# ZEROTH-ORDER SGD WITH MOMENTUM RESETTING ACHIEVES DIMENSION-DEPENDENT OPTIMALITY IN NONSMOOTH NONCONVEX OPTIMIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We investigate the complexity of zeroth-order stochastic gradient methods with momentum for finding $(\delta, \epsilon)$-stationary points of Lipschitz-continuous objectives that may be both nonsmooth and nonconvex, where the algorithm has access only to function values. Recent work has developed specialized zeroth-order schemes that achieve the best-known dependence on the key problem parameters in this setting. However, these guarantees rely on particular algorithmic constructions and do not directly apply to classical SGD-type methods, whose existing analyses typically inherit an additional dependence on the smoothness constant of the smoothed objective. In this work, we study a simple variant of zeroth-order SGD with momentum (ZO-SGDM), which retains the standard SGDM update while incorporating an epoch-wise outer loop with periodic momentum resetting. By exploiting the recursive averaging structure of momentum, we show that this ZO-SGDM-type method achieves an oracle complexity of $\mathcal{O}\left(d\delta^{-1}\epsilon^{-3}\right)$ for finding a $(\delta, \epsilon)$-stationary point, matching the best-known dependence on $d$, $\delta$, and $\epsilon$ attained by specialized zeroth-order methods. Beyond establishing the complexity bound, our analysis offers a clearer understanding of the origin of dimension dependence in zeroth-order optimization. It shows that the dimensional penalty can be attributed entirely to zeroth-order gradient estimation, while the underlying SGDM optimization dynamics introduce no additional dependence on $d$. This distinction helps separate the intrinsic cost of zeroth-order information from artifacts introduced by smoothness-based analyses.

## 1 INTRODUCTION

In this work, we study the stochastic optimization problem

$$\min_{w \in \mathbb{R}^d} f(w) := \mathbb{E}_{Z \sim \mathcal{D}}[\ell(w; Z)], \tag{1}$$

where $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}_+$ is a nonnegative loss, and $\ell(w; z)$ denotes the loss incurred at data point $z \in \mathcal{Z}$ under parameter $w \in \mathbb{R}^d$; $\mathcal{D}$ is a distribution over the measurable space $\mathcal{Z}$. Throughout, we assume that $\ell(\cdot; z)$ is Lipschitz continuous in its first argument, while allowing it to be potentially nonconvex and nonsmooth.

To solve the problem in (1), we consider the following generalized zeroth-order stochastic gradient descent with momentum (ZO-SGDM) update:

$$w_t = w_{t-1} - \eta\, m_{t-1}, \tag{2}$$

$$v_t = w_t + \gamma\,(w_t - w_{t-1}), \tag{3}$$

$$m_t = (1-\beta)\, m_{t-1} + \beta\, \hat{g}_\nu(v_t; z_t), \qquad z_t \sim \mathcal{D}. \tag{4}$$

The zeroth-order (ZO) stochastic gradient estimator $\hat{g}_\nu$ we consider is based on two-point finite differences with a single random direction. Let $\nu > 0$ be the smoothing radius, $d$ the dimension. For any $(w, z)$, sample $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and define

$$\hat{g}_\nu(w; z) := \frac{d}{2\nu}\Big(\ell(w + \nu u; z) - \ell(w - \nu u; z)\Big)\, u. \tag{5}$$



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Equivalently, the update can be re-expressed as

$$w_{t+1} = w_t - \eta\beta\,\hat{g}_\nu(v_t; z_t) + (1-\beta)(w_t - w_{t-1}).$$

When $\gamma = 0$, (2)–(4) reduces to the standard ZO-SGDM. When $\gamma = 1-\beta$, it corresponds to a ZO analogue of Nesterov-type momentum where the ZO estimator is evaluated at the extrapolated point $w_t + (1-\beta)(w_t - w_{t-1})$. Throughout this paper, we assume without loss of generality that $\gamma \in [0, 1]$.

Such algorithms are particularly important, since zeroth-order SGD and its momentum-based variants are widely adopted in settings where gradients are unavailable, unreliable, or prohibitively expensive to compute and only function-value queries are accessible. Representative applications include memory-efficient fine-tuning of large language models (Malladi et al., 2023; Chen et al., 2024; Zhang et al., 2024), black-box adversarial evaluation (Ilyas et al., 2018; Chen et al., 2017), and simulation-driven optimization (Salimans et al., 2017).

Owing to its widespread use, research on this topic has continued to advance in recent years. In the convex setting, these problems have been extensively studied and are by now well understood (Duchi et al. (2015)). For nonsmooth nonconvex objectives, however, the landscape is markedly different: since Zhang et al. (2020) show that, for general nonsmooth nonconvex functions, one cannot guarantee finding an $\epsilon$-stationary point in finite time, attention has shifted in recent years toward weaker yet meaningful notions of stationarity, spurring a growing body of work (Davis et al. (2022); Kong & Lewis (2025); Lin et al. (2022); Kornowski & Shamir (2024); Tian & So (2024); Jordan et al. (2023)). It is also worth noting that methodological advances in first-order optimization often directly influence the development of zeroth-order methods: ideas first introduced for nonsmooth nonconvex first-order algorithms were subsequently adapted to the zeroth-order nonsmooth nonconvex regime, ultimately yielding state-of-the-art dimension dependence by Kornowski & Shamir (2024). For a concrete overview of recent progress in zeroth-order nonsmooth nonconvex optimization, we refer the reader to the related work in Section 2 and further related work in Section A.

Since first-order methods do not incur an explicit dimension dependence, one may expect that zeroth-order optimization should introduce dimension dependence primarily through the gradient estimator itself. However, existing sharp zeroth-order guarantees rely on the specialized algorithmic framework and do not directly recover the update rule of ZO-SGDM. This raises a natural question:

> Can one obtain the optimal dimension dependence while making only mild modifications to the classical ZO-SGDM?

We introduce two modifications to standard ZO-SGDM: periodic momentum resetting and epoch averaging, and refer to the resulting method as ZO-SGDM-PR; see Algorithm 1. Momentum resetting limits long-term momentum accumulation, placing ZO-SGDM-PR heuristically between ZO-SGD and standard ZO-SGDM: with frequent resets, the dynamics become more SGD-like, whereas with longer reset periods they increasingly approach those of standard ZO-SGDM. Epoch averaging is a standard iterate-averaging device commonly used in stochastic optimization and does not alter the SGDM updates within each epoch. Thus, ZO-SGDM-PR preserves the local SGDM dynamics while introducing inter-epoch modifications that facilitate the analysis.

**Contributions.** We show that a ZO-SGDM-type method achieves an oracle complexity of $\mathcal{O}(d\delta^{-1}\epsilon^{-3})$ for finding a $(\delta, \epsilon)$-stationary point in stochastic nonsmooth nonconvex zeroth-order optimization. The method augments classical ZO-SGDM with periodic momentum resetting and epoch averaging while retaining the standard SGDM recursion within each epoch. The resulting guarantee matches the best-known dependence on $d$, $\delta$, and $\epsilon$.

Our main technical contribution is a direct analysis of the SGDM recursion through the weak convexity of the smoothed objective. By controlling the spread of iterates within each epoch and relating the terminal momentum to a weighted average of smoothed gradients, we bound the regularized stationarity measure of the epoch average without invoking the smoothness constant of the smoothed objective. Consequently, the dimension dependence in the leading oracle complexity comes solely from the second moment of the zeroth-order estimator.

Table 1: Comparison of zeroth-order methods for nonsmooth nonconvex optimization.

| Setting | Reference | Algorithm | Algorithmic $d$-factor | $(\delta, \epsilon)$-stationarity |
|---|---|---|---|---|
| Deterministic | Lin et al. (2022) | (2-) GFM | $\mathcal{O}(d^{1/2})$ | $\mathcal{O}(d^{3/2}\,\delta^{-1}\epsilon^{-4})$ |
| | Rando et al. (2023) | O-ZD | $\mathcal{O}(d^{1/2})$ | $\mathcal{O}(d^{3/2}\,\delta^{-1}\epsilon^{-3})$ |
| Stochastic | Kornowski & Shamir (2024) | O2NC | $\mathcal{O}(1)$ | $\mathcal{O}(d\,\delta^{-1}\epsilon^{-3})$ |
| | Lin et al. (2022) | (2-) SGFM | $\mathcal{O}(d^{1/2})$ | $\mathcal{O}(d^{3/2}\,\delta^{-1}\epsilon^{-4})$ |
| | Chen et al. (2023) | GFM+ | $\mathcal{O}(d^{1/2})$ | $\mathcal{O}(d^{3/2}\,\delta^{-1}\epsilon^{-3})$ |
| | **Ours** | ZO-SGDM-PR | **None** | $\mathcal{O}(d\,\delta^{-1}\epsilon^{-3})$ |

Table 1 highlights how this result compares with existing guarantees. The algorithmic $d$-factor denotes the additional multiplicative dimension dependence introduced by the optimization analysis beyond that inherent to zeroth-order estimation. Several earlier methods listed in the table incur an additional $\mathcal{O}(d^{1/2})$ factor beyond zeroth-order estimation, resulting in an $\mathcal{O}(d^{3/2})$ dimension dependence. Kornowski and Shamir (2024) already attain $\mathcal{O}(d\delta^{-1}\epsilon^{-3})$ through a smoothness-independent online-to-nonconvex conversion. Our result attains the same parameter dependence through a direct analysis of the SGDM recursion. The distinction concerns both the algorithmic design and the analysis. O2NC employs random line-segment interpolation and displacement projection, so that its favorable scaling $\mathcal{O}(1)$ is built into a specialized update rule. In contrast, our method retains the standard SGDM-style update within each epoch, yet our analysis avoids introducing a direct multiplicative dependence on the smoothness constant of the smoothed objective into the leading stationarity bound. Thus, rather than removing the additional dependence through a specially designed optimization procedure, we show that it can be avoided at the level of analysis even for classical momentum dynamics.

Our analysis also allows smoothing radii as small as $\nu = \Theta(d^{-1/2}\delta\epsilon)$, subject to $\nu \le \delta/2$ for transferring stationarity to the original objective. For smooth objectives, the same framework recovers $\mathcal{O}(d\epsilon^{-4})$ complexity for gradient-norm stationarity; when $\delta = \Theta(\epsilon)$, the nonsmooth Goldstein-stationarity bound has the same dependence on $d$ and $\epsilon$.

Empirical comparisons in Appendix E show that ZO-SGDM-PR remains close to standard ZO-SGDM in the evaluated settings and preserves its empirical behavior more faithfully than O2NC.

## 2 Related Work

In recent years, a major goal in zeroth-order (ZO) nonsmooth nonconvex optimization has been to derive increasingly tight complexity guarantees for producing Goldstein $(\delta, \epsilon)$-stationary points, with particular emphasis on the dependence on the dimension $d$. .

**The $\mathcal{O}(d^{3/2})$ barrier in ZO methods** Lin et al. (2022) studied gradient-free methods for nonsmooth nonconvex optimization in both deterministic and stochastic settings. Their proposed deterministic and stochastic algorithms find a $(\delta, \epsilon)$-stationary point using $\mathcal{O}\big(d^{3/2}\delta^{-1}\epsilon^{-4}\big)$ function-evaluation queries. In the deterministic setting, Rando et al. (2023) subsequently improved the dependence on $\epsilon$ from $\epsilon^{-4}$ to $\epsilon^{-3}$, while the dimension dependence remained $d^{3/2}$. In the stochastic setting, Chen et al. (2023) improved the $\epsilon$-dependence from $\epsilon^{-4}$ to $\epsilon^{-3}$ for nonsmooth nonconvex stochastic objectives, while still retaining the $d^{3/2}$ dimension scaling.

**Achieving $\mathcal{O}(d)$ via the framework of O2NC** Kornowski & Shamir (2024) identified a key source of the $\mathcal{O}(d^{3/2})$ scaling in conventional smoothness-based analyses, namely the interaction between dimension-dependent estimator variance ($\mathcal{O}(d)$) and the induced smoothness of the smoothed surrogate ($\mathcal{O}(\sqrt{d})$), and, inspired by the online-to-nonconvex conversion (O2NC) framework of Cutkosky et al. (2023) (which avoids explicit dependence on smoothness constants for nonsmooth

nonconvex objectives), they obtained the first rate of $\mathcal{O}\left(d\delta^{-1}\epsilon^{-3}\right)$, matching the best-known dependence on $d$, $\delta$, and $\epsilon$ separately. This result addresses the open problem posed by Lin et al. (2022) by establishing that, in the stochastic zeroth-order setting, nonsmooth nonconvex optimization admits algorithms with only linear dependence on the dimension $d$, matching the dimensional dependence known for smooth nonconvex optimization. For completeness, we present the core update rule of O2NC in Appendix B and compare it with the standard SGDM update to clarify their structural similarities and differences. In particular, O2NC relies on an online-to-nonconvex conversion and, in the projected-OGD instantiation considered here, additionally employs random line-segment interpolation and projection of the displacement variable.

**Notation** We write $\|\cdot\| := \|\cdot\|_2$ for the Euclidean norm and $\operatorname{dist}(x, S) := \inf_{s\in S}\|x - s\|$ for the distance from a point $x$ to a set $S$. For a set $V \subseteq \mathbb{R}^d$, we use $\sup_{v\in V}\|v - w\|$ and $\operatorname{dist}(0, \partial_V f)$, where $\partial_V f := \operatorname{conv}\big(\bigcup_{v\in V}\partial f(v)\big)$, $\operatorname{conv}(\cdot)$ denotes the convex hull. We use standard big-O notation, with $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ hiding absolute constants independent of problem parameters. We write $\operatorname{conv}(\cdot)$ for the convex hull and $\operatorname{Unif}(A)$ for the uniform measure on a set $A$, $\mathbb{B}_\delta(w) = \{x \in \mathbb{R}^d : \|x - w\|_2 \le \delta\}$ the (closed) $\delta$-ball centered at $w$. we say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is $G$-Lipschitz continuous if $|f(w) - f(w')| \le G\|w - w'\|$ for all $w, w' \in \mathbb{R}^d$; $f$ is $H$-smooth if it is differentiable and its gradient $\nabla f$ is $H$-Lipschitz continuous, i.e., $\|\nabla f(w) - \nabla f(w')\| \le H\|w - w'\|$ for all $w, w' \in \mathbb{R}^d$. For a nonsmooth function $f$, we denote its Clarke subdifferential at $w \in \mathbb{R}^d$ by $\partial f(w)$ (Clarke, 1990). Recall that $f$ is said to be $\rho$-weakly convex if the quadratically regularized function $f(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is convex, or equivalently, if $f(w) \ge f(w') + \langle g', w - w'\rangle - \frac{\rho}{2}\|w - w'\|^2$ for all $w, w' \in \mathbb{R}^d$ and $g' \in \partial f(w')$. We abbreviate $\mathbb{E}_{Z,U}[\cdot]$ as $\mathbb{E}[\cdot]$ when clear from context. For $a, b \ge 0$, $a \lesssim b$ (equivalently $b \gtrsim a$) means $a \le Cb$ for a universal constant $C > 0$.

## 3 PRELIMINARIES

In this section, we introduce the definitions, lemmas, and assumptions that will be used in the analysis of the next section.

For generic nonsmooth nonconvex functions as we are studying in this work, the Goldstein $(\delta, \epsilon)$-stationarity (Goldstein, 1977) is a standard criterion for convergence analysis, as defined below.

**Definition 1** ($(\delta, \epsilon)$-Stationarity)**.** *The Goldstein $\delta$-subdifferential of a Lipschitz function $f$ at a point $w \in \mathbb{R}^d$ is the convex hull of all Clarke subgradients at points in a $\delta$-ball around $w$, i.e.,*

$$\partial_\delta f(w) := \operatorname{conv}\left\{\cup_{v\in\mathbb{B}_\delta(w)}\partial f(v)\right\}.$$

*A point $w$ is called a $(\delta, \epsilon)$-stationary point if* $\operatorname{dist}\left(0, \partial_\delta f(w)\right) \le \epsilon$.

**$(\mu, \epsilon)$-Regularized Stationarity** We next introduce a linear-radius variant of the regularized stationarity notion of Ji & Yuan (2026), whose original formulation uses a squared-radius penalty. Our variant remains equivalent to Goldstein $(\delta, \epsilon)$-stationarity and is more amenable to our analysis. Given a subset $V \subseteq \mathbb{R}^d$, we denote $\partial_V f := \operatorname{conv}\{\cup_{v\in V}\partial f(v)\}$. Let us define

$$\|\partial f(w)\|_{+\mu} := \inf_{V\subseteq\mathbb{R}^d}\left\{\operatorname{dist}(0, \partial_V f) + \mu\sup_{v\in V}\|v - w\|\right\}.$$

So we have the following definition.

**Definition 2** ($(\mu, \epsilon)$-Regularized Stationarity)**.** *A point $w$ is said to be a $(\mu, \epsilon)$-regularized stationary point of a Lipschitz function $f$ if $\|\partial f(w)\|_{+\mu} \le \epsilon$.*

The notion of $(\mu, \epsilon)$-regularized stationarity enjoys several desirable properties that are useful for our proof. We refer readers to Section C for a brief introduction.

**Assumptions** Next, we make the following assumptions to ensure that the problem under study is tractable and well-defined.

**Assumption 1.** *For any $z \in \mathcal{Z}$, the loss function $\ell(\cdot; z)$ is $G$-Lipschitz with respect to its first argument.*

**Assumption 2** (Stochastic oracle)**.** *For each $w \in \mathbb{R}^d$, it holds that $g(w) = \mathbb{E}_{Z\sim\mathcal{D}}[\hat{g}(w;Z)] \in \partial f(w)$, where $\hat{g}(w;z) \in \partial\ell(w;z)$ for any $z \in \mathcal{Z}$.*

**Assumption 3** (Zeroth-order stochastic oracle)**.** *At each iteration $t$, the oracle draws $z_t \sim \mathcal{D}$ and a random direction $u_t \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and returns function values $\ell(v_t \pm \nu u_t; z_t)$, from which $\hat{g}_\nu(v_t; z_t)$ is computed as in* (5)*.*

**Randomized smoothing** Since we study zeroth-order algorithms, it is common to work with a smoothed surrogate objective, as it enjoys more desirable regularity properties. In this paper, we adopt the unit-ball randomized smoothing scheme.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be $G$-Lipschitz. For a smoothing radius $\nu > 0$, define the *unit-ball smoothing* of $f$ by

$$f_\nu(w) := \mathbb{E}_{U\sim\mathrm{Unif}(\mathbb{B}^d)}\big[f(w + \nu U)\big],$$

It is standard that the unit-ball smoothing inherits the Lipschitzness and becomes differentiable, with:

$$(i)\ f_\nu \text{ is also } G\text{-Lipschitz}; (ii)\ f_\nu \text{ is } \mathcal{O}\left(\frac{\sqrt{d}G}{\nu}\right)\text{-smooth}; (iii)\ |f(w) - f_\nu(w)| \le \nu G,\ \forall w \in \mathbb{R}^d.$$

These properties are proposed by Yousefian et al. (2012). The smoothed surrogate objective inherits the $L_0$-Lipschitz property of the original function and, moreover, becomes smooth. By the definition of smoothness, it follows immediately that the surrogate objective is $\rho$-weakly convex ($\rho = \frac{\sqrt{d}G}{\nu}$).

In what follows, we keep the notation $\rho$ explicit rather than substituting it by the smoothness constant of $f_\nu$, so as to highlight later why our method saves a $\sqrt{d}$ factor in the dimension dependence.

**Connection to the original problem** As mentioned earlier, although zeroth-order methods lack direct access to gradient information, they can naturally extract richer information than first-order methods via function evaluations. However, this advantage is meaningful only if we can relate convergence on the surrogate problem back to the original objective. Kornowski & Shamir (2024) established the following lemma to do so.[1]

**Lemma 1** (Kornowski & Shamir (2024), Lemma 4)**.** *If $x$ is a $(\delta, \epsilon)$-stationary point of $f_\nu$, then it is a $(\nu + \delta, \epsilon)$-stationary point of $f$. In particular, as long as $\nu + \delta \le \delta'$, it is a $(\delta', \epsilon)$-stationary point of $f$.*

**The second moment of the zeroth-order estimator** Also, the zeroth-order estimator introduces an explicit dependence on the dimension when constructing an approximate first-order oracle, for every $w \in \mathbb{R}^d$, $g_\nu(w) := \mathbb{E}[\hat{g}_\nu(w;Z)] = \nabla f_\nu(w)$. The following lemma (Shamir (2017)) provides a second-moment bound on the 2-norm of a standard zeroth-order gradient estimator with unit-sphere perturbations, showing that it depends only linearly on the dimension $d$.

**Lemma 2** (Shamir (2017), Lemma 10)**.** *For the two-point (antithetic) estimator in* (5)*, we have for every $w \in \mathbb{R}^d$,*

$$\mathbb{E}[\hat{g}_\nu(w;Z) \mid w] = g_\nu(w), \mathbb{E}\big[\|\hat{g}_\nu(w;Z)\|_2^2 \mid w\big] \le c\, d\, G^2,$$

*where $c$ is some numerical constant.*[2]

The above lemma indicates that dimension dependence of stochastic zeroth-order (ZO) optimization is largely governed by the second moment (variance) of the underlying gradient estimator.

## 4 Main results

In this section, we present the main theoretical results of this paper. We first introduce ZO-SGDM-PR, a zeroth-order SGDM-type method with periodic momentum resetting and epoch averaging.

[1] This lemma is instrumental in their dimension-improvement analysis, in this paper we use it only to transfer guarantees from the smoothed surrogate back to the original objective.

[2] In recent work, Lin et al. (2022) pointed out that $c = 16\sqrt{2\pi}$

**Algorithm 1** ZO-SGDM-PR

**Require:** Momentum $\beta \in (0, 1]$, stepsize $\eta > 0$, extrapolation $\gamma \in [0, 1]$, smoothing radius $\nu > 0$, inner length $K \in \mathbb{N}$, number of epochs $J \in \mathbb{N}$, initial point $w_0$.

$w_K^{(0)} \leftarrow w_0$.
**for** $j = 1, 2, \ldots, J$ **do**
  $w_0^{(j)} \leftarrow w_K^{(j-1)}$; $\quad m_0^{(j)} \leftarrow 0$.
  **for** $k = 1, 2, \ldots, K$ **do**
    $w_k^{(j)} \leftarrow w_{k-1}^{(j)} - \eta\, m_{k-1}^{(j)}$.
    $v_k^{(j)} \leftarrow w_k^{(j)} + \gamma\big(w_k^{(j)} - w_{k-1}^{(j)}\big)$.
    Sample $z_k^{(j)} \sim \mathcal{D}$.
    $g_k^{(j)} \leftarrow \text{ZOORACLE}\Big(v_k^{(j)}, z_k^{(j)}; \nu\Big)$
    $m_k^{(j)} \leftarrow (1-\beta) m_{k-1}^{(j)} + \beta\, g_k^{(j)}$.
  **end for**
  $\bar{w}^{(j)} \leftarrow \frac{1}{K} \sum_{k=1}^{K} w_k^{(j)}$.
**end for**
**Output:** $\bar{w} \sim \text{Unif}\big(\{\bar{w}^{(j)} : j \in [J]\}\big)$.

**Algorithm 2** ZOORACLE$(v, z; \nu)$

**Require:** Point $v \in \mathbb{R}^d$, sample $z$, smoothing radius $\nu > 0$.

Sample $u \sim \text{Unif}(\mathbb{S}^{d-1})$.
Query function values $\ell(v + \nu u; z)$ and $\ell(v - \nu u; z)$.
**Return** $\hat{g}_\nu(v; z) \leftarrow \frac{d}{2\nu}\big(\ell(v + \nu u; z) - \ell(v - \nu u; z)\big)\, u$.

We then establish its stationarity guarantee and show that it attains the best-known dimension dependence for stochastic nonsmooth nonconvex optimization. Finally, we provide further insight into the mechanism underlying the analysis and specialize the resulting guarantees to the smooth setting.

### 4.1 STATIONARITY ANALYSIS FOR ZO-SGDM-PR

We begin with the stationarity analysis of Algorithm 1, which applies the standard SGDM-style recursion within each epoch while resetting the momentum every $K$ iterations.

**Our algorithm** Algorithm 1 presents ZO-SGDM with periodic momentum resetting, where the gradient estimate $g_k^{(j)}$ is computed by the zeroth-order oracle $\text{ZOORACLE}(v_k^{(j)}, z_k^{(j)}; \nu)$. At the beginning of each epoch $j$, the momentum variable $m_0^{(j)}$ is reset to zero, which prevents the accumulation of stale momentum across different epochs. Within each epoch, the iterate $w_k^{(j)}$ is first updated using the previous momentum direction $m_{k-1}^{(j)}$, and an extrapolated point $v_k^{(j)}$ is then constructed to query the zeroth-order oracle. The oracle estimates a gradient of the smoothed stochastic loss by using a two-point finite-difference estimator $\hat{g}_\nu(v_k^{(j)}; z_k^{(j)})$ along a random direction $u \sim \text{Unif}(\mathbb{S}^{d-1})$. The resulting zeroth-order gradient estimate $g_k^{(j)}$ is then incorporated into the momentum recursion $m_k^{(j)} = (1-\beta)m_{k-1}^{(j)} + \beta g_k^{(j)}$. After each epoch, the algorithm forms the average of the inner iterates $\bar{w}^{(j)} = \frac{1}{K}\sum_{k=1}^{K} w_k^{(j)}$, and the final output $\bar{w}$ is sampled uniformly from these epoch-wise averages $\{\bar{w}^{(j)}\}_{j=1}^{J}$ to get averaged-iterate convergence.

The inner-loop dynamics retain the standard exponentially weighted SGDM recursion, while periodic resetting and epoch averaging are introduced as explicit outer-loop mechanisms in our theoretical construction. The role of momentum in our analysis is to provide a recursively averaged zeroth-order gradient estimate whose energy can be controlled over each epoch. This leads to a momentum-energy inequality that couples the cumulative step lengths with the decrease of the smoothed objective and the second moment of the zeroth-order estimator. Such a relation is the key ingredient that allows us to isolate the explicit dimension dependence to the zeroth-order estimator.

The following lemma formalizes this momentum-energy control and serves as the main intermediate result for Theorem 1.

**Lemma 3.** *Suppose that Assumption 1 and Assumption 3 hold. Assume that $\eta \le \frac{1}{3\rho}$ and $\gamma \le 1$. Then it holds that*

$$\mathbb{E}\Big[\beta \sum_{t=1}^{T} \|m_{t-1}\|^2 + \|m_T\|^2\Big] \le \mathbb{E}\Big[\|m_0\|^2 + \frac{2\beta(1-\beta)}{\eta}\Big(f_\nu(w_0) - f_\nu(w_T)\Big)\Big] + c\,d\,\beta^2 G^2\,T.$$

*Proof sketch.* Expanding the momentum recursion and using $w_t = w_{t-1} - \eta m_{t-1}$, the cross term can be expressed through $\langle w_{t-1} - w_t, g_\nu(v_t)\rangle$ after taking conditional expectation. The $\rho$-weak convexity of $f_\nu$ bounds this term by the objective decrease $f_\nu(w_{t-1}) - f_\nu(w_t)$ plus a quadratic term in $\|w_t - w_{t-1}\|$, which is absorbed under $\eta \le 1/(3\rho)$ and $\gamma \le 1$. Summing the resulting inequality over $t$ yields the claimed bound. The complete proof is given in Section D.1. □

This inequality provides the key control on the cumulative momentum magnitude used in the subsequent stationarity analysis.

For any $j \in [J]$, for the gradient of the smoothed objective, we define $\bar{g}_\nu^{(j)} := \frac{1}{1-(1-\beta)^K} \sum_{k=1}^{K} (1-\beta)^{K-k} \beta\, g_\nu(v_k^{(j)})$. This definition is motivated by the momentum recursion: the end-of-epoch momentum corresponds (in expectation) to an exponentially weighted average of smoothed gradients. The normalization makes the weights sum to one, so $\bar{g}_\nu^{(j)}$ is a convex combination that is convenient for our subsequent stationarity analysis.

Moreover, we assume $f^* := \inf_{w\in\mathbb{R}^d} f(w) > -\infty$. and denote $\Delta f_0 := f(w_0) - f^*$, since $\|f - f_\nu\|_\infty \le \nu G$, it follows that $f_\nu(w_0) - f_\nu^* \le \Delta f_0 + \nu G =: \Delta f_{\nu,0}$. We initialize $m_0 = 0$.

**Theorem 1.** *Suppose that Assumption 1 and Assumption 3 hold. Assume that $\eta \le \frac{1}{3\rho}$ and $\gamma \le 1$. Then for any $\mu \le \frac{1}{\eta(1-(1-\beta)^K)}\sqrt{\frac{\beta}{K+1}}$, the sequence $\{\bar{w}^{(j)}\}_{j=1}^{J}$ generated by Algorithm 1 satisfies*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} \left\|\partial f_\nu(\bar{w}^{(j)})\right\|_{+\mu}^2\right] \le \frac{8\beta(1-\beta)\Delta f_{\nu,0}}{\eta\,(1-(1-\beta)^K)^2\,J} + \frac{4c\beta d G^2\,(\beta K+1)}{(1-(1-\beta)^K)^2}.$$

*Proof sketch.* Fix an epoch $j$. Applying Lemma 3 with $m_0^{(j)} = 0$ and $w_0^{(j+1)} = w_K^{(j)}$ gives a one-epoch bound on $\beta\sum_{k=1}^{K}\|m_{k-1}^{(j)}\|^2 + \|m_K^{(j)}\|^2$ by the decrease $f_\nu(w_0^{(j)}) - f_\nu(w_0^{(j+1)})$ and the accumulated ZO variance. The first term controls the spread of the extrapolation points around $\bar{w}^{(j)}$, while the terminal momentum $m_K^{(j)}$ controls the weighted average gradient $\bar{g}_\nu^{(j)}$, up to the same variance error. Averaging over $j = 1, \dots, J$ telescopes the objective decrease and yields a bound on the average of $\|\bar{g}_\nu^{(j)}\|^2$ plus the deviation term $\max_k \|v_k^{(j)} - \bar{w}^{(j)}\|^2$. Finally, since $\bar{g}_\nu^{(j)} \in \partial_{V^{(j)}} f_\nu$ and $\mu \le \mu'$, Lemma 6 converts this bound into the desired estimate for $\frac{1}{J}\sum_{j=1}^{J}\|\partial f_\nu(\bar{w}^{(j)})\|_{+\mu}^2$. Details can be found in Section D.3. □

**Remark** Theorem 1 reveals the key mechanism behind our analysis. The objective-decrease term controls the cumulative momentum energy and hence the within-epoch spread, while the terminal momentum controls an exponentially weighted average of zeroth-order gradients. The resulting bound separates the optimization error from the zeroth-order estimation error: the former is dimension-free, whereas the latter contributes the factor $\mathcal{O}(d)$ through $\mathbb{E}\|\hat{g}_\nu\|^2 = \mathcal{O}(dG^2)$. In particular, although the weak-convexity parameter $\rho$ may itself depend on $d$ after randomized smoothing, it affects only the admissible stepsize and does not directly appear in the stationarity bound.

We can further establish the following result on the complexity of Algorithm 1 for producing $(\mu, \epsilon)$-regularized stationary points.

**Corollary 1.** *Suppose that Assumption 1 and Assumption 3 hold. Assume that $\gamma, \nu \le 1$. Let $\mu, \epsilon > 0$ be the desired regularized-stationarity parameters and $T$ be the total budget of iterates. Set*

$$K = \left\lceil \frac{d^{1/2}T^{1/2}}{\mu^{1/2}} \right\rceil, J = \left\lceil \frac{T}{K} \right\rceil, \beta = \frac{4}{K+1}, \eta = \frac{2}{(K+1)\mu}.$$

*Recall* $f_\nu(w_0) - f_\nu^* \le \Delta f_0 + \nu G =: \Delta f_{\nu,0} \le \Delta f_0 + G$. *Suppose that*

$$T \gtrsim \frac{d\mu(\Delta f_0 + G + G^2)^2}{\epsilon^4} + \frac{\rho^2}{d\mu}.$$

*Then the output* $\bar{w}$ *by Algorithm 1 satisfies* $\mathbb{E}\left[\|\partial f_\nu(\bar{w})\|_{+\mu}\right] \le \epsilon$.

*Proof.* The proof is deferred to Section D.4. □

**Remark** By setting $\mu = \epsilon/\delta$ and $\rho = \mathcal{O}(\frac{\sqrt{d}G}{\nu})$, Corollary 1 yields $\mathcal{O}\big(\frac{d}{\delta\epsilon^3} + \frac{\delta}{\nu^2\epsilon}\big)$ iterations, up to problem-dependent constants, for obtaining the corresponding Goldstein-stationarity guarantee for the smoothed objective $f_\nu$. Therefore, the first term gives the desired $\mathcal{O}(d\delta^{-1}\epsilon^{-3})$ dependence, while the second term remains lower order provided $\nu = \Omega\Big(\frac{\delta\epsilon}{\sqrt{d}}\Big)$. Finally, Lemma 1 transfers the guarantee from $f_\nu$ to the original objective $f$. In particular, by running the above argument with an internal radius $\delta/2$ and choosing $\nu \le \delta/2$, we obtain the desired $(\delta, \epsilon)$-stationarity guarantee for $f$.

We next make the parameter scaling more explicit. For simplicity, consider the commonly used regime $\delta = \epsilon$. Since the dominant complexity becomes $T = \mathcal{O}(d\epsilon^{-4})$, we have $\epsilon = \delta = \Theta\Big(\big(\frac{d}{T}\big)^{1/4}\Big)$ and hence $\mu = \frac{\epsilon}{\delta} = 1$. Ignoring problem-dependent constants, our parameter choices therefore scale as $K = \Theta\Big(\sqrt{dT}\Big), J = \Theta\Big(\sqrt{\frac{T}{d}}\Big), \beta = \Theta\Big(\frac{1}{\sqrt{dT}}\Big), \eta = \Theta\Big(\frac{1}{\sqrt{dT}}\Big)$, while the smoothing radius only needs to satisfy $\nu = \Omega\Big(\frac{1}{\sqrt{T}}\Big)$.

### 4.2 First-order counterpart

A useful way to understand the dimension dependence of our result is to replace the zeroth-order oracle in Algorithm 1 with a first-order stochastic oracle. Specifically, suppose that $\hat{g}\Big(v_k^{(j)}; z_k^{(j)}\Big) \in \partial\ell\Big(v_k^{(j)}; z_k^{(j)}\Big)$ is used in place of the zeroth-order estimator. The resulting method is the first-order counterpart of ZO-SGDM-PR applied to a weakly convex and Lipschitz objective.

The analysis remains unchanged except for the second-moment term of the stochastic oracle. In the zeroth-order setting, this term scales as $\mathbb{E}\|\hat{g}_\nu(w; Z)\|^2 \le cdG^2$, whereas for the first-order oracle it is bounded by $\mathbb{E}\|\hat{g}(w; Z)\|^2 \le G^2$. Consequently, the explicit dimension dependence disappears from the analysis, leading to the following corollary.

**Corollary 2.** *Suppose that Assumption 1 and Assumption 2 hold. Assume that* $\gamma \le 1$. *Let* $\mu, \epsilon > 0$ *be the desired regularized-stationarity parameters and* $T$ *be the total budget of iterates. Set*

$$K = \left\lceil \frac{T^{1/2}}{\mu^{1/2}} \right\rceil, J = \left\lceil \frac{T}{K} \right\rceil, \ \beta = \frac{4}{K+1}, \ \eta = \frac{2}{(K+1)\mu}.$$

*Suppose that*

$$T \gtrsim \frac{(G^2 + \Delta f_0)^2\mu}{\epsilon^4} + \frac{\rho^2}{\mu}.$$

*Then the output* $\bar{w}$ *by using first-order oracle in Algorithm 1 satisfies* $\mathbb{E}\left[\|\partial f(\bar{w})\|_{+\mu}\right] \le \epsilon$.

*Proof.* The proof is deferred to Section D.5. □

**Remark** Setting $\mu = \epsilon/\delta$ and applying Lemma 5, Corollary 2 yields $\mathcal{O}\Big(\frac{1}{\delta\epsilon^3} + \frac{\rho^2\delta}{\epsilon}\Big)$ iterations for finding a $(\delta, \epsilon)$-stationary point. This recovers the best-known first-order dependence for weakly convex nonsmooth optimization when the weak-convexity term does not dominate.

More importantly, the first-order counterpart serves to isolate the source of the dimension dependence in our zeroth-order result. A related periodically restarted first-order scheme that is closely related to restarted SGDM dynamics was studied by Ji & Yuan (2026). In our analysis, replacing the

zeroth-order oracle by a first-order oracle removes the $O(d)$ second-moment factor without otherwise changing the argument. This shows that the dimension dependence in Corollary 1 comes solely from the zeroth-order estimator, rather than from the SGDM-PR analysis itself.

Interestingly, although both the first- and zeroth-order analyses involve a weak-convexity parameter, the latter becomes dimension-dependent after randomized smoothing. Nevertheless, this dependence does not translate into an additional dimension factor in the final complexity bound.

### 4.3 ZO-SGDM-PR IN THE SMOOTH CASE

We next specialize our result to smooth nonconvex objectives. In the stochastic zeroth-order setting, the standard complexity for finding an $\epsilon$-stationary point of a smooth nonconvex objective is $\mathcal{O}(d\epsilon^{-4})$, up to problem-dependent constants (Ghadimi & Lan, 2013). We show that ZO-SGDM-PR recovers the same dependence on $d$ and $\epsilon$.

The following result of Zhang et al. (2020) connects Goldstein stationarity with the usual gradient-based stationarity notion under smoothness.

**Lemma 4** (Zhang et al. (2020), Proposition 6)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and let $\epsilon > 0$.*

- *If $\|\nabla f(x)\| \le \epsilon$, then $x$ is a $(\delta, \epsilon)$-stationary point for every $\delta \ge 0$.*
- *Suppose in addition that $f$ is $H$-smooth. If $x$ is a $\left(\frac{\epsilon}{3H}, \frac{\epsilon}{3}\right)$-stationary point, then $\|\nabla f(x)\| \le \epsilon$.*

Since $f$ is $H$-smooth, its randomized smoothing $f_\nu$ is also $H$-smooth and hence $H$-weakly convex. Therefore, Corollary 1 applies with $\rho = H$.

**Corollary 3.** *Suppose that $f$ is $H$-smooth, and conditions in Corollary 1 hold. By taking $\mu = 2H$, $\nu = \frac{\epsilon}{6H}$, and applying Corollary 1 with target accuracy $\frac{\epsilon}{3}$, if*

$$T \gtrsim \frac{d\,H\,(\Delta f_0 + G + G^2)^2}{\epsilon^4} + \frac{H}{d},$$

*then it suffices to run for $T = \mathcal{O}\big(d\,\epsilon^{-4}\big)$ iterations, up to problem-dependent constants, to obtain the corresponding $\epsilon$-stationarity guarantee.*

*Proof.* Since $f_\nu$ is $H$-smooth, we may set $\rho = H$ in Corollary 1. Taking $\mu = 2H$ and the target regularized-stationarity accuracy as $\epsilon/3$, we obtain the corresponding Goldstein radius $\frac{\epsilon/3}{2H} = \frac{\epsilon}{6H}$. By further taking $\nu = \frac{\epsilon}{6H}$, Lemma 1 gives the Goldstein radius $\frac{\epsilon}{6H} + \frac{\epsilon}{6H} = \frac{\epsilon}{3H}$ for the original objective $f$. Therefore, Lemma 4 yields the desired $\epsilon$-stationarity guarantee. Substituting $\mu = 2H$ and $\rho = H$ into Corollary 1 gives $T = \mathcal{O}\big(d\,\epsilon^{-4}\big)$ up to problem-dependent constants. □

Thus, ZO-SGDM-PR attains the same leading dependence on $d$ and $\epsilon$ in the nonsmooth and smooth settings, namely $\mathcal{O}(d\epsilon^{-4})$ when $\delta$ is chosen on the order of $\epsilon$. Corollary 3 implies that, in the stochastic zeroth-order setting, nonsmooth nonconvex optimization is essentially as easy as smooth nonconvex optimization, the result especially from our algorithm in this paper. The experiments in Appendix E further illustrate similar empirical convergence behavior in the two settings.

## 5 CONCLUSION

In this paper, we develop a new analysis based on an equivalent formulation of Goldstein $(\delta, \epsilon)$-stationarity and establish an $\mathcal{O}!\left(d\delta^{-1}\epsilon^{-3}\right)$ oracle complexity for ZO-SGDM-PR. This matches the best-known dependence on $d$, $\delta$, and $\epsilon$ for stochastic zeroth-order nonsmooth nonconvex optimization. Our analysis isolates the explicit dimension dependence to the second moment of the zeroth-order estimator, while avoiding the additional dimension factor introduced by conventional smoothness-based analyses. The dependence on $d$, $\delta$, and $\epsilon$ is known to be optimal when considered separately, while a matching joint lower bound remains open.

For future work, it would be interesting to establish the joint information-theoretic optimality of this rate, to obtain the same dimension dependence for plain ZO-SGD or unmodified ZO-SGDM without periodic resetting and epoch averaging, and to extend the analysis to other momentum-based zeroth-order methods such as ZO-AdaMM (Chen et al., 2019).

### AI USE STATEMENT

Generative AI tools were used at several stages of manuscript preparation. For drafting, they assisted with restructuring selected passages of the motivation, related work, descriptions of theoretical results, and experimental setup. For writing assistance, they were used for language polishing, improving clarity and conciseness, content condensation, and formatting. For technical checking, they were used to identify potential errors in proofs and to review the manuscript for completeness and logical consistency; all mathematical arguments and conclusions were independently checked by the authors. For retrieval and discovery, they were used to suggest potentially relevant related work and references beyond the primary comparison methods, which were subsequently examined and verified by the authors. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final manuscript.

### ETHICS STATEMENT

This work does not involve human participants or personal data.

### REPRODUCIBILITY STATEMENT

The algorithms, assumptions, parameter choices, and complete proofs are provided in the main text and appendices. The experimental environments, baselines, hyperparameter selection, and evaluation protocols are also described in detail in the appendix.

## REFERENCES

Lesi Chen, Jing Xu, and Luo Luo. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. In *International Conference on Machine Learning*, pp. 5219–5233. PMLR, 2023.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. ZO-AdaMM: Zeroth-order adaptive momentum method for black-box optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yiming Chen, Yuan Zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order finetuning for language models with low-rank structures. *arXiv preprint arXiv:2410.07698*, 2024.

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

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning (ICML)*, pp. 6643–6670. PMLR, 2023.

Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *Advances in neural information processing systems*, 35:6692–6703, 2022.

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

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

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

Eduard Gorbunov, Adel Bibi, Ozan Sener, El Houcine Bergou, and Peter Richtárik. A stochastic derivative free optimization method with momentum. *arXiv preprint arXiv:1905.13278*, 2019.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pp. 2137–2146. PMLR, 2018.

Fanfan Ji and Xiao-Tong Yuan. Derandomized online-to-non-convex conversion for stochastic weakly convex optimization. *International Conference on Learning Representations*, 2026.

Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International conference on machine learning*, pp. 3100–3109. PMLR, 2019.

Michael Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *Conference on Learning Theory*, pp. 4570–4597. PMLR, 2023.

Siyu Kong and Adrian S Lewis. Lipschitz minimization and the goldstein modulus. *Mathematical Programming*, pp. 1–30, 2025.

Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *Journal of Machine Learning Research*, 25(122): 1–14, 2024.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

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

Tianyi Lin, Zeyu Zheng, and Michael Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 35: 26160–26175, 2022.

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

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

Egor Petrov, Grigoriy Evseev, Aleksey Antonov, Andrey Veprikov, Nikolay Bushkov, Stanislav Moiseev, and Aleksandr Beznosikov. Leveraging coordinate momentum in signsgd and muon: Memory-optimized zero-order. *arXiv preprint arXiv:2506.04430*, 2025.

Marco Rando, Cesare Molinari, Lorenzo Rosasco, and Silvia Villa. An optimal structured zeroth-order algorithm for non-smooth optimization. *Advances in Neural Information Processing Systems*, 36:36738–36767, 2023.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *Mathematical Programming*, 208(1):51–74, 2024.

Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pp. 11173–11182. PMLR, 2020.

Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D Lee, Wotao Yin, Mingyi Hong, et al. Revisiting zeroth-order optimization for memory-efficient llm fine-tuning: A benchmark. *arXiv preprint arXiv:2402.11592*, 2024.

# A FURTHER RELATED WORK

In this section, we review developments in zeroth-order optimization beyond the nonsmooth nonconvex setting, as well as theoretical and applied progress on ZO-SGDM.

**Progress in zeroth-order optimization** Zeroth-order optimization has been extensively studied in several settings beyond the stochastic nonsmooth nonconvex regime considered in this paper. For nonsmooth convex optimization, the optimal zeroth-order oracle complexity is $\Theta(d\epsilon^{-2})$, with the upper bound established by Shamir (2017) and the matching lower bound shown by Duchi et al. (2015). In the smooth nonconvex regime, zeroth-order methods can achieve an oracle complexity of $\mathcal{O}(d\epsilon^{-2})$ when exact function values are available (Nesterov & Spokoiny, 2017). Under noisy function-value oracles, however, the complexity worsens to $\mathcal{O}(d\epsilon^{-4})$ (Ghadimi & Lan, 2013). Under additional structural assumptions, such as the finite-sum setting, variance-reduction techniques can be leveraged (Liu et al., 2018; Ji et al., 2019; Chen et al., 2019), leading to the complexity of $\mathcal{O}(d^{3/4}\epsilon^{-3})$.

**Theory and applications of ZO-SGDM-Type Methods** Regarding zeroth-order momentum methods' theoretical results, Gorbunov et al. (2019) studied derivative-free heavy-ball momentum and proposed a stochastic three-point algorithm, establishing a series of convergence results under smoothness assumptions. Recently, Petrov et al. (2025) studied a sign-based zeroth-order momentum method and provided a convergence analysis for ZO-SignSGD with momentum in the stochastic nonconvex setting. As for ZO-SGDM-type algorithms in applications, one representative work, studied by Zhang et al. (2024), systematically benchmarked multiple zeroth-order optimizers, including ZO-SGDM . Their empirical results suggest that ZO-SGDM can provide improved stability in LLM fine-tuning.

# B COMPARISON BETWEEN O2NC AND SGDM

For completeness, we present the core update rule of O2NC (Kornowski & Shamir, 2024) and compare it with the standard SGDM update. For a direct comparison, we consider the O2NC instantiation with projected online gradient descent (OGD) as the online learner.

**Algorithm 3** O2NC with projected OGD: core update

**Require:** Stepsize $\eta > 0$, displacement radius $D > 0$, smoothing radius $\nu > 0$, initial point $x_0$.
1: Initialize $\Delta_1 \leftarrow 0$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad x_t \leftarrow x_{t-1} + \Delta_t$.
4: $\quad$ Sample $s_t \sim \mathrm{Unif}[0, 1]$ and set $v_t \leftarrow x_{t-1} + s_t\Delta_t$.
5: $\quad$ Sample $z_t \sim \mathcal{D}$ and compute $g_t \leftarrow \mathrm{ZOORACLE}(v_t, z_t; \nu)$.
6: $\quad \Delta_{t+1} \leftarrow \Pi_{\mathbb{B}(0,D)}(\Delta_t - \eta g_t)$.
7: **end for**

For comparison, the standard ZO-SGDM update with $\gamma = 0$ can be written as

**Algorithm 4** Standard ZO-SGDM: core update

**Require:** Momentum parameter $\beta \in (0, 1]$, stepsize $\eta > 0$, smoothing radius $\nu > 0$, initial point $w_0$.
1: Initialize $m_0 \leftarrow 0$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad w_t \leftarrow w_{t-1} - \eta m_{t-1}$.
4: $\quad$ Sample $z_t \sim \mathcal{D}$ and compute $g_t \leftarrow \mathrm{ZOORACLE}(w_t, z_t; \nu)$.
5: $\quad m_t \leftarrow (1 - \beta)m_{t-1} + \beta g_t$.
6: **end for**

The two updates share a momentum-like structure in the sense that both maintain an auxiliary state that aggregates past stochastic zeroth-order gradients and uses this state to generate the next iterate. In particular, if the projection in Algorithm 3 is removed and $s_t = 1$, the O2NC displacement satisfies

$$\Delta_{t+1} = \Delta_t - \eta g_t,$$

which resembles a heavy-ball-type update with unit momentum.

There are, however, several structural differences. First, O2NC evaluates the stochastic gradient estimator at a random point

$$v_t = x_{t-1} + s_t \Delta_t$$

along the line segment between two consecutive iterates, whereas standard ZO-SGDM evaluates it at the current iterate. This random interpolation is central to the O2NC conversion analysis. Second, O2NC explicitly projects the displacement $\Delta_t$ onto a bounded ball, while standard SGDM does not impose such a projection on its momentum variable. Third, O2NC accumulates the previous displacement with coefficient one before projection, whereas SGDM exponentially discounts past gradients through the factor $1 - \beta$.

Our ZO-SGDM-PR preserves the standard SGDM recursion within each epoch and introduces periodic momentum resetting and epoch averaging only at the outer-loop level. Therefore, although O2NC and ZO-SGDM-PR both admit momentum-like interpretations, their update mechanisms and analyses are different: O2NC relies on random line-segment interpolation and displacement projection together with the online-to-nonconvex conversion, whereas our analysis is based on the exponentially weighted SGDM momentum recursion and its associated momentum-energy inequality.

## C SOME USEFUL LEMMAS

In this section, we establish several basic properties of the regularized stationarity measure used in this paper. Our definition is motivated by the regularized stationarity notion of Ji & Yuan (2026), but differs in that we use a linear radius penalty, whereas their formulation uses a squared-radius penalty. We show that this linear-radius variant remains closely related to Goldstein stationarity and satisfies a useful monotonicity property.

The following lemma establishes the connection between $(\mu, \epsilon)$-regularized stationarity and Goldstein $(\delta, \epsilon)$-stationarity.

**Lemma 5.** *Let $\delta, \epsilon, \mu > 0$ be arbitrary positive values. Consider a Lipschitz function $f$.*

*(a) If $w$ is a $(\delta, \epsilon)$-stationary point, then it is also a $\left(\frac{\epsilon}{\delta}, 2\epsilon\right)$-regularized stationary point.*

*(b) If $w$ is a $(\mu, \epsilon)$-regularized stationary point, then it is also a $\left(\frac{\epsilon}{\mu}, \epsilon\right)$-stationary point.*

As a consequence, controlling the regularized stationarity measure $\|\partial f(w)\|_{+\mu}$ is sufficient for deriving Goldstein-type guarantees for the original objective, which is the perspective we adopt in the sequel.

The $(\mu, \epsilon)$-regularized stationarity measure enjoys favorable properties. In particular, we have the following lemma, which shows that $\|\partial f(w)\|_{+\mu}$ is monotone with respect to the regularization strength $\mu$.

*Proof.* Part(a): Let $w$ be a Goldstein $(\delta, \epsilon)$-stationary point of $f$. Consider $\mu = \frac{\epsilon}{\delta}$. Then it follows from Definition 2 that

$$\begin{aligned}
\|\partial f(w)\|_{+\mu} \leq & \operatorname{dist}(0, \partial_{\mathbb{B}_\delta(w)} f) + \mu \sup_{v \in \mathbb{B}_\delta(w)} \|v - w\| \\
= & \operatorname{dist}(0, \partial_\delta f(w)) + \mu \sup_{v \in \mathbb{B}_\delta(w)} \|v - w\| \\
\leq & \epsilon + \mu\delta = 2\epsilon,
\end{aligned}$$

where in the second inequality we have used the definition of Goldstein $(\delta, \epsilon)$-stationarity. Then by definition $w$ must be a $\left(\frac{\epsilon}{\delta}, 2\epsilon\right)$-regularized stationary point of $f$.

Part(b): Let us now consider the case that $w$ is a $(\mu, \epsilon)$-regularized stationary point of $f$. Let $\delta = \frac{\epsilon}{\mu}$ and $\varepsilon > 0$ be arbitrary. Since $\|\partial f(w)\|_{+\mu} \leq \epsilon$, it follows from Definition 2 that there exists some $V^*(\varepsilon)$ such that

$$\epsilon \geq \|\partial f(w)\|_{+\mu} \geq \operatorname{dist}(0, \partial_{V^*(\varepsilon)} f) + \mu \sup_{v \in V^*(\varepsilon)} \|v - w\| - \varepsilon,$$

which then directly implies

$$\operatorname{dist}(0, \partial_{V^*(\varepsilon)} f) \leq \epsilon + \varepsilon, \quad \sup_{v \in V^*(\varepsilon)} \|v - w\| \leq \frac{\epsilon + \varepsilon}{\mu} = \delta + \frac{\varepsilon}{\mu}.$$

The second inequality in the above implies that $V^*(\varepsilon) \subseteq \mathbb{B}_{\delta + \frac{\varepsilon}{\mu}}(w)$ and thus $\partial_{V^*(\varepsilon)} f \subseteq \partial_{\delta + \frac{\varepsilon}{\mu}} f$. Then we have

$$\operatorname{dist}\left(0, \partial_{\delta + \frac{\varepsilon}{\mu}} f\right) \leq \operatorname{dist}\left(0, \partial_{V^*(\varepsilon)} f\right) \leq \epsilon + \varepsilon.$$

Since $\varepsilon$ is allowed to be arbitrarily small and recall that $\delta = \frac{\epsilon}{\mu}$, the above inequality implies that $w$ deems a Goldstein $\left(\frac{\epsilon}{\mu}, \epsilon\right)$-stationary point. □

**Lemma 6.** *Let $f$ be a Lipschitz function. Then for any $w \in \mathbb{R}^d$ and $0 < \mu_1 \leq \mu_2$, it holds that $\|\partial f(w)\|_{+\mu_1} \leq \|\partial f(w)\|_{+\mu_2}$.*

*Proof.* Consider a fixed vector $w$. Let $\varepsilon > 0$ be arbitrary. By definition we know that there exists a subset $V_2^*(\varepsilon) \subseteq \mathbb{R}^d$ such that $\|\partial f(w)\|_{+\mu_2} \geq \operatorname{dist}(0, \partial_{V_2^*(\varepsilon)} f) + \mu_2 \sup_{v \in V_2^*(\varepsilon)} \|v - w\| - \varepsilon$. Again, by definition and the condition $\mu_1 \leq \mu_2$ we can see that

$$\begin{aligned}
\|\partial f(w)\|_{+\mu_1} &\leq \operatorname{dist}(0, \partial_{V_2^*(\varepsilon)} f) + \mu_1 \sup_{v \in V_2^*(\varepsilon)} \|v - w\| \\
&\leq \operatorname{dist}(0, \partial_{V_2^*(\varepsilon)} f) + \mu_2 \sup_{v \in V_2^*(\varepsilon)} \|v - w\| \\
&\leq \|\partial f(w)\|_{+\mu_2} + \varepsilon.
\end{aligned}$$

By noting that $\varepsilon$ can be arbitrarily small, we must have $\|\partial f(w)\|_{+\mu_1} \leq \|\partial f(w)\|_{+\mu_2}$. □

As in the squared-radius formulation of Ji & Yuan (2026), our linear-radius measure is monotone in $\mu$.

**Remark** Lemma 6 states that the regularized stationarity measure $\|\partial f(w)\|_{+\mu}$ is monotone non-decreasing in the regularization strength $\mu$. Intuitively, increasing $\mu$ places a larger penalty on the "radius" term $\sup_{v \in V} \|v - w\|$ in the definition of $\|\partial f(w)\|_{+\mu}$, making the criterion more conservative and thus yielding a larger value.

Directly constructing a suitable $\mu$ can be difficult, but the above lemma allows us to upper bound the objective by a family of expressions that conform to the $(\mu, \epsilon)$-regularized stationarity structure, thereby converting the original challenging construction problem into an existence problem.

For naturally weakly convex functions in this paper, the following lemma is also useful.

**Lemma 7.** *If $f$ is $\rho$-weakly convex, then for any pair of $w, v \in \mathbb{R}^d$ it holds that*

$$\langle w - v, g(w) - g(v) \rangle \geq -\rho \|w - v\|^2,$$

*where $g(w) \in \partial f(w)$, $g(v) \in \partial f(v)$.*

## D Convergence analysis

In this section, we give our proofs for the main results.

To reduce the dimension dependence, a simple idea is to run multiple queries and average them. For example, Kornowski & Shamir (2024) average over $2q$ oracle queries to form a gradient estimator.[3]

[3] As noted in their discussion, the idea of averaging multiple gradient estimators to mitigate the dimension dependence in the complexity can also be applied to smooth or convex objectives. However, even in these more benign settings, the resulting overall complexity remains substantially worse, most notably in its dependence on $\epsilon$, than that of optimal deterministic first-order methods.

In a similar spirit, we adopt periodic momentum resetting every $K$ iterations. The following simple lemma about averaging is useful in our analysis.

**Lemma 8.** *Let $w_1, w_2, \ldots, w_K$ be a set of vectors and $\bar{w} = \frac{1}{K}\sum_{k=1}^{K} w_k$. Then the following holds for all $k \in [K]$:*

$$\|w_k - \bar{w}\|^2 \le \frac{1}{K}\sum_{j=1}^{K}\|w_k - w_j\|^2 \le K\sum_{j=1}^{K}\|\Delta_j\|^2,$$

*where $\Delta_j := w_j - w_{j-1}$ and $w_0 = w_1$. Moreover, let $v_k = w_k + \gamma(w_k - w_{k-1})$ for some $\gamma \in [0,1]$. Then for all $k \in [K]$, it holds that*

$$\|v_k - \bar{w}\|^2 \le 2(K+1)\sum_{j=1}^{K}\|\Delta_j\|^2.$$

**Remark** Lemma 8 relates the dispersion of iterates within an epoch to the cumulative squared increments. In particular, it shows that the deviation of any iterate $w_k$ (and the extrapolated point $v_k$) from the epoch average $\bar{w}$ can be controlled by $\sum_{j=1}^{K}\|\Delta_j\|^2$, where $\Delta_j = w_j - w_{j-1}$. Since in our algorithm $\Delta_j = -\eta m_{j-1}$, this bound converts control of the momentum energy (cf. Lemma 3) into control of the within-epoch spread $\max_{k\in[K]}\|v_k - \bar{w}\|^2$. This connection is crucial for establishing $(\mu,\epsilon)$-regularized stationarity of the averaged iterate.

### D.1 Proof for Lemma 3

We consider the filtration $\mathcal{F}_t := \sigma\Big(\{z_s\}_{s=1}^{t}, \{u_s\}_{s=1}^{t}\Big)$, where $z_t \sim \mathcal{D}$ and $u_t \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ are freshly sampled at each iteration. Note that $w_t$ and hence $v_t = w_t + \gamma(w_t - w_{t-1})$ are $\mathcal{F}_{t-1}$-measurable.

*Proof.* Expanding the momentum recursion gives

$$\begin{aligned}
\|m_t\|^2 &= \|(1-\beta)m_{t-1} + \beta\hat{g}_\nu(v_t; z_t)\|^2 \\
&= (1-\beta)^2\|m_{t-1}\|^2 + 2\beta(1-\beta)\langle m_{t-1}, \hat{g}_\nu(v_t; z_t)\rangle + \beta^2\|\hat{g}_\nu(v_t; z_t)\|^2 \\
&= (1-\beta)^2\|m_{t-1}\|^2 + \frac{2\beta(1-\beta)}{\eta}\langle w_{t-1} - w_t, \hat{g}_\nu(v_t; z_t)\rangle + \beta^2\|\hat{g}_\nu(v_t; z_t)\|^2.
\end{aligned}$$

The last step comes from (2). Taking conditional expectation, conditioned on $\mathcal{F}_{t-1}$, on both sides yields

$$
\begin{aligned}
&\mathbb{E}\left[\|m_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
&\leq (1-\beta)^2\|m_{t-1}\|^2 + \frac{2\beta(1-\beta)}{\eta}\Big\langle w_{t-1}-w_t,\, \mathbb{E}[\hat{g}_\nu(v_t;z_t) \mid \mathcal{F}_{t-1}]\Big\rangle + \beta^2\mathbb{E}\big[\|\hat{g}_\nu(v_t;z_t)\|^2 \mid \mathcal{F}_{t-1}\big] \\
&\overset{\zeta_1}{\leq} (1-\beta)^2\|m_{t-1}\|^2 + \frac{2\beta(1-\beta)}{\eta}\langle w_{t-1}-w_t, g_\nu(v_t)\rangle + cd\beta^2G^2 \\
&= (1-\beta)^2\|m_{t-1}\|^2 + \frac{2\beta(1-\beta)}{\eta}\Big(\langle w_{t-1}-w_t, g_\nu(w_t)\rangle \\
&\qquad\qquad + \langle w_{t-1}-w_t, g_\nu(v_t)-g_\nu(w_t)\rangle\Big) + cd\beta^2G^2 \\
&\overset{\zeta_2}{\leq} (1-\beta)^2\|m_{t-1}\|^2 + \frac{2\beta(1-\beta)}{\eta}\Big(f_\nu(w_{t-1}) - f_\nu(w_t) + \frac{\rho}{2}\|w_t-w_{t-1}\|^2 \\
&\qquad\qquad + \rho\gamma\|w_t-w_{t-1}\|^2\Big) + cd\beta^2G^2 \\
&= (1-\beta)^2\|m_{t-1}\|^2 + \beta(1-\beta)\rho\eta(1+2\gamma)\|m_{t-1}\|^2 \\
&\qquad + \frac{2\beta(1-\beta)}{\eta}\left(f_\nu(w_{t-1}) - f_\nu(w_t)\right) + cd\beta^2G^2 \\
&\overset{\zeta_3}{\leq} (1-\beta)^2\|m_{t-1}\|^2 + \beta(1-\beta)\|m_{t-1}\|^2 \\
&\qquad + \frac{2\beta(1-\beta)}{\eta}\left(f_\nu(w_{t-1}) - f_\nu(w_t)\right) + cd\beta^2G^2 \\
&= (1-\beta)\|m_{t-1}\|^2 + \frac{2\beta(1-\beta)}{\eta}\left(f_\nu(w_{t-1}) - f_\nu(w_t)\right) + cd\beta^2G^2.
\end{aligned}
$$

where in "$\zeta_1$" we used $\mathbb{E}[\hat{g}_\nu(v_t;z_t) \mid \mathcal{F}_{t-1}] = g_\nu(v_t)$ and $\mathbb{E}\big[\|\hat{g}_\nu(v_t;z_t)\|^2 \mid \mathcal{F}_{t-1}\big] \leq cdG^2$, in "$\zeta_2$" we used the $\rho$-weak convexity of $f_\nu$ and $v_t - w_t = -\gamma(w_{t-1}-w_t)$; when $\gamma = 0$, the corresponding term vanishes directly, and in "$\zeta_3$" we used the conditions $\gamma \leq 1$ and $\eta \leq \frac{1}{3\rho}$. Then it follows that

$$
\mathbb{E}\left[\|m_t\|^2\right] \leq \mathbb{E}\left[(1-\beta)\|m_{t-1}\|^2 + \frac{2\beta(1-\beta)}{\eta}\left(f_\nu(w_{t-1}) - f_\nu(w_t)\right)\right] + cd\beta^2G^2. \tag{6}
$$

Rearranging yields

$$
\mathbb{E}\left[\beta\|m_{t-1}\|^2\right] \leq \mathbb{E}\left[\|m_{t-1}\|^2 - \|m_t\|^2 + \frac{2\beta(1-\beta)}{\eta}\left(f_\nu(w_{t-1}) - f_\nu(w_t)\right)\right] + cd\beta^2G^2.
$$

Summing over $t = 1, \ldots, T$ gives

$$
\mathbb{E}\left[\beta\sum_{t=1}^{T}\|m_{t-1}\|^2\right] \leq \mathbb{E}\left[\|m_0\|^2 - \|m_T\|^2 + \frac{2\beta(1-\beta)}{\eta}\left(f_\nu(w_0) - f_\nu(w_T)\right)\right] + cd\beta^2G^2\,T,
$$

which implies the claim. □

### D.2 Proof for Lemma 8

*Proof.* Fix some $k \in [K]$. It can be shown that

$$\begin{aligned}
\|w_k - \bar{w}\|^2 &= \left\| w_k - \frac{1}{K}\sum_{k'=1}^{K} w_{k'} \right\|^2 \le \frac{1}{K}\sum_{k'=1}^{K} \|w_k - w_{k'}\|^2 \\
&= \frac{1}{K}\sum_{k'=1}^{K} \left\| \sum_{j=k\wedge k'+1}^{k\vee k'} (w_{j-1} - w_j) \right\|^2 \\
&\le \frac{1}{K}\sum_{k'=1}^{K} \left( \sum_{j=k\wedge k'+1}^{k\vee k'} \|\Delta_j\| \right)^2 \\
&\le \left( \sum_{j=1}^{K} \|\Delta_j\| \right)^2 \le K \sum_{j=1}^{K} \|\Delta_j\|^2.
\end{aligned}$$

Additionally,

$$\begin{aligned}
\|v_k - \bar{w}\|^2 &= \|v_k - w_k + w_k - \bar{w}\|^2 \\
&= \|\gamma(w_k - w_{k-1}) + w_k - \bar{w}\|^2 \\
&\le 2\gamma^2 \|w_k - w_{k-1}\|^2 + 2\|w_k - \bar{w}\|^2 \\
&\le 2\gamma^2 \|\Delta_k\|^2 + 2K \sum_{j=1}^{K} \|\Delta_j\|^2 \\
&\le 2(K+1) \sum_{j=1}^{K} \|\Delta_j\|^2,
\end{aligned}$$

where we used $\gamma \le 1$. □

### D.3 Proof for Theorem 1

*Proof.* For any fixed $j \in [J]$, by applying Lemma 3 we get

$$\begin{aligned}
&\mathbb{E}\left[ \underbrace{\beta \sum_{k=1}^{K} \left\| m_{k-1}^{(j)} \right\|^2}_{\mathtt{A}} + \underbrace{\left\| m_K^{(j)} \right\|^2}_{\mathtt{B}} \right] \\
&\le \mathbb{E}\left[ \|m_0^{(j)}\|^2 + \frac{2\beta(1-\beta)}{\eta} \left( f_\nu(w_0^{(j)}) - f_\nu(w_K^{(j)}) \right) \right] + cd\beta^2 G^2 K \\
&= \mathbb{E}\left[ \frac{2\beta(1-\beta)}{\eta} \left( f_\nu(w_0^{(j)}) - f_\nu(w_0^{(j+1)}) \right) \right] + cd\beta^2 G^2 K,
\end{aligned} \tag{7}$$

where we used $m_0^{(j)} = 0$ and $w_0^{(j+1)} = w_K^{(j)}$. Next we lower bound $\mathtt{A}$ and $\mathtt{B}$, respectively. Recall that $\bar{w}^{(j)} = \frac{1}{K}\sum_{k=1}^{K} w_k^{(j)}$. For $\mathtt{A}$, Lemma 8 implies

$$\mathtt{A} = \beta \sum_{k=1}^{K} \left\| m_{k-1}^{(j)} \right\|^2 = \beta \sum_{k=1}^{K} \frac{\left\| w_k^{(j)} - w_{k-1}^{(j)} \right\|^2}{\eta^2} \ge \frac{\beta}{2(K+1)\eta^2} \max_{k\in[K]} \left\| v_k^{(j)} - \bar{w}^{(j)} \right\|^2. \tag{8}$$

For $\mathtt{B}$, note that

$$m_K^{(j)} = (1-\beta)^K m_0^{(j)} + \sum_{k=1}^{K} (1-\beta)^{K-k} \beta \hat{g}_\nu(v_k^{(j)}; z_k^{(j)}) = \sum_{k=1}^{K} (1-\beta)^{K-k} \beta \hat{g}_\nu(v_k^{(j)}; z_k^{(j)}).$$

It follows that

$$\begin{aligned}
\mathtt{B} &= \mathbb{E}\left[\left\|m_K^{(j)}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\sum_{k=1}^{K}(1-\beta)^{K-k}\beta\hat{g}_\nu(v_k^{(j)};z_k^{(j)})\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\sum_{k=1}^{K}(1-\beta)^{K-k}\beta\left(\hat{g}_\nu(v_k^{(j)};z_k^{(j)})-g_\nu(v_k^{(j)})+g_\nu(v_k^{(j)})\right)\right\|^2\right] \\
&\overset{\zeta_1}{\geq} \mathbb{E}\left[\frac{1}{2}\left\|\sum_{k=1}^{K}(1-\beta)^{K-k}\beta g_\nu(v_k^{(j)})\right\|^2 - \left\|\sum_{k=1}^{K}(1-\beta)^{K-k}\beta\left(\hat{g}_\nu(v_k^{(j)};z_k^{(j)})-g_\nu(v_k^{(j)})\right)\right\|^2\right] \\
&\overset{\zeta_2}{\geq} \mathbb{E}\left[\frac{1}{2}\left\|\sum_{k=1}^{K}(1-\beta)^{K-k}\beta g_\nu(v_k^{(j)})\right\|^2\right] - c\sum_{k=1}^{K}d(1-\beta)^{2(K-k)}\beta^2G^2 \\
&= \mathbb{E}\left[\frac{(1-(1-\beta)^K)^2}{2}\left\|\bar{g}_\nu^{(j)}\right\|^2\right] - c\sum_{k=1}^{K}d(1-\beta)^{2(K-k)}\beta^2G^2,
\end{aligned} \tag{9}$$

where in "$\zeta_1$" we used $\|x+y\|^2 \geq \frac{1}{2}\|x\|^2 - \|y\|^2$, and in "$\zeta_2$" we used $\mathbb{E}\left[\left\|\sum_{k=1}^{K}(1-\beta)^{K-k}\beta\left(\hat{g}_\nu(v_k^{(j)};z_k^{(j)})-g_\nu(v_k^{(j)})\right)\right\|^2\right] = \sum_{k=1}^{K}(1-\beta)^{2(K-k)}\beta^2\,\mathbb{E}\left[\left\|\hat{g}_\nu(v_k^{(j)};z_k^{(j)})-g_\nu(v_k^{(j)})\right\|^2\right] \leq c\sum_{k=1}^{K}d(1-\beta)^{2(K-k)}\beta^2\,G^2$.

By submitting (8) and (9) to (7) and rearranging, we get

$$\begin{aligned}
&\mathbb{E}\left[\frac{\left(1-(1-\beta)^K\right)^2}{2}\left\|\bar{g}_\nu^{(j)}\right\|^2 + \frac{\beta}{2(K+1)\eta^2}\max_{k\in[K]}\left\|v_k^{(j)}-\bar{w}^{(j)}\right\|^2\right] \\
&\leq \mathbb{E}\left[\frac{2\beta(1-\beta)}{\eta}\left(f_\nu(w_0^{(j)})-f_\nu(w_0^{(j+1)})\right)\right] + cd\beta^2G^2K + c\sum_{k=1}^{K}d(1-\beta)^{2(K-k)}\beta^2G^2.
\end{aligned}$$

Averaging over $j\in[J]$ yields

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}\left(\left\|\bar{g}_\nu^{(j)}\right\|^2 + \frac{\beta}{(K+1)\eta^2\left(1-(1-\beta)^K\right)^2}\max_{k\in[K]}\left\|v_k^{(j)}-\bar{w}^{(j)}\right\|^2\right)\right] \\
&\leq \frac{4\beta(1-\beta)\Delta f_{\nu,0}}{\eta\left(1-(1-\beta)^K\right)^2J} + \frac{2cd\left(\beta^2G^2K+\beta G^2\right)}{\left(1-(1-\beta)^K\right)^2}.
\end{aligned} \tag{10}$$

For any $j\in[J]$, consider $V^{(j)} = \left\{v_k^{(j)}\right\}_{k\in[K]}$. Clearly, we have $\bar{g}_\nu^{(j)} \in \partial_V f_\nu$. Since $\mu \leq \mu' = \frac{1}{\eta(1-(1-\beta)^K)}\sqrt{\frac{\beta}{K+1}}$, in view of Lemma 6 we get

$$\begin{aligned}
\left\|\partial f_\nu(\bar{w}^{(j)})\right\|_{+\mu}^2 &\leq \left\|\partial f_\nu(\bar{w}^{(j)})\right\|_{+\mu'}^2 \\
&\leq \left(\left\|\bar{g}_\nu^{(j)}\right\| + \mu'\max_{k\in[K]}\left\|v_k^{(j)}-\bar{w}^{(j)}\right\|\right)^2 \\
&\leq 2\left(\left\|\bar{g}_\nu^{(j)}\right\|^2 + (\mu')^2\max_{k\in[K]}\left\|v_k^{(j)}-\bar{w}^{(j)}\right\|^2\right).
\end{aligned} \tag{11}$$

Submitting (11) to (10) leads to

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}\left\|\partial f_\nu(\bar{w}^{(j)})\right\|_{+\mu}^2\right] \leq \frac{8\beta(1-\beta)\Delta f_{\nu,0}}{\eta\left(1-(1-\beta)^K\right)^2J} + \frac{4cd\beta G^2\left(\beta K+1\right)}{\left(1-(1-\beta)^K\right)^2}.$$

□

### D.4 Proof for Corollary 1

*Proof.* Since $\eta = 2(K+1)^{-1}\mu^{-1}$, $K = \lceil d^{1/2}T^{1/2}\mu^{-1/2} \rceil$ and $T \geq 36d^{-1}\rho^2\mu^{-1}$, it can be verified that

$$\eta = \frac{2}{(K+1)\mu} \leq \frac{2}{d^{1/2}T^{1/2}\mu^{1/2}} \leq \frac{1}{3\rho}.$$

Also, we have

$$\frac{1}{\eta\left(1-(1-\beta)^K\right)}\sqrt{\frac{\beta}{K+1}} \geq \frac{1}{\eta}\sqrt{\frac{\beta}{K+1}} = \mu.$$

Then by applying Theorem 1 we obtain that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}\left\|\partial f_\nu(\bar{w}^{(j)})\right\|_{+\mu}^2\right] \leq & \frac{8\beta(1-\beta)\Delta f_{\nu,0}}{\eta\left(1-(1-\beta)^K\right)^2 J} \\
& + \frac{4cd\beta G^2\left(\beta K+1\right)}{\left(1-(1-\beta)^K\right)^2} \\
\overset{\zeta_1}{\lesssim} & \frac{\beta\Delta f_{\nu,0}}{\eta J} + d\beta G^2 \\
\lesssim & \frac{d^{1/2}\mu^{1/2}(\Delta f_0 + G + G^2)}{T^{1/2}},
\end{aligned}$$

then we have

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}\left\|\partial f_\nu(\bar{w}^{(j)})\right\|_{+\mu}^2\right] \leq \epsilon^2,$$

where we have used in "$\zeta_1$" the choice of $\beta = 4/(K+1)$ which implies that $1-(1-\beta)^K \geq 1-\exp\{-\beta K\} = 1-\exp\{-\frac{4K}{K+1}\} \geq 1-\exp\{-2\} \geq 0.86$ and $\beta K \leq 4$, and in the last two steps the choices of $\beta, \eta, J$ and the condition on $T$. Since $\bar{w} \sim \mathrm{Unif}(\{\bar{w}^{(j)} : j \in [J]\})$, we further have

$$\begin{aligned}
\mathbb{E}\left[\left\|\partial f_\nu(\bar{w})\right\|_{+\mu}\right] &\leq \sqrt{\mathbb{E}\left[\left\|\partial f_\nu(\bar{w})\right\|_{+\mu}^2\right]} \\
&= \sqrt{\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}\left\|\partial f_\nu(\bar{w}^{(j)})\right\|_{+\mu}^2\right]} \leq \epsilon.
\end{aligned}$$

This proves the desired bound. □

### D.5 Proof for Corollary 2

The proof for the first-order weakly convex case is nearly identical to its zeroth-order counterpart. The corresponding algorithm is presented in Algorithm 5. It suffices to replace the smoothed surrogate objective and the second-moment bound for the zeroth-order gradient estimator with the corresponding first-order objective and the expectation of the first-order stochastic gradient, respectively. We therefore omit the details of other theorems.

*Proof.* Since $\eta = 2(K+1)^{-1}\mu^{-1}$, $K = \lceil T^{1/2}\mu^{-1/2} \rceil$ and $T \geq 36\rho^2\mu^{-1}$, it can be verified that

$$\eta = \frac{2}{(K+1)\mu} \leq \frac{2}{T^{1/2}\mu^{1/2}} \leq \frac{1}{3\rho}.$$

Also, we have

$$\frac{1}{\eta\left(1-(1-\beta)^K\right)}\sqrt{\frac{\beta}{K+1}} \geq \frac{1}{\eta}\sqrt{\frac{\beta}{K+1}} = \mu.$$

**Algorithm 5** SGDM with periodic momentum resetting

**Require:** Momentum $\beta \in (0,1]$, stepsize $\eta > 0$, extrapolation $\gamma \in [0,1]$, inner length $K \in \mathbb{N}$, number of epochs $J \in \mathbb{N}$, initial point $w_0$.

$w_K^{(0)} \leftarrow w_0$.
**for** $j = 1, 2, \ldots, J$ **do**
  $w_0^{(j)} \leftarrow w_K^{(j-1)}$; $\quad m_0^{(j)} \leftarrow 0$.
  **for** $k = 1, 2, \ldots, K$ **do**
    $w_k^{(j)} \leftarrow w_{k-1}^{(j)} - \eta\, m_{k-1}^{(j)}$.
    $v_k^{(j)} \leftarrow w_k^{(j)} + \gamma\big(w_k^{(j)} - w_{k-1}^{(j)}\big)$.
    Randomly sample $z_k^{(j)} \sim \mathcal{D}$ and compute $\widehat{g}\Big(v_k^{(j)}; z_k^{(j)}\Big) \in \partial \ell\Big(v_k^{(j)}; z_k^{(j)}\Big)$.
    $m_k^{(j)} \leftarrow (1-\beta) m_{k-1}^{(j)} + \beta\, \widehat{g}\Big(v_k^{(j)}; z_k^{(j)}\Big)$.
  **end for**
  $\bar{w}^{(j)} \leftarrow \frac{1}{K} \sum_{k=1}^{K} w_k^{(j)}$.
**end for**
**Output:** $\bar{w} \sim \mathrm{Unif}\big(\{\bar{w}^{(j)} : j \in [J]\}\big)$.

Then similar to Corollary 1 we obtain that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}\Big\|\partial f(\bar{w}^{(j)})\Big\|_{+\mu}^2\right] \leq& \frac{8\beta(1-\beta)\Delta f_0}{\eta\left(1-(1-\beta)^K\right)^2 J} + \frac{4(\beta^2 G^2 K + \beta G^2)}{\left(1-(1-\beta)^K\right)^2} \\ \overset{\zeta_1}{\leq}& \frac{11\beta\Delta f_0}{\eta J} + 28\beta G^2 \\ \leq& \frac{\mu^{1/2}(44\Delta f_0 + 112 G^2)}{T^{1/2}} \leq \epsilon^2,\end{aligned}$$

where we have used in $\zeta_1$ the choice $\beta = 4/(K+1)$, which implies

$$1-(1-\beta)^K \geq 1 - \exp\{-\beta K\} = 1 - \exp\left\{-\frac{4K}{K+1}\right\} \geq 1 - \exp\{-2\} \geq 0.86,$$

and $\beta K \leq 4$. In the last two steps, we use the choices of $\beta, \eta, J$ and the condition on $T$. Since $\bar{w} \sim \mathrm{Unif}\big(\{\bar{w}^{(j)} : j \in [J]\}\big)$, we further have

$$\mathbb{E}\Big[\|\partial f(\bar{w})\|_{+\mu}\Big] \leq \sqrt{\mathbb{E}\Big[\|\partial f(\bar{w})\|_{+\mu}^2\Big]} = \sqrt{\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}\Big\|\partial f(\bar{w}^{(j)})\Big\|_{+\mu}^2\right]} \leq \epsilon.$$

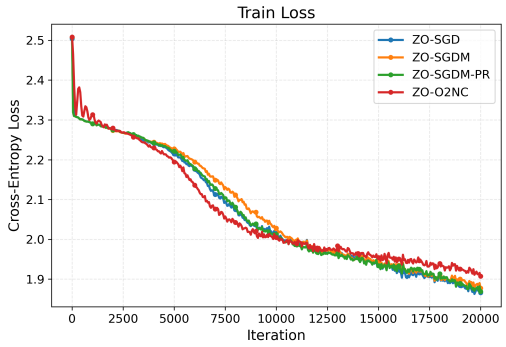
This proves the desired bound.

□

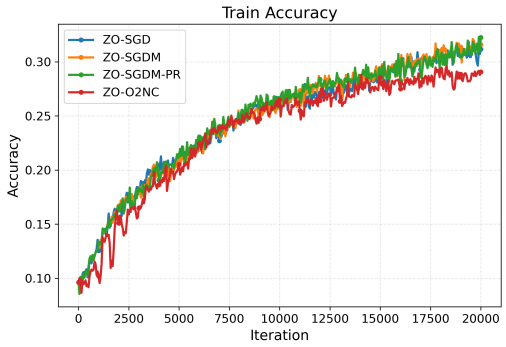
## E EXPERIMENT

We conduct two sets of experiments. The first experiment evaluates the same algorithmic idea on a larger-scale CIFAR-10 [4] (Krizhevsky & Hinton (2009)) classification task with ResNet-18, where we compare zeroth-order SGD, zeroth-order SGDM, and zeroth-order SGDM with periodic momentum resetting. The second experiment is designed to examine the empirical convergence-rate behavior of the proposed zeroth-order momentum method under smooth and nonsmooth objectives using MNIST [5](LeCun et al. (1998)). Similar setups can be found in Lin et al. (2022). For each algorithm and each evaluation point, we compute the mean and empirical standard deviation across random seeds using NumPy. The solid curve represents the mean, and the shaded region is plotted as $\bar{y}_t \pm \sigma_t$ using Matplotlib's `fill_between`.

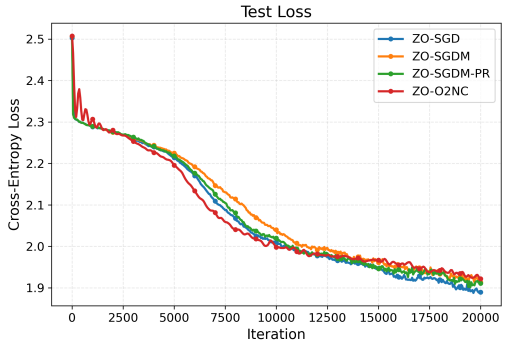
[4] https://www.cs.toronto.edu/ kriz/cifar.html
[5] http://yann.lecun.com/exdb/mnist



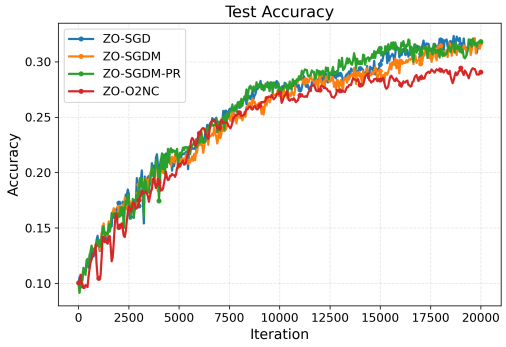

(a) Train loss.




(b) Train accuracy.




(c) Test loss.




(d) Test accuracy.

### E.1 Algorithmic Comparison on CIFAR-10

**Goal** This experiment compares ZO-SGDM-PR with standard zeroth-order SGD-type methods and O2NC, with particular emphasis on whether the additional mechanisms in ZO-SGDM-PR substantially alter the empirical behavior of standard ZO-SGDM.

**Dataset and model** We use the standard CIFAR-10 training and test splits. The model is ResNet-18 adapted to $32 \times 32$ images: the first convolution is replaced by a $3 \times 3$ convolution with stride 1, the initial max-pooling layer is removed, and the final fully connected layer is modified for 10 classes.

**Compared methods** We compare four zeroth-order methods: ZO-SGD, ZO-SGDM, ZO-SGDM-PR, and O2NC. Here ZO-SGDM-PR denotes zeroth-order SGDM with periodic momentum resetting and epoch averaging.

**Training and evaluation protocol** All methods are initialized from the same model parameters and use the same zeroth-order query configuration, with smoothing radius $\nu = 10^{-4}$, $q = 8$ random directions, batch size 256, 20,000 training iterations, and evaluation every 50 iterations. For ZO-SGD, ZO-SGDM, and ZO-SGDM-PR, we use learning rate $\eta = 3 \times 10^{-4}$. The momentum parameter is set to 0.1 for the momentum-based methods, and ZO-SGDM-PR resets its momentum every 1000 iterations.

For O2NC, we use learning rate $\eta = 10^{-6}$, displacement projection radius $D = 0.05$, and output-averaging parameter $\nu_{\mathrm{O2NC}} = 1$, corresponding to one output block every 20 iterations, we tune its numerical value for the present experiment. All methods are evaluated under the same function-query budget.

The loss is the standard multiclass cross-entropy loss.

**Result** As shown in Figure 1a–1d, ZO-SGDM-PR exhibits empirical behavior close to that of standard ZO-SGDM, indicating that periodic resetting and epoch averaging do not substantially alter the underlying SGDM dynamics in this experiment. In contrast, O2NC exhibits a noticeably different optimization trajectory. This comparison further supports the view that ZO-SGDM-PR provides a closer recovery of standard ZO-SGDM behavior, while achieving the same best-known dimension dependence as the more specialized O2NC framework.

### E.2 Empirical comparison of decay behavior between smooth and nonsmooth

**Goal** This experiment is intended only as a qualitative comparison of empirical decay behavior and is not a direct implementation of the theoretical nonsmooth setting analyzed above. Since the two settings involve different losses and regularizers, we focus on the evolution of their respective stationarity surrogates rather than their absolute objective values. The log-log plots and fitted slopes provide a descriptive summary of the observed convergence trends over the tested budget range.

**Dataset and model** We construct a binary classification task from MNIST by keeping only digits 3 and 8, and encode the labels as $+1$ and $-1$, respectively. Unless otherwise specified, we use $4000$ training examples and $1000$ test examples. The model is a two-hidden-layer multilayer perceptron with hidden widths $128$ and $64$, using $\tanh$ activations.

**Objectives** We consider a smooth objective and a nonsmooth composite objective. In the smooth setting, we minimize the logistic loss with $\ell_2$ regularization,

$$f_{\mathrm{sm}}(w) = \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-y_i h_w(x_i))) + \frac{\mu}{2}\|w\|_2^2.$$

In the nonsmooth setting, we minimize the hinge loss with $\ell_1$ regularization,

$$f_{\mathrm{nsm}}(w) = \frac{1}{n}\sum_{i=1}^{n} \max\{1-y_i h_w(x_i), 0\} + \lambda\|w\|_1.$$

The hinge loss and the $\ell_1$ penalty make the second objective nonsmooth. In the nonsmooth composite case, the zeroth-order estimator is applied to the hinge-loss data-fitting term, while the $\ell_1$ term is handled by the soft-thresholding proximal operator.

**Algorithm** We evaluate the proposed zeroth-order SGDM method with periodic momentum resetting. At the beginning of each epoch, the momentum variable is reset to zero. Within each epoch, the momentum recursion is updated using the zeroth-order estimator. At the beginning of each epoch, the momentum variable is reset to zero and the iterate is initialized from the last iterate of the previous epoch. The epoch average is used only for output and does not affect the subsequent SGDM updates. For the nonsmooth composite objective, the update additionally applies the soft-thresholding proximal map for the $\ell_1$ penalty.

**Stationarity measures and empirical decay exponent** For the smooth objective, we use the squared full-batch gradient norm

$$\mathcal{M}_{\mathrm{sm}}(w) = \|\nabla f_{\mathrm{sm}}(w)\|^2.$$

For the nonsmooth objective, we use a proximal-gradient-type stationarity surrogate

$$\mathcal{M}_{\mathrm{nsm}}(w) = \|\mathcal{G}_{\eta_{\mathrm{eval}}}(w)\|^2,$$

where

$$\mathcal{G}_{\eta_{\mathrm{eval}}}(w) = \frac{1}{\eta_{\mathrm{eval}}}\Big(w - \mathrm{prox}_{\eta_{\mathrm{eval}}\lambda\|\cdot\|_1}\big(w - \eta_{\mathrm{eval}} g(w)\big)\Big).$$

Here $g(w)$ denotes the subgradient of the hinge-loss data-fitting term returned by automatic differentiation. We use a fixed evaluation stepsize $\eta_{\mathrm{eval}}$ when computing this surrogate so that changes in the training stepsize do not change the scale of the stationarity measure.

To estimate the empirical convergence exponent, we fit a linear model to the log-log stationarity trajectory:

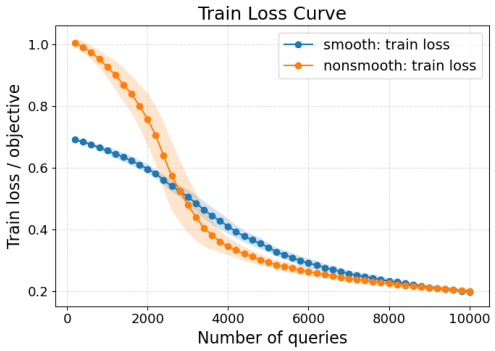
$$\log \mathcal{M}(w_t) = -\alpha \log Q_t + c,$$

where $Q_t$ denotes the cumulative number of function queries. The fitted exponent $\alpha$ is the negative slope of the least-squares regression line. To reduce the influence of the initial transient phase, we fit the slope using the last portion of the trajectory and report results for $40\%$ tail ratios. The main reported exponent is the median over random seeds.

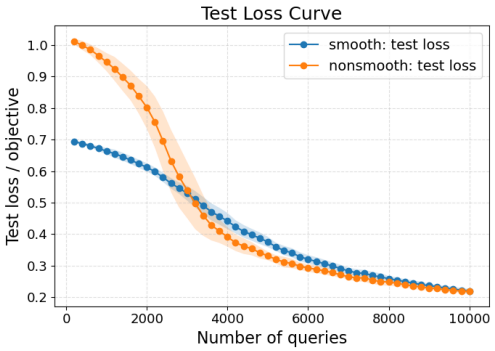
**Evaluation metrics** In the smooth setting, the reported objective is the logistic loss with $\ell_2$ regularization. In the nonsmooth setting, the reported objective is the hinge loss with $\ell_1$ regularization. These objective values have different numerical scales and should not be directly compared across settings. They are used to verify descent within each setting, while classification accuracy provides a common predictive-performance metric.

**Hyperparameters** In our implementation, we use batch size 128, smoothing radius $\nu = 10^{-2}$, momentum parameter $\beta = 0.1$, $\mu = 10^{-4}$, and $\lambda = 10^{-5}$. The learning rate is set to $0.01$ in the smooth setting and $0.02$ in the nonsmooth setting. We use inner-loop length $K = 50$ and number of epochs $J = 20$.

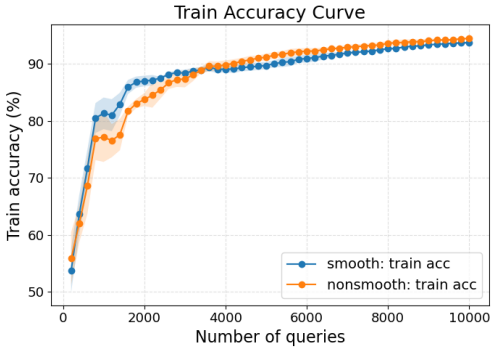
**Result** As shown in Figure 2e, as algorithms convergence, the iteration-dependent term gradually dominates the convergence behavior. Consequently, the slope of the log-log stationarity curve becomes close to the fitted decay exponent. The fitted exponents are $\alpha = 2.068$ for the nonsmooth setting and $\alpha = 2.217$ for the smooth setting. These values are close to each other, indicating that ZO-SGDM achieves comparable empirical stationarity decay in both settings.



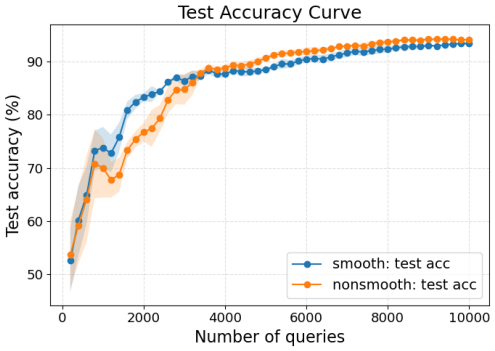

(a) Train loss



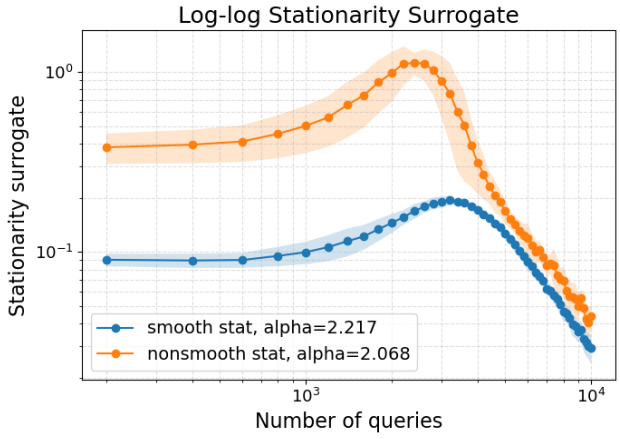

(b) Test loss




(c) Train acc.




(d) Test acc.




(e) Stationarity