^2 d^{4/3}\delta^{-2/3}\epsilon^{-2}\right)$ stochastic zeroth-order oracle complexity, where R is the distance of the initial point to the optimal solution set. We summarize the results for convex case in Table 2.

2. Preliminaries

In this section, we introduce the background for nonsmooth nonconvex function and randomized smoothing technique in zeroth-order optimization.

2.1. Notation and Assumptions

Throughout this paper, we use $\|\cdot\|$ to represent the Euclidean norm of a vector. $\mathbb{B}_\delta(x) \triangleq \{y \in \mathbb{R}^d : \|y - x\| \le \delta\}$ represents the Euclidean ball centered at point x with radius δ. We also define $\mathbb{B} = \mathbb{B}_1(0)$ and $\mathbb{S} \triangleq \{x \in \mathbb{R}^d : \|x\| = 1\}$ as the unit Euclidean ball and sphere centered at origin respectively. We let $\mathrm{conv}\{A\}$ be the convex hull of set A and $\mathrm{dist}(x, A) = \inf_{y\in A} \|x - y\|$ be the distance between vector x and set A. When measuring complexity, we use the notion $\tilde{\mathcal{O}}(\,\cdot\,)$ to hide the logarithmic factors.

Following the standard setting of stochastic optimization, we assume the objective f can be expressed as an expectation of some stochastic components.

Assumption 1. *We suppose that the objective function has the form of $f(x) = \mathbb{E}_\xi[F(x;\xi)]$, where ξ denotes the random index.*

We focus on the Lipschitz function with finite infimum, which satisfies the following assumptions.

Assumption 2. *We suppose that the stochastic component $F(\,\cdot\,;\xi) : \mathbb{R}^d \to \mathbb{R}$ is $L(\xi)$-Lipschitz for any ξ, i.e. it holds that*

$$\|F(x;\xi) - F(y;\xi)\| \le L(\xi)\|x - y\| \tag{2}$$

for any $x, y \in \mathbb{R}^d$ and $L(\xi)$ has bounded second-order moment, i.e. there exists some constant $L > 0$ such that

$$\mathbb{E}_\xi\left[L(\xi)^2\right] \le L^2.$$

Assumption 3. *We suppose that the objective $f : \mathbb{R}^d \to \mathbb{R}$ is lower bounded and define*

$$f^* := \inf_{x\in\mathbb{R}^d} f(x).$$

The inequality (2) in Assumption 2 implies the mean-squared Lipschitz continuity of $F(x;\xi)$.

Proposition 1 (mean-squared continuity)**.** *Under Assumption 2, for any $x, y \in \mathbb{R}^d$ it holds that*

$$\mathbb{E}_\xi\|F(x;\xi) - F(y;\xi)\|^2 \le L^2\|x - y\|^2.$$

Table 1. We present the complexities of finding a (δ, ϵ)-Goldstein stationary point for d-dimensional L-Lipschitz objective under both deterministic and stochastic settings, where we denote $\Delta = f(x_0) - f^*$ and x_0 is the initial point of the algorithm. We remark for "*" that the algorithms proposed by Zhang et al. (2020) use the non-standard first-order oracle called the Hadamard directional derivative.

Setting	Oracle	Method	Reference	Complexity
Stochastic	1st*	SINGD	Zhang et al. (2020)	$\tilde{\mathcal{O}}\left(\frac{\Delta L^3}{\delta\epsilon^4}\right)$
	1st	PSINGD	Tian et al. (2022)	$\tilde{\mathcal{O}}\left(\frac{\Delta L^3}{\delta\epsilon^4}\right)$
	1st	O2NCC	Cutkosky et al. (2023)	$\mathcal{O}\left(\frac{\Delta L^2}{\delta\epsilon^3}\right)$
	0th	GFM	Lin et al. (2022)	$\mathcal{O}\left(d^{3/2}\left(\frac{L^4}{\epsilon^4} + \frac{\Delta L^3}{\delta\epsilon^4}\right)\right)$
	0th	GFM$^+$	Theorem 1	$\mathcal{O}\left(d^{3/2}\left(\frac{L^3}{\epsilon^3} + \frac{\Delta L^2}{\delta\epsilon^3}\right)\right)$
Deterministic	0th & 1st*	INGD	Zhang et al. (2020)	$\tilde{\mathcal{O}}\left(\frac{\Delta L^2}{\delta\epsilon^3}\right)$
	0th & 1st	PINGD	Davis et al. (2022)	$\tilde{\mathcal{O}}\left(\frac{\Delta L^2}{\delta\epsilon^3}\right)$
	0th & 1st	cutting plane	Davis et al. (2022)	$\tilde{\mathcal{O}}\left(\frac{\Delta L d}{\delta\epsilon^2}\right)$

All of above assumptions follow the setting of Lin et al. (2022). We remark that Assumption 2 is weaker than assuming each stochastic component $F(\,\cdot\,;\xi)$ is L-Lipschitz.

2.2. Goldstein Stationary Point

According to Rademacher's theorem, a Lipschitz function is differentiable almost everywhere. Thus, we can define the Clarke subdifferential (Clarke, 1990) as well as approximate Clarke stationary point as follows.

Definition 1 (Clarke subdifferential)**.** *The Clarke subdifferential of a Lipschitz function f at a point x is defined by*

$$\partial f(x) := \operatorname{conv}\left\{g : g = \lim_{x_k \to x} \nabla f(x_k)\right\}.$$

Definition 2 (Approximate Clarke stationary point)**.** *Given a Lipschitz function f, we say the point x is an ϵ-Clarke stationary point of f if it holds that*

$$\operatorname{dist}(0, \partial f(x)) \le \epsilon.$$

However, finding an ϵ-Clarke stationary point is computationally intractable (Zhang et al., 2020). Furthermore, finding a point that is δ-close to an ϵ-stationary point may have an inevitable exponential dependence on the problem dimension (Kornowski & Shamir, 2021). As a relaxation, we pursue the (δ, ϵ)-Goldstein stationary point (Zhang et al., 2020), whose definition is based on the following Goldstein subdifferential (Goldstein, 1977).

Definition 3 (Goldstein subdifferential)**.** *Given $\delta > 0$, the Goldstein δ-subdifferential of aLipschitz function f at a point x is defined by*

$$\partial_\delta f(x) := \operatorname{conv}\left\{\cup_{y \in \mathbb{B}_\delta(x)} \partial f(y)\right\},$$

where $\partial f(x)$ is the Clarke subdifferential.

The (δ, ϵ)-Goldstein stationary point (Zhang et al., 2020) is formally defined as follows.

Definition 4 (Approximate Goldstein stationary point)**.** *Given a Lipschitz function f, we say the point x is a (δ, ϵ)-Goldstein stationary point of f if it holds that*

$$\operatorname{dist}(0, \partial_\delta f(x)) \le \epsilon. \tag{3}$$

Our goal is to design efficient stochastic algorithms to find a (δ, ϵ)-Goldstein stationary point in expectation.

2.3. Randomized Smoothing

Recently, Lin et al. (2022) established the relationship between uniform smoothing and Goldstein subdifferential. We first present the definition of the smoothed surrogate.

Table 2. We present the complexities of finding a (δ, ϵ)-Goldstein stationary point for d-dimensional L-Lipschitz objectives with zeroth-order oracles under the stochastic setting, where we denote $R = \mathrm{dist}(x_0, \mathcal{X}^*)$ and x_0 is the initial point of the algorithm.

Method	Reference	Complexity
GFM	Lin et al. (2022)	$\mathcal{O}\left(\frac{d^{3/2}L^4}{\epsilon^4} + \frac{d^{3/2}\,L^4\,R}{\delta\epsilon^4}\right)$
WS-GFM	Theorem 4	$\mathcal{O}\left(\frac{d^{3/2}L^4}{\epsilon^4} + \frac{d^{4/3}L^{8/3}R^{2/3}}{\delta^{2/3}\epsilon^{4/3}}\right)$
GFM$^+$	Theorem 1	$\mathcal{O}\left(\frac{d^{3/2}L^3}{\epsilon^3} + \frac{d^{3/2}\,L^3\,R}{\delta\epsilon^3}\right)$
WS-GFM$^+$	Theorem 3	$\mathcal{O}\left(\frac{d^{3/2}L^3}{\epsilon^3} + \frac{d^{4/3}L^2R^{2/3}}{\delta^{2/3}\epsilon^2}\right)$

Definition 5 (uniform smoothing)**.** *Given a Lipschitz function f, we denote its smoothed surrogate as*

$$f_\delta(x) := \mathbb{E}_{u\sim\mathcal{P}}[f(x+\delta u)],$$

where $\mathcal{P}$ is the uniform distribution on unit ball $\mathbb{B}$.

The smoothed surrogate has the following properties (Lin et al., 2022, Proposition 2.3 and Theorem 3.1).

Proposition 2. *Suppose that function $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz, then it holds that:*

- $|f_\delta(\,\cdot\,) - f(\,\cdot\,)| \le \delta L$.
- *$f_\delta(\,\cdot\,)$ is L-Lipschitz.*
- *$f_\delta(\,\cdot\,)$ is differentiable and with $c\sqrt{d}L\delta^{-1}$-Lipschitz gradient for some numeric constant $c > 0$.*
- *$\nabla f_\delta(\,\cdot\,) \in \partial_\delta f(\,\cdot\,)$, where $\partial_\delta f(\,\cdot\,)$ is the Goldstein subdifferentiable.*

Flaxman et al. (2004) showed that we can obtained an unbiased estimate of $\nabla f_\delta(x)$ by using the two function value evaluations of points sampled on unit sphere $\mathbb{S}$, which leads to the zeroth-order gradient estimator.

Definition 6 (zeroth-order gradient estimator)**.** *Given a stochastic component $F(\,\cdot\,;\xi) : \mathbb{R}^d \to \mathbb{R}$, we denote its stochastic zeroth-order gradient estimator at $x \in \mathbb{R}^d$ by:*

$$g(x; w, \xi) = \frac{d}{2\delta}(F(x+\delta w;\xi) - F(x-\delta w;\xi))w$$

where $w \in \mathbb{R}^d$ is sampled from a uniform distribution on a unit sphere $\mathbb{S}$.

We also introduce the mini-batch zeroth-order gradient estimator that plays an important role in the stochastic zeroth-order algorithms.

Definition 7 (Mini-batch zeroth-order gradient estimator)**.** *Let $S = \{(\xi_i, w_i)\}_{i=1}^b$, where vectors $w_1, \ldots, w_b \in \mathbb{R}^d$ are i.i.d. sampled from a uniform distribution on unit sphere $\mathbb{S}$ and random indices $\xi_1, \ldots, \xi_b$ are i.i.d. We denote the mini-batch zeroth-order gradient estimator of $f(\,\cdot\,) : \mathbb{R}^d \to \mathbb{R}$ in terms of S at $x \in \mathbb{R}^d$ by*

$$g(x; S) = \frac{1}{b}\sum_{i=1}^{b} g(x; w_i, \xi_i).$$

Next we present some properties for zeroth-order gradient estimators.

Proposition 3 (Lemma D.1 of Lin et al. (2022))**.** *Under Assumption 1 and 2, it holds that*

$$\mathbb{E}_{w,\xi}[g(x; w, \xi)] = \nabla f_\delta(x)$$

and

$$\mathbb{E}_{w,\xi}\|g(x; w, \xi)\|^2 \le 16\sqrt{2\pi}dL^2.$$

Corollary 1. *Let $S = \{(\xi_i, w_i)\}_{i=1}^b$, then under Assumption 1 and 2 it holds that*

$$\mathbb{E}_S\|g(x; S) - \nabla f_\delta(x)\|^2 \le \frac{16\sqrt{2\pi}dL^2}{b}.$$

The following continuity condition of gradient estimator is an essential element for variance reduction.

Proposition 4. *Under Assumption 1 and 2, for any $w \in \mathbb{S}$ and $x, y \in \mathbb{R}^d$, it holds that*

$$\mathbb{E}_\xi\|g(x; w, \xi) - g(y; w, \xi)\|^2 \le \frac{d^2L^2}{\delta^2}\|x-y\|^2.$$

Algorithm 1 GFM (x_0, η, T)

for $t = 0, 1, \cdots, T-1$
  sample a random direction $w_t \in \mathbb{S}$ and a random index ξ_t
  update $x_{t+1} = x_t - \eta g(x_t; w_t, \xi_t)$
end for
return x_{out} chosen uniformly from $\{x_t\}_{t=0}^{T-1}$

To simplify the presentation, we introduce the following notations:

$$L_\delta = \frac{c\sqrt{d}L}{\delta}, \quad \sigma_\delta^2 = 16\sqrt{2\pi}dL^2, \quad \Delta = f(x_0) - f^*$$
$$M_\delta = \frac{dL}{\delta}, \quad f_\delta^* := \inf_{x\in\mathbb{R}^d} f_\delta(x), \quad \Delta_\delta = \Delta + L\delta, \tag{4}$$

where x_0 is the initial point of the algorithm.

Then the above results can be written as

- $\|\nabla f_\delta(x) - \nabla f_\delta(y)\| \le L_\delta \|x - y\|$ for any $x, y \in \mathbb{R}^d$, where f_δ follows Definition 5;
- $\mathbb{E}_\xi \|g(x; w, \xi) - g(y; w, \xi)\|^2 \le M_\delta^2 \|x - y\|^2$ for any $x, y \in \mathbb{R}^d$, where w and ξ follow Definition 6;
- $\mathbb{E}_S \|g(x; S) - \nabla f_\delta(x)\|^2 \le \sigma_\delta^2 / b$ for any $x \in \mathbb{R}^d$, where S and b follow Definition 7;
- $f_\delta(x_0) - f_\delta^* \le \Delta_\delta$.

In Appendix B, we also prove the orders of Lipschitz constants $L_\delta = \mathcal{O}(\sqrt{d}L\delta^{-1})$ and $M_\delta = \mathcal{O}(dL\delta^{-1})$ are tight in general. However, it remains unknown whether the order of σ_δ could be improved.

3. Algorithms and Main Results

This section introduces GFM$^+$ for nonsmooth nonconvex stochastic optimization problem (1). We also provide complexity analysis to show the algorithm has better theoretical guarantee than GFM (Lin et al., 2022).

3.1. The Algorithms

We propose GFM$^+$ in Algorithm 2. Different from GFM (Algorithm 1) (Lin et al., 2022) that uses a vanilla zeroth-order gradient estimator $g(x_t; w_t, \xi_t)$, GFM$^+$ approximates $\nabla f_\delta(x_t)$ by recursive gradient estimator v_t with update rule

$$v_t = v_{t-1} + g(x_t; S) - g(x_{t-1}; S).$$

The estimator v_t properly reduces the variance in estimating $\nabla f_\delta(x)$, which leads to a better stochastic zeroth-order oracle upper complexity than GFM.

Algorithm 2 GFM$^+$$(x_0, \eta, T, m, b, b')$

$v_0 = g(x_0; S')$
for $t = 0, 1, \cdots, T-1$
  if $t \bmod m = 0$
    sample $S' = \{(\xi_i, w_i)\}_{i=1}^{b'}$
    calculate $v_t = g(x_t; S')$
  else
    sample $S = \{(\xi_i, w_i)\}_{i=1}^{b}$
    calculate $v_t = v_{t-1} + g(x_t; S) - g(x_{t-1}; S)$
  end if
  update $x_{t+1} = x_t - \eta v_t$
end for
return x_{out} chosen uniformly from $\{x_t\}_{t=0}^{T-1}$

The variance reduction is widely used to design stochastic first-order and zeroth-order algorithms for smooth nonconvex optimization (Fang et al., 2018; Pham et al., 2020; Wang et al., 2019; Nguyen et al., 2017; Ji et al., 2019; Huang et al., 2022; Liu et al., 2018; Levy et al., 2021; Cutkosky & Orabona, 2019). The existing variance-reduced algorithms for nonsmooth problem (Pham et al., 2020; Reddi et al., 2016; Xiao & Zhang, 2014; Li & Li, 2022) require the objective function to have a composite structure, which does not include our setting that each stochastic component $F(x; \xi)$ could be nonsmooth.

3.2. Complexity Analysis

This subsection considers the upper bound complexity of proposed GFM$^+$. First of all, we present the descent lemma.

Lemma 1 (Li et al. (2021, Lemma 2)). *Under Assumption 1 and 2 , Algorithm 2 holds that*

$$f_\delta(x_{t+1}) \le f_\delta(x_t) - \frac{\eta}{2}\|\nabla f_\delta(x_t)\|^2 + \frac{\eta}{2}\|v_t - \nabla f_\delta(x_t)\|^2 - \left(\frac{\eta}{2} - \frac{L_\delta \eta^2}{2}\right)\|v_t\|^2,$$

where L_δ follows the definition in (4).

Secondly, we show the variance bound of stochastic recursive gradient estimators for smooth surrogate $f_\delta(x)$, which is similar to Lemma 1 of Fang et al. (2018).

Lemma 2 (Variance bound). *Under Assumption 1, 2 , for Algorithm 2 it holds that*

$$\mathbb{E}\|v_{t+1} - \nabla f_\delta(x_{t+1})\|^2 \le \mathbb{E}\left[\|v_t - \nabla f_\delta(x_t)\|^2 + \frac{\eta^2 M_\delta^2}{b}\|v_t\|^2\right],$$

where M_δ follows the definition in (4).

For convenience, we denote $n_t \triangleq \lceil t/m \rceil$ as the index of epoch such that $(n_t - 1)m \leq t \leq n_t m - 1$. Then Lemma 2 leads to the following corollary.

Corollary 2. *Under Assumption 1 and 2, for Algorithm 2 it holds that*

$$\mathbb{E}\|v_{t+1} - \nabla f_\delta(x_{t+1})\|^2 \leq \frac{\sigma_\delta^2}{b'} + \frac{\eta^2 M_\delta^2}{b} \sum_{i=m(n_t-1)}^{t} \mathbb{E}\|v_i\|^2.$$

Combing the descent Lemma 1 and Corollary 2, we obtain the progress of one epoch.

Lemma 3 (One epoch progress)**.** *Under Assumption 1 and 2, for Algorithm 2 it holds that*

$$\begin{aligned}
&\mathbb{E}[f_\delta(x_{mn_t})] \\
&\leq f_\delta(x_{m(n_t-1)}) - \frac{\eta}{2} \sum_{i=m(n_t-1)}^{mn_t-1} \mathbb{E}\|\nabla f_\delta(x_i)\|^2 + \frac{m\eta\sigma_\delta^2}{2b'} \\
&\quad - \left(\frac{\eta}{2} - \frac{L_\delta \eta^2}{2} - \frac{m\eta^3 M_\delta^2}{2b}\right) \sum_{i=m(n_t-1)}^{mn_t-1} \mathbb{E}\|v_i\|^2,
\end{aligned}$$

where $L_\delta, M_\delta, \sigma_\delta$ follow the definition in (4).

Connecting the progress for all of T epochs leads to the following result.

Corollary 3. *Under Assumption 1 and 2 , for Algorithm 2 it holds that*

$$\begin{aligned}
\mathbb{E}[f_\delta(x_T)] \leq f_\delta(x_0) &- \frac{\eta}{2} \sum_{i=0}^{T-1} \mathbb{E}\|\nabla f_\delta(x_i)\|^2 + \frac{T\eta\sigma_\delta^2}{2b'} \\
&- \underbrace{\left(\frac{\eta}{2} - \frac{L_\delta \eta^2}{2} - \frac{m\eta^3 M_\delta^2}{2b}\right)}_{(*)} \sum_{i=0}^{T-1} \mathbb{E}\|v_i\|^2.
\end{aligned} \tag{5}$$

Using Corollary 3 with

$$\eta = \frac{\sqrt{b'}}{mM_\delta}, \quad m = \left\lceil \frac{L_\delta \sqrt{b'}}{M_\delta} \right\rceil \quad \text{and} \quad b = \left\lceil \frac{2b'}{m} \right\rceil,$$

we know that the term (*) in inequality (5) is positive and obtain

$$\mathbb{E}[f_\delta(x_T)] \leq f_\delta(x_0) - \frac{\eta}{2} \sum_{i=0}^{T-1} \mathbb{E}\|\nabla f_\delta(x_i)\|^2 + \frac{T\eta\sigma_\delta^2}{2b'}, \tag{6}$$

which means

$$\mathbb{E}\|\nabla f_\delta(x_{\text{out}})\|^2 \leq \frac{2\Delta\delta}{\eta T} + \frac{\sigma_\delta^2}{b'}. \tag{7}$$

Applying inequality (7) with

$$T = \left\lceil \frac{4\Delta_\delta}{\eta\epsilon^2} \right\rceil \quad \text{and} \quad b' = \left\lceil \frac{2\sigma_\delta^2}{\epsilon^2} \right\rceil, \tag{8}$$

we conclude that the output x_{out} is an ϵ-stationary point of the smooth surrogate $f_\delta(\,\cdot\,)$ in expectation.

Finally, the relationship $\nabla f_\delta(x_{\text{out}}) \in \partial_\delta f(x_{\text{out}})$ shown in Proposition 2 indicates that Algorithm 2 with above parameter settings could output a (δ, ϵ)-Goldstein stationary point of $f(\,\cdot\,)$ in expectation, and the total stochastic zeroth-order oracle complexity is

$$T\left(\left\lceil \frac{b'}{m} \right\rceil + 2b\right) = \mathcal{O}\left(d^{3/2}\left(\frac{L^3}{\epsilon^3} + \frac{\Delta L^2}{\delta\epsilon^3}\right)\right).$$

We formally summarize the result of our analysis as follows.

Theorem 1 (GFM$^+$)**.** *Under Assumption 1, 2 and 3, we run GFM$^+$ (Algorithm 2) with*

$$T = \left\lceil \frac{4\Delta_\delta}{\eta\epsilon^2} \right\rceil, \; b' = \left\lceil \frac{2\sigma_\delta^2}{\epsilon^2} \right\rceil, \; b = \left\lceil \frac{2b'}{m} \right\rceil,$$
$$\eta = \frac{\sqrt{b'}}{mM_\delta} \quad \text{and} \quad m = \left\lceil \frac{L_\delta\sqrt{b'}}{M_\delta} \right\rceil,$$

where $L_\delta, \sigma_\delta, M_\delta$ and Δ_δ follows the definition in (4). Then it outputs a (δ, ϵ)-Goldstein stationary point of $f(\cdot)$ in expectation and the total stochastic zeroth-order oracle complexity is at most

$$\mathcal{O}\left(d^{3/2}\left(\frac{L^3}{\epsilon^3} + \frac{\Delta L^2}{\delta\epsilon^3}\right)\right), \tag{9}$$

where $\Delta = f(x_0) - f^*$.

Theorem 1 achieves the best known stochastic zeroth-order oracle complexity for finding a (δ, ϵ)-Goldstein stationary point of nonsmooth nonconvex stochastic optimization problem, which improves upon GFM (Lin et al., 2022) in the dependency of both ϵ and L.

3.3. The Results for Convex Optimization

This subsection extends the idea of GFM$^+$ to find (δ, ϵ)-Goldstein stationary points for nonsmooth convex optimization problem. We propose warm-started GFM$^+$ (WS-GFM$^+$) in Algorithm 3, which initializes GFM$^+$ with GFM.

The complexity analysis of WS-GFM$^+$ requires the following assumption.

Assumption 4. *We suppose that objective $f : \mathbb{R}^d \to \mathbb{R}$ is convex and the set $\mathcal{X}^* := \arg\min_{x\in\mathbb{R}^d} f(x)$ is non-empty.*

We remark the assumption of non-empty $\mathcal{X}^*$ is stronger than Assumption 3 since the Lipschitzness of f implies $f(x_0) - \inf_{x\in\mathbb{R}^d} f(x) \leq L\,\mathrm{dist}(x_0, \mathcal{X}^*)$. Therefore, under Assumption 4, directly using GFM$^+$ (Algorithm 2) requires

$$\mathcal{O}\left(d^{3/2}\left(\frac{L^3}{\epsilon^3} + \frac{L^3 R}{\delta\epsilon^3}\right)\right) \tag{10}$$

Algorithm 3 WS-GFM$^+$$(x_0, \eta_0, T_0, \eta, T, m, b, b')$

1: $x_1 = \text{GFM}(x_0, \eta_0, T_0)$
2: $x_T = \text{GFM}^+(x_1, \eta, T, m, b, b')$
3: **return** x_T

Algorithm 4 WS-GFM $(x_0, \eta_0, T_0, \eta, T)$

1: $x_1 = \text{GFM}(x_0, \eta_0, T_0)$
2: $x_T = \text{GFM}(x_1, \eta, T)$
3: **return** x_T

iterations to find a (δ, ϵ)-Goldstein stationary point of f, where we denote $R = \text{dist}(x_0, \mathcal{X}^*)$.

Next we show the warm-start strategy can improve the complexity in (10). It is based on the fact that GFM obtains the optimal stochastic zeroth-order oracle complexity in the measure of function value (Shamir, 2017). We include the proof in the appendix for completeness.

Theorem 2 (GFM in function value)**.** *Under Assumption 1, 2 and 4, if we run GFM (Algorithm 1) with* $T = \lceil 2\sigma_\delta R^2/\zeta \rceil$, $\eta = R/(\sigma_\delta \sqrt{T})$ *and* $\delta = \zeta/(4L)$ *where* σ_δ *follows definition in (4), then the output satisfies* $\mathbb{E}[f(x_{\text{out}}) - f^*] \leq \zeta$ *and the total stochastic zeroth-order oracle complexity is at most* $\mathcal{O}(dL^2R^2\zeta^{-2})$*, where* $R = \text{dist}(x_0, \mathcal{X}^*)$.

Theorem 2 means using the output of GFM as the initialization for GFM$^+$ can make the term Δ in (9) be small. We denote $\zeta = f(x_1) - f^*$, then the total complexity in WS-GFM$^+$ is

$$\mathcal{O}\left(\frac{d^{3/2}L^3}{\epsilon^3} + \frac{d^{3/2}L^3\zeta}{\delta\epsilon^3} + \frac{dL^2R^2}{\zeta^2}\right).$$

Then an appropriate choice of ζ leads to the following result.

Theorem 3 (WS-GFM$^+$)**.** *Under Assumption 1, 2 and 4, Algorithm 3 with an appropriate parameter setting can output a* (δ, ϵ)*-Goldstein stationary point in expectation within the stochastic zeroth-order oracle complexity of*

$$\mathcal{O}\left(\frac{d^{3/2}L^3}{\epsilon^3} + \frac{d^{4/3}L^2R^{2/3}}{\delta^{2/3}\epsilon^2}\right),$$

where $R = \text{dist}(x_0, \mathcal{X}^*)$.

Naturally, we can also use the idea of warm-start to obtain the complexity of GFM (Lin et al., 2022) for convex case. We present warm-started GFM (WS-GFM) in Algorithm 4 and provide its theoretical guarantee as follows.

Theorem 4 (WS-GFM)**.** *Under Assumption 1, 2 and 4, Algorithm 4 with an appropriate parameter setting can output a* (δ, ϵ)*-Goldstein stationary point in expectation within the stochastic zeroth-order oracle complexity of*

$$\mathcal{O}\left(\frac{d^{3/2}L^4}{\epsilon^4} + \frac{d^{4/3}L^{8/3}R^{2/3}}{\delta^{2/3}\epsilon^{8/3}}\right),$$

where $R = \text{dist}(x_0, \mathcal{X}^*)$.

4. Numerical Experiments

In this section, we conduct the numerical experiments on nonconvex penalized SVM and black-box attack to show the empirical superiority of proposed GFM$^+$.

4.1. Nonconvex Penalized SVM

We consider the nonconvex penalized SVM with capped-ℓ_1 regularizer (Zhang, 2010b). The model targets to train the binary classifier $x \in \mathbb{R}^d$ on dataset $\{(a_i, b_i)\}_{i=1}^n$, where $a_i \in \mathbb{R}^d$ and $b_i \in \{1, -1\}$ are the feature of the i-th sample and its corresponding label. It is formulated as the following nonsmooth nonconvex problem

$$\min_{x \in \mathbb{R}^d} f(x) \triangleq \frac{1}{n}\sum_{i=1}^n l(b_i a_i^\top x) + r(x),$$

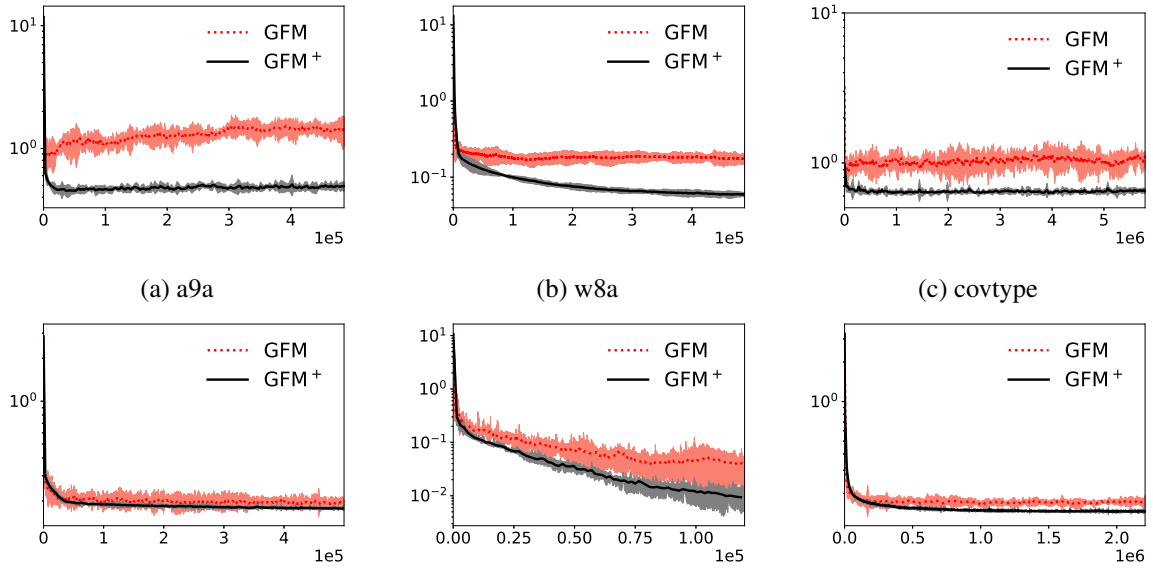
where $l(x) = \max\{1-x, 0\}$, $r(x) = \lambda \sum_{j=1}^d \min\{|x_j|, \alpha\}$ and $\lambda, \alpha > 0$ are hyperparamters. We take $\lambda = 10^{-5}/n$ and $\alpha = 2$ in our experiments.

We compare the proposed GFM$^+$ with GFM (Lin et al., 2022) on LIBSVM datasets "a9a", "w8a", "covtype", "ijcnn1", "mushrooms" and "phishing" (Chang & Lin, 2011). We set $\delta = 0.001$ and tune the stepsize η from $\{0.1, 0.01, 0.001\}$ for the two algorithms. For GFM$^+$, we tune both m and b in $\{1, 10, 100\}$ and set $b' = mb$ by following our theory. We run the algorithms with twenty different random seeds for each dataset and demonstrate the results in Figure 1, where the vertical axis represents the mean of loss calculated by the twenty runs and the horizontal axis represents the zeroth-order complexity. It can be seen that GFM$^+$ leads to faster convergence and improve the stability in the training.

4.2. Black-Box Attack on CNN

We consider the untargeted black-box adversarial attack on image classification with convolutional neural network (CNN). For a given sample $z \in \mathbb{R}^d$, we aim to find a sample $x \in \mathbb{R}^d$ that is close to z and leads to misclassification. We formulate the problem as the following nonsmooth nonconvex problem (Carlini & Wagner, 2017):

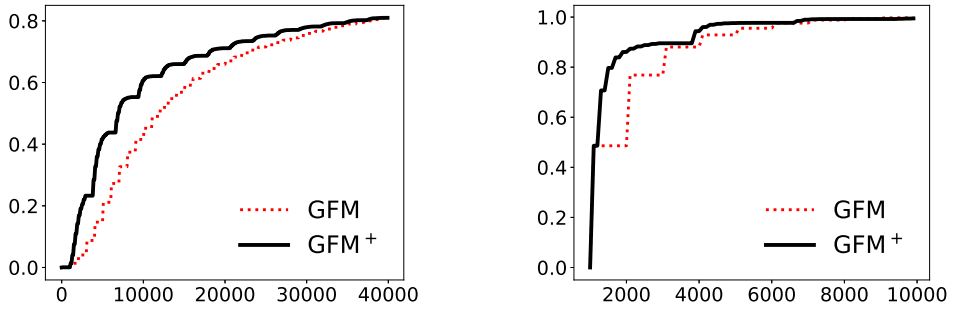
$$\min_{\|x-z\|_\infty \leq \kappa} \max\{\log[Z(x)]_t - \max_{i \neq t} \log[Z(x)]_i, -\theta\},$$



(a) a9a (b) w8a (c) covtype

(e) ijcnn1 (f) mushrooms (g) phishing

Figure 1. For nonconvex penalized SVM, we present the results for **complexity *vs.* loss** on datasets "a9a", "w8a", "covtype", "ijcnn1", "mushrooms" and "phishing". The result for each method is averaged over 20 independent runs.



(a) MNIST (b) Fashion-MNIST

Figure 2. For black-box attack, we present the results for **complexity *vs.* success rate** on datasets "MNIST" and "Fashion-MNIST".

where κ is the constraint level of the distortion, t is the class of sample z and $Z(x)$ is the logit layer representation after softmax in the CNN for x such that $[Z(x)]_i$ represents the predicted probability that x belongs to class i. We set $\theta = 4$ and $\kappa = 0.2$ for our experiments. To address the constraint in the problem, we heuristically conduct additional projection step for the update of x in GFM and GFM$^+$.

We compare the proposed GFM$^+$ with GFM (Lin et al., 2022) on datasets MNIST and Fashion-MNIST. We use a LeNet (LeCun et al., 1998) type CNN which consists of two 5x5 convolution layers two fully-connected layers; each convolution layer is associated with max pooling and ReLU activation; each fully-connected layer is associated with ReLU activation. The detail of the network is presented in Appendix I. We train the CNN with SGD by 100 epochs with stepsize starting with 0.1 and decaying by $1/2$ every 20 epochs. It achieves classification accuracies 98.95% on the MNIST and 91.94% on the Fashion-MNIST. We use GFM$^+$ and GFM to attack the trained CNNs on all of 10,000 images on testsets and set $\delta = 0.01$. For both GFM and GFM$^+$, we tune b' from $\{500, 1000, 2000\}$. For GFM$^+$, we additionally tune m from $\{10, 20, 50\}$ and set $b = b'/m$. For both the two algorithms, we tune the initial learning rate η in $\{0.5, 0.05, 0.005\}$ and decay by $1/2$ if there is no improvement in 10 iterations at one attack. The experiment results are shown in Figure 2, where the vertical axis represents the success attack rate in the whole dataset and the horizontal axis represents the zeroth-order complexity. It shows that the GFM$^+$ can achieve a better attack success rate than the GFM, within the same number of zeroth-order oracles.

5. Conclusion

This paper proposes the novel zeroth-order algorithm GFM$^+$ for stochastic nonsmooth nonconvex optimization. We prove that the algorithm could find a (ϵ, δ)-Goldstein stationary point of the Lipschitz objective in expectation within at most $\mathcal{O}(L^3 d^{3/2}\epsilon^{-3} + \Delta L^2 d^{3/2}\delta^{-1}\epsilon^{-3})$ stochastic zeroth-order oracle complexity, which improves the best known result of GFM (Lin et al., 2022). The numerical experiments on nonconvex penalized SVM and black-box attack on CNN also validate the advantage of GFM$^+$. However, the tightness for the complexity of GFM$^+$ is still unclear. It is interesting to study the zeroth-order and first-order oracle lower bounds for the task of finding approximate Goldstein stationary points.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62206058) and Shanghai Sailing Program (22YF1402900). The authors thank Zhichao Huang for sharing the code for black-box attack on CNN.

References

Allen-Zhu, Z. How to make the gradients small stochastically: Even faster convex and nonconvex SGD. *NeurIPS*, 2018.

Bach, F. and Perchet, V. Highly-smooth zero-th order online optimization. In *COLT*. PMLR, 2016.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. Towards minimax policies for online linear optimization with bandit feedback. In *COLT*, 2012.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017.

Chang, C.-C. and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology*, 2(3):1–27, 2011. URL https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Workshop on AISec*, pp. 15–26, 2017.

Choromanski, K., Rowland, M., Sindhwani, V., Turner, R., and Weller, A. Structured evolution with compact architectures for scalable policy optimization. In *ICML*, 2018.

Clarke, F. H. *Optimization and nonsmooth analysis*. SIAM, 1990.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex SGD. In *NeurIPS*, 2019.

Cutkosky, A., Mehta, H., and Orabona, F. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. *arXiv preprint arXiv:2302.03775*, 2023.

Davis, D., Drusvyatskiy, D., Lee, Y. T., Padmanabhan, S., and Ye, G. A gradient sampling method with complexity guarantees for Lipschitz functions in high and low dimensions. In *NeurIPS*, 2022.

Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Duffie, D. *Dynamic asset pricing theory*. Princeton University Press, 2010.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *NeurIPS*, 2018.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

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

Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In *AISTATS*, 2011.

Goldstein, A. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.

Hong, L. J., Nelson, B. L., and Xu, J. Discrete optimization via simulation. *Handbook of simulation optimization*, pp. 9–44, 2015.

Huang, F., Gao, S., Pei, J., and Huang, H. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal of Machine Learning Research*, 23(36):1–70, 2022.

Ji, K., Wang, Z., Zhou, Y., and Liang, Y. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *ICML*, 2019.

Jing, G., Bai, H., George, J., Chakrabortty, A., and Sharma, P. K. Asynchronous distributed reinforcement learning for lqr control via zeroth-order block coordinate descent. *arXiv preprint arXiv:2107.12416*, 2021.

Jordan, M. I., Lin, T., and Zampetakis, M. On the complexity of deterministic nonsmooth and nonconvex optimization. *arXiv preprint arXiv:2209.12463*, 2022.

Kornowski, G. and Shamir, O. Oracle complexity in nonsmooth nonconvex optimization. *NeurIPS*, 2021.

Kornowski, G. and Shamir, O. On the complexity of finding small subgradients in nonsmooth optimization. *arXiv preprint arXiv:2209.10346*, 2022.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, J., Park, C., and Ryu, E. A geometric structure of acceleration and its role in making gradients small fast. *NeurIPS*, 2021.

Levy, K., Kavis, A., and Cevher, V. STORM$^+$: Fully adaptive SGD with recursive momentum for nonconvex optimization. In *NeurIPS*, 2021.

Li, Z. SSRGD: Simple stochastic recursive gradient descent for escaping saddle points. In *NeurIPS*, 2019.

Li, Z. and Li, J. Simple and optimal stochastic gradient methods for nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(239):1–61, 2022.

Li, Z., Bao, H., Zhang, X., and Richtárik, P. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *ICML*, 2021.

Lin, T., Zheng, Z., and Jordan, M. I. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *arXiv preprint arXiv:2209.05045*, 2022.

Liu, S., Kailkhura, B., Chen, P.-Y., Ting, P., Chang, S., and Amini, L. Zeroth-order stochastic variance reduction for nonconvex optimization. In *NeurIPS*, 2018.

Mania, H., Guy, A., and Recht, B. Simple random search provides a competitive approach to reinforcement learning. *arXiv preprint arXiv:1803.07055*, 2018.

Mazumder, R., Friedman, J. H., and Hastie, T. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

Nelson, B. L. Optimization via simulation over discrete decision variables. *Risk and optimization in an uncertain world*, pp. 193–207, 2010.

Nesterov, Y. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, 88:10–11, 2012.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.

Pham, N. H., Nguyen, L. M., Phan, D. T., and Tran-Dinh, Q. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.

Reddi, S. J., Sra, S., Poczos, B., and Smola, A. J. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *NIPS*, 2016.

Shamir, O. On the complexity of bandit and derivative-free stochastic convex optimization. In *COLT*. PMLR, 2013.

Shamir, O. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(1):1703–1713, 2017.

Stadtler, H. Supply chain management—an overview. *Supply chain management and advanced planning*, pp. 9–36, 2008.

Suh, H. J., Simchowitz, M., Zhang, K., and Tedrake, R. Do differentiable simulators give better policy gradients? In *ICML*, 2022.

Tian, L. and So, A. M.-C. On the hardness of computing near-approximate stationary points of clarke regular nonsmooth nonconvex problems and certain DC programs. In *ICML Workshop on Beyond First-Order Methods in ML Systems*, 2021.

Tian, L. and So, A. M.-C. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *arXiv preprint arXiv:2210.06907*, 2022.

Tian, L., Zhou, K., and So, A. M.-C. On the finite-time complexity and practical computation of approximate stationarity concepts of lipschitz functions. In *ICML*, 2022.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. SpiderBoost and momentum: Faster variance reduction algorithms. In *NeurIPS*, 2019.

Xiao, L. and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Ye, H., Huang, Z., Fang, C., Li, C. J., and Zhang, T. Hessian-aware zeroth-order optimization for black-box adversarial attack. *arXiv preprint arXiv:1812.11377*, 2018.

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2): 894–942, 2010a.

Zhang, H. H., Ahn, J., Lin, X., and Park, C. Gene selection using support vector machines with non-convex penalty. *bioinformatics*, 22(1):88–95, 2006.

Zhang, J., Lin, H., Jegelka, S., Sra, S., and Jadbabaie, A. Complexity of finding stationary points of nonsmooth nonconvex functions. In *ICML*, 2020.

Zhang, T. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11(3), 2010b.

A. The Proof of Proposition 4

Proof. According to the definition of $g(x; w, \xi)$, we use Young's inequality and Proposition 1 to obtain

$$\begin{aligned}
&\mathbb{E}_\xi \|g(x; w, \xi) - g(y; w, \xi)\|^2 \\
\leq\ & \frac{d^2}{2\delta^2}\mathbb{E}_\xi \|F(x + \delta w; \xi) - F(y + \delta w; \xi)\|^2 + \frac{d^2}{2\delta^2}\mathbb{E}_\xi \|F(x - \delta w; \xi) - F(y - \delta w; \xi)\|^2 \\
\leq\ & \frac{d^2 L^2}{\delta^2}\|x - y\|^2.
\end{aligned}$$

□

B. The Tightness of L_δ and M_δ

We formally show the tightness of L_δ and M_δ as follows.

Proposition 5. *The order of Lipschitz constant $L_\delta = \mathcal{O}(\sqrt{d}L\delta^{-1})$ for the gradient of $f_\delta(\,\cdot\,)$ is tight.*

Proof. According to Lemma 10 of Duchi et al. (2012), it holds that

$$E_U(f)E_V(f) \geq c' L^2 \sqrt{d}$$

for some constant $c' > 0$, where $E_U(f)$ and $E_V(f)$ are defined as

$$E_U(f) = \inf_{\kappa \in \mathbb{R}} \sup_{x \in \mathbb{R}^d} \{|f(x) - f_\delta(x)| \leq \kappa\}$$

and

$$E_V(f) = \inf_{\kappa \in \mathbb{R}} \sup_{x, y \in \mathbb{R}^d} \{\|\nabla f_\delta(x) - \nabla f_\delta(y)\| \leq \kappa \|x - y\|\}.$$

The above lower bound can be taken by the function

$$f(x) = \frac{L}{2}\|x\| + \frac{L}{2}\left|\frac{x^\top y}{\|y\|^2} - \frac{1}{2}\right|.$$

Using Proposition 3, we know that $E_U(f) \leq L\delta$ and it immediately implies our claim. □

Proposition 6. *The order of mean-squared Lipschitz constant $M_\delta = \mathcal{O}(dL\delta^{-1})$ for the stochastic zeroth-order gradient estimator $g(\,\cdot\,; w, \xi)$ is tight.*

Proof. We construct a function $f(x) = F(x; \xi) = \frac{L}{\sqrt{d}} \sum_{i=1}^d |x_i|$ for any random vector ξ. It is clear that each $F(x; \xi)$ is L-Lipschitz for any random index ξ by noting:

$$|F(x; \xi) - F(y; \xi)| = \frac{L}{\sqrt{d}}|\|x\|_1 - \|y\|_1| \leq \frac{L}{\sqrt{d}}\|x - y\|_1 \leq L\|x - y\|_2.$$

Choosing $x = \delta \mathbf{1}$ and $y = -\delta \mathbf{1}$, we know that $\|x - y\|_2 = 2\sqrt{d}\delta$. Hence, we further know that (a) $\triangleq M_\delta \|x - y\|_2 = 2d^{3/2}L$. Next, we calculate the zeroth order gradient for $w = 1/\sqrt{d} \cdot \mathbf{1}$ and verify that $g(x; w, \xi) = Ld\mathbf{1}$. Similarly, we also have $g(y; w, \xi) = -Ld\mathbf{1}$. Therefore, (b) $\triangleq \|g(x; w, \xi) - g(y; w, \xi)\|_2 = 2d^{3/2}L$. Finally, we complete the proof by noting that term (a) is exactly the same as term (b).

□

C. The Proof of Lemma 1

Proof. We have

$$\begin{aligned} f_\delta(x_{t+1}) &\le f_\delta(x_t) + \nabla f_\delta(x_t)^\top (x_{t+1} - x_t) + \frac{L_\delta}{2}\|x_{t+1} - x_t\|^2 \\ &= f_\delta(x_t) - \eta \nabla f_\delta(x_t)^\top v_t + \frac{L_\delta \eta^2}{2}\|v_t\|^2 \\ &= f_\delta(x_t) - \frac{\eta}{2}\|\nabla f_\delta(x_t)\|^2 - \left(\frac{\eta}{2} - \frac{L_\delta \eta^2}{2}\right)\|v_t\|^2 + \frac{\eta}{2}\|v_t - \nabla f_\delta(x_t)\|^2, \end{aligned}$$

where the first inequality uses the L_δ-smoothness of f_δ in Proposition 2; the second line follows the update $x_{t+1} = x_t - \eta v_t$; the last step uses fact $2a^\top b = \|a\|^2 + \|b\|^2 - \|a-b\|^2$ for any $a, b \in \mathbb{R}^d$. □

D. The Proof of Lemma 2

Proof. The update of v_{t+1} leads to

$$\begin{aligned} \mathbb{E}\|v_{t+1} - \nabla f_\delta(x_{t+1})\|^2 &= \mathbb{E}\|v_t - g(x_t; S) + g(x_{t+1}; S) - \nabla f_\delta(x_{t+1})\|^2 \\ &= \mathbb{E}\|v_t - \nabla f_\delta(x_t)\|^2 + \mathbb{E}\|g(x_{t+1}; S) - g(x_t; S) + \nabla f_\delta(x_t) - \nabla f_\delta(x_{t+1})\|^2 \\ &\le \mathbb{E}\|v_t - \nabla f_\delta(x_t)\|^2 + \frac{1}{b}\mathbb{E}\|g(x_{t+1}; w_1, \xi_1) - g(x_t; w_1, \xi_1)\|^2 \\ &\le \mathbb{E}\|v_t - \nabla f_\delta(x_t)\|^2 + \frac{\eta^2 M_\delta^2}{b}\mathbb{E}\|v_t\|^2, \end{aligned}$$

where the first line follows the update rule; the second line use the property of martingale (Fang et al., 2018, Proposition 1); the first inequality uses the fact $\mathbb{E}\left\|\sum_{i=1}^m a_i\right\|^2 = m\mathbb{E}\|a_1\|^2$ for any i.i.d random vector $a_1, a_2 \cdots, a_m \in \mathbb{R}^d$ with zero-mean and Definition 7; the second inequality is obtained by Proposition 4 and the update $x_{t+1} = x_t - \eta v_t$. □

E. The Proof of Lemma 3 and Corollary 3

Proof. We just need to plug the result of Corollary 2 into Lemma 1. □

F. The Proof of Theorem 1

Proof. The parameter settings of this theorem guarantee that the term $(*)$ in Corollary 3 be positive and we can drop it to obtain inequalities (6) and (7). Using the Jensen's inequality $\mathbb{E}\|\nabla f_\delta(x_{\text{out}})\| \le \sqrt{\mathbb{E}\|\nabla f_\delta(x_{\text{out}})\|^2}$ and relationship $\nabla f_\delta(x_{\text{out}}) \in \partial_\delta f(x_{\text{out}})$ shown in Proposition 2, we know that the output satisfies $\mathbb{E}\text{dist}(0, \partial_\delta f(x)) \le \epsilon$. The overall stochastic zeroth-order oracle complexity is obtained by plugging definitions in (4) into $T\left(\lceil b'/m \rceil + 2b\right)$. □

G. The Proof of Theorem 2

Proof. For any $x^* \in \mathcal{X}^*$, it holds that

$$\begin{aligned} &\mathbb{E}[f_\delta(x_t) - f_\delta(x^*)] \\ &\le \mathbb{E}\left[\nabla f_\delta(x_t)^\top (x_t - x^*)\right] \\ &= \frac{1}{\eta}\mathbb{E}\left[(x_t - x_{t+1})^\top (x_t - x^*)\right] \\ &= \frac{1}{2\eta}\mathbb{E}\left[\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 + \|x_{t+1} - x_t\|^2\right] \\ &\le \frac{1}{2\eta}\mathbb{E}\left[\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right] + \frac{\eta\sigma_\delta^2}{2}, \end{aligned}$$

where the first inequality uses the convexity of $f_\delta(\,\cdot\,)$ and the second inequality uses Proposition 3. Combining the Jensen's inequality, we have

$$\mathbb{E}[f_\delta(x_{\rm out}) - f_\delta(x^*)] \le \frac{R^2}{2\eta T} + \frac{\eta\sigma_\delta^2}{2} = \frac{R\sigma_\delta}{2\sqrt{T}}. \tag{11}$$

Since Proposition 2 means $\|f - f_\delta\|_\infty \le L\delta$, the results of (11) implies

$$\mathbb{E}[f(x_{\rm out}) - f^*] \le \mathbb{E}[f_\delta(x_{\rm out}) - f_\delta(x^*)] + 2L\delta \le \frac{R\sigma_\delta}{2\sqrt{T}} + 2L\delta.$$

□

H. Proof of Theorem 3 and 4

Proof. For Algorithm 4, the complexity is given by

$$\mathcal{O}\left(\frac{d^{3/2}L^4}{\epsilon^4} + \frac{d^{3/2}L^4\zeta}{\delta\epsilon^4} + \frac{dR^2L^2}{\zeta^2}\right).$$

For Algorithm 3, the complexity is given by

$$\mathcal{O}\left(\frac{d^{3/2}L^3}{\epsilon^3} + \frac{d^{3/2}L^3\zeta}{\delta\epsilon^3} + \frac{dR^2L^2}{\zeta^2}\right).$$

Then we tune ζ to achieve the best upper bound using the inequality $3\sqrt[3]{a^2b} \le 2a\zeta + b/\zeta^2$ for any constant $a > 0, b > 0$. □

Table 3. Summary of datasets used in our SVM experiments

Dataset	n	d
a9a	48,842	123
w8a	64,700	300
covtype	581,012	54
ijcnn1	49,990	22
mushrooms	8,142	112
phishing	11,055	68

Table 4. Network used in the experiments

Layer	Size
Convolution	$5 \times 5 \times 16$
ReLU	-
Max Pooling	2×2
Convolution	$5 \times 5 \times 16$
ReLU	-
Max Pooling	2×2
Linear	3136×128
ReLU	-
Linear	128×10

I. The Details of Experiments

The details of the datasets for the experiments of nonconvex penalized SVM is shown in Table 3. The network used in the experiments of black box attack is shown in Table 4.