Stochastic Zeroth-Order Optimization Under Heavy-Tailed Noise

Taha El Bakkali[1]* **El Mahdi Chayti**[2] **Qiuyi Zhang**[3] **Imane Rahali**[1] **Omar Saadi**[1]
[1]College of Computing, Mohammed VI Polytechnic University (UM6P)
[2]Machine Learning and Optimization Laboratory (MLO), EPFL
[3] Elorian AI

Abstract

We study stochastic zeroth-order (ZO) optimization of smooth nonconvex objectives under heavy-tailed sample-gradient noise, motivated by evidence that gradient noise in modern machine learning can violate bounded-variance assumptions. Establishing high-probability guarantees in this setting requires transferring tail control of the unobserved sample gradients to directional finite-difference estimates constructed from function values. We carry out this analysis under a weak-L_p tail condition for $p \in (1, 2]$, which is weaker than the standard bounded p-th moment assumption.

We propose the Robust Scalar-Clipped Zeroth-Order method (**RSC-ZO**), which clips each scalar two-point directional estimate before aggregation. Under sample-wise smoothness and shared randomness within each pair of function evaluations, we establish a high-probability stationarity guarantee with query complexity

$$\widetilde{O}\left(\frac{d^{\frac{p}{2(p-1)}}}{\varepsilon^{\frac{3p-2}{p-1}}}\right),$$

suppressing logarithmic and problem-dependent factors. This matches the ε-dependence of the general first-order heavy-tailed benchmark. At $p = 2$, the bound becomes $\widetilde{O}(d\varepsilon^{-4})$, recovering the classical stochastic ZO dimension–accuracy dependence with a high-probability guarantee under weak-L_2 noise, which can have infinite variance. We also analyze a momentum variant and quantify its batch-size/stepsize tradeoff. With an explicit choice of momentum, the running batch size can be reduced to one after a warm start while preserving the total query order up to logarithmic factors.

1 Introduction

We consider the stochastic optimization problem

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}_{\xi \sim \mathbb{P}}\big[F(x; \xi)\big],$$

where $(\Xi, \mathcal{F}, \mathbb{P})$ is a probability space. For each ξ, $x \mapsto F(x, \xi)$ is differentiable, and for each x, it is measurable with respect to ξ. We assume that f is finite-valued and bounded below. The finite-sum case is recovered by taking $\Xi = \{1, \dots, n\}$ and $\mathbb{P}$ to be the uniform distribution.

In this paper, we consider the zeroth-order (ZO) setting where the algorithm can query only noisy function values $F(x; \xi)$, with no access to gradients. This oracle model arises naturally in black-box adversarial attacks [Chen et al., 2017, Ilyas et al., 2018], simulation-based optimization [Fu, 2015],

*Corresponding author: taha.elbakkali@um6p.ma

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and, more recently, memory-efficient fine-tuning of large language models, where backpropagation is prohibitive and ZO methods such as MeZO achieve competitive performance using only forward passes [Malladi et al., 2023].

For smooth nonconvex objectives, stochastic ZO methods [Ghadimi and Lan, 2013, Chayti et al., 2025b] establish the benchmark $O(d\,\varepsilon^{-4})$ query complexity for finding an ε-stationary point in expectation, when each sample function $x \mapsto F(x;\xi)$ is smooth and the gradient noise has bounded variance.

The bounded-variance assumption is, however, increasingly at odds with the regimes in which ZO methods are deployed. Empirical evidence in modern machine learning indicates that stochastic gradient noise is often *heavy-tailed*. For example, Zhang et al. [2020] provide evidence of heavy-tailed stochastic-gradient noise in BERT pretraining, with empirical variance estimates that fail to stabilize as the sample size increases. Similar behavior has been reported in image classification: Şimşekli et al. [2019] estimate stochastic-gradient tail indices below 2, consistent with infinite-variance noise. Heavy-tailed gradients have also been documented in reinforcement learning [Garg et al., 2021]. In such regimes, the second moment may be infinite or poorly controlled. This poses a challenge not only for first-order gradient-based stochastic methods, but also for ZO methods, whose updates rely on gradient estimates reconstructed from noisy function evaluations.

The first-order literature has responded with a substantial line of work showing that *gradient clipping*, possibly combined with momentum or normalization, can restore convergence guarantees under bounded p-th moment noise on the sample gradients for $p \in (1, 2]$ [Gorbunov et al., 2020, Zhang et al., 2020, Cutkosky and Mehta, 2021, Sadiev et al., 2023, Nguyen et al., 2023, Liu et al., 2023, Chayti, 2026]. For smooth nonconvex first-order stochastic optimization under bounded p-th moment noise on the sample gradients, the standard benchmark is the optimal in-expectation ε-dependence $\Theta(\varepsilon^{-(3p-2)/(p-1)})$ Zhang et al. [2020]. Matching high-probability guarantees were later obtained by Cutkosky and Mehta [2021].

The zeroth-order picture is much less developed. For zeroth-order optimization, results under heavy-tailed noise are scarce and almost entirely confined to the convex setting. Kornilov et al. [2024] obtain complexity bounds for nonsmooth convex ZO under symmetric heavy-tailed noise using mini-batched median estimates of two-point gradient differences followed by vector clipping; Kornilov et al. [2023a] treat convex compact domains; and Kornilov et al. [2023b] give an accelerated method when the variance is infinite. To the best of our knowledge, no prior result appears to give a high-probability complexity bound for smooth nonconvex stochastic ZO that matches the classical finite-variance expectation benchmark. This work closes this gap under a *weak-L_p* tail condition on the sample gradient noise $\nabla F(x;\xi) - \nabla f(x)$ (Assumption 1.3), which is strictly weaker than the classical bounded p-th moment assumption and allows for infinite variance even when $p = 2$.

Why the nonconvex ZO case is not a corollary. A natural temptation is to view ZO under heavy tails as an immediate consequence of first-order results. In first-order SGD, the noise assumption is imposed directly on the stochastic gradient noise $\nabla F(x;\xi) - \nabla f(x)$, and the update uses $\nabla F(x;\xi)$ itself. Thus the stochastic error appearing in the update is exactly the quantity controlled by the assumption. In zeroth-order optimization, however, the algorithm does not observe $\nabla F(x;\xi)$. It only observes noisy function values and forms directional finite-difference estimates. Consequently, the error in the update is not the original sample-gradient noise vector, but a scalar directional derivative multiplied by a random direction, together with finite-difference and smoothing effects. Therefore, weak-L_p control of $\nabla F(x;\xi) - \nabla f(x)$ must first be transferred to the scalar two-point estimator. This directional transfer is the main ZO-specific step in our proof and is absent from standard first-order analyses. The difficulty is amplified by the fact that our assumption is a weak-L_p tail condition, which is strictly weaker than the classical bounded p-th moment assumption.

Contributions. We study a robust two-point ZO method, **RSC-ZO**, that combines spherical smoothing, shared-randomness directional estimates, and *scalar clipping applied to each directional derivative before aggregation* (Algorithm 1). Our contributions are as follows.

- **High-probability ZO stationarity under weak-L_p noise.** Under sample-wise smoothness (Assumption 1.2) and a weak-L_p tail assumption on the sample-gradient noise (Assumption 1.3),

we prove that RSC-ZO finds an ε-stationary point with high probability using $\widetilde{O}\left(\frac{d^{\frac{p}{2(p-1)}}}{\varepsilon^{\frac{3p-2}{p-1}}}\right)$ noisy function evaluations, up to logarithmic and problem-dependent factors (Theorem B.11 and Remark B.12). At $p = 2$, this gives $\widetilde{O}(d\,\varepsilon^{-4})$, matching the benchmark dimension–accuracy dependence of stochastic ZO methods under bounded-variance noise [Ghadimi and Lan, 2013, Chayti et al., 2025b]. Unlike these classical results, which are typically established in expectation, our guarantee holds with high probability and under a weaker weak-L_2 tail condition. Moreover, the ε-dependence matches the general first-order heavy-tailed benchmark established by Zhang et al. [2020].

- **A scalar-clipping analysis tailored to derivative-free estimation.** A key step in the proof is to transfer weak-L_p control of the sample-gradient noise to clipped two-point directional estimates. This step is specific to the zeroth-order setting: the algorithm never observes $\nabla F(x;\xi)$ directly, but only noisy function values. The resulting analysis separates smoothing bias, directional-estimation error, and heavy-tailed stochastic fluctuations before aggregation.
- **Momentum extension with a quantified batch-size/stability tradeoff.** We extend the analysis to a momentum variant (Algorithm 1). Using a localized Lyapunov argument, we show that momentum reduces the required running batch size by a factor of $1-\beta$ after warm start, at the price of the stability coupling $\alpha \lesssim (1-\beta)/L$. With an explicit choice of the momentum parameter, the running batch size can be reduced to one after a warm start while preserving the total query order up to logarithmic factors (Corollary C.9).

Comparison with related work. Our setting sits at the intersection of heavy-tailed first-order optimization and classical stochastic ZO theory.

Heavy-tailed first-order optimization. Zhang et al. [2020] analyze clipped SGD under bounded p-th moment sample-gradient noise and achieve the optimal in-expectation ε-dependence for smooth nonconvex first-order stochastic optimization. Cutkosky and Mehta [2021] obtain high-probability bounds using clipping, momentum, and normalization. More recently, Hübler et al. [2024], Liu and Zhou [2025] show that normalization alone, without clipping, can achieve the optimal rate in related nonconvex first-order settings. These results do not address the zeroth-order oracle, where the update is constructed from noisy function values rather than observed stochastic gradients.

Heavy-tailed zeroth-order optimization. The closest line of work is Kornilov et al. [2024, 2023a,b], which studies *convex*, typically nonsmooth, zeroth-order optimization under heavy-tailed oracle noise. These works often rely on additional structure, such as symmetry of the oracle noise, and use median-based gradient-difference estimators followed by vector clipping. In contrast, we treat *smooth nonconvex* objectives, do not require symmetry, and use scalar clipping directly on each directional derivative before aggregation.

Stochastic ZO in the bounded-variance regime. Ghadimi and Lan [2013] and Chayti et al. [2025b] establish the benchmark $O(d\,\varepsilon^{-4})$ query complexity for stochastic ZO methods under sample smoothness and bounded-variance noise, with guarantees stated in expectation.

1.1 Assumptions

We now state the assumptions used throughout the paper. Our analysis relies on two ingredients: sample-wise smoothness of the stochastic objective and a direct probability control on the sample gradient noise.

Assumption 1.1 (Well-defined problem)**.** *The function f is bounded from below, i.e.: $f_* := \inf_{x\in\mathbb{R}^d} f(x) > -\infty$ and we define the initial gap $\Delta_0 := f(x_0) - f_*$.*

Assumption 1.2 (Sample smoothness)**.** *For $\mathbb{P}$-almost every ξ, the map $x \mapsto F(x;\xi)$ is L-smooth, i.e., $\forall x, y \in \mathbb{R}^d, \quad \|\nabla F(x;\xi) - \nabla F(y;\xi)\|_2 \le L\,\|x-y\|_2$.*

Assumption 1.3 (Weak-L_p control of the sample gradient noise)**.** *There exist $p \in (1,2]$ and $\sigma > 0$ such that, for every $x \in \mathbb{R}^d$ and every $t > 0$, $\mathbb{P}_\xi\big(\|\nabla F(x;\xi) - \nabla f(x)\|_2 > t\big) \le \big(\frac{\sigma}{t}\big)^p$.*

Remark 1.4. Assumption 1.3 is weaker than the bounded p-th moment condition used in previous works. Indeed, a uniform bound

$$\mathbb{E}_\xi\,\|\nabla F(x;\xi) - \nabla f(x)\|_2^p \le \sigma^p$$

implies Assumption 1.3 by Markov's inequality, while the converse need not hold: weak-L_p control may allow an infinite p-th moment (see Proposition A.1).

Remark 1.5. Under Assumption 1.2, in order to justify $\nabla f(x) = \mathbb{E}_\xi[\nabla F(x;\xi)]$ and conclude that f is L-smooth, it is enough to assume that there exists $x_0 \in \mathbb{R}^d$ such that $\mathbb{E}_\xi \|\nabla F(x_0;\xi)\|_2 < \infty$. Indeed, for any compact set $K \subset \mathbb{R}^d$, $\sup_{x\in K} \|\nabla F(x;\xi)\|_2 \le \|\nabla F(x_0;\xi)\|_2 + L \sup_{x\in K} \|x - x_0\|_2$, so dominated convergence yields $\nabla f(x) = \mathbb{E}_\xi[\nabla F(x;\xi)]$. Then

$$\|\nabla f(x) - \nabla f(y)\|_2 \le \mathbb{E}_\xi \|\nabla F(x;\xi) - \nabla F(y;\xi)\|_2 \le L \|x - y\|_2,$$

which shows that f is L-smooth. Throughout the paper, we assume that such a point x_0 exists.

1.2 Algorithm

We now introduce RSC-ZO and its momentum variant. Both methods use the same scalar-clipped two-point estimator. Given a point $x \in \mathbb{R}^d$, batch size M, and clipping threshold τ, sample $u_\ell \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and $\xi_\ell \sim \mathbb{P}$, using the same ξ_ℓ in the two function evaluations, and define

$$Y_\ell(x) := \frac{F(x + \mu u_\ell; \xi_\ell) - F(x - \mu u_\ell; \xi_\ell)}{2\mu}.$$

We clip each scalar directional estimate by $\psi_\tau(z) := z \min\{1, \tau/|z|\}$ and form the clipped zeroth-order estimator $G_\mu(x; M, \tau) := \frac{d}{M} \sum_{\ell=1}^{M} \psi_\tau(Y_\ell(x)) u_\ell$. The base method corresponds to $\beta = 0$ and updates directly with $G_\mu(x_t; M, \tau)$. The momentum variant corresponds to $\beta > 0$: in this case, we first initialize the momentum variable using a separate warm-start estimate at x_0, with batch size M_0 and threshold τ_0, and then run the momentum recursion. The warm start is used only in the momentum case. Algorithm 1 summarizes both variants.

Algorithm 1 Robust Scalar-Clipped Zeroth-Order Method (RSC-ZO)

Require: $x_0 \in \mathbb{R}^d$, stepsize $\alpha > 0$, smoothing parameter $\mu > 0$, running batch size M, threshold τ, momentum parameter $\beta \in [0, 1)$

if $\beta = 0$ **then**
Base RSC-ZO
  for $t = 0, 1, 2, \dots$ **do**
    $g_t \leftarrow G_\mu(x_t; M, \tau)$
    $x_{t+1} \leftarrow x_t - \alpha g_t$
  end for
else
Momentum RSC-ZO
Additional inputs: warm-start batch size M_0, warm-start threshold τ_0
  $g_0 \leftarrow G_\mu(x_0; M_0, \tau_0)$
  $m_0 \leftarrow g_0$
  $x_1 \leftarrow x_0 - \alpha m_0$
  for $t = 1, 2, \dots$ **do**
    $g_t \leftarrow G_\mu(x_t; M, \tau)$
    $m_t \leftarrow \beta m_{t-1} + (1 - \beta) g_t$
    $x_{t+1} \leftarrow x_t - \alpha m_t$
  end for
end if

1.3 Spherical smoothing and notation

For $\mu > 0$, define the spherical smoothing of f by $f_\mu(x) := \mathbb{E}_{v\sim\mathrm{Unif}(\mathbb{B}^d)}[f(x + \mu v)]$, where $\mathbb{B}^d := \{v \in \mathbb{R}^d : \|v\|_2 \le 1\}$ is the unit Euclidean ball. It is a classical result [Flaxman et al., 2004] that

$$\nabla f_\mu(x) = \frac{d}{\mu} \mathbb{E}_{u\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[f(x + \mu u) u].$$

Using the symmetry of the uniform distribution on $\mathbb{S}^{d-1}$, this can be rewritten in the two-point form

$$\nabla f_\mu(x) = d\, \mathbb{E}_{u\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[\frac{f(x + \mu u) - f(x - \mu u)}{2\mu} u\right].$$

Accordingly, for $x \in \mathbb{R}^d$ and $u \in \mathbb{S}^{d-1}$, we define $D_\mu f(x,u) := \frac{f(x+\mu u)-f(x-\mu u)}{2\mu}$, so that

$$\nabla f_\mu(x) = d\,\mathbb{E}_{u\sim \mathrm{Unif}(\mathbb{S}^{d-1})}[D_\mu f(x,u)u].$$

2 High-Probability Analysis of Base RSC-ZO

Notation. Let $\bar{\Delta}_0 := \Delta_0 + \frac{L\mu^2}{2}$. For brevity, define $S_\mu := \frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu$. This scale collects the three quantities that control the scalar directional finite-difference estimates. The term $\sqrt{L\bar{\Delta}_0}/\sqrt{d}$ controls the noiseless directional derivative, the term $\sigma/\sqrt{d}$ controls the stochastic gradient noise after projection onto the random direction u, and the term $L\mu$ controls the finite-difference smoothing error. After multiplication by the zeroth-order factor d, the corresponding estimator scale is dS_μ.

Goal. We first establish a high-probability bound on the estimator error $g(x) - \nabla f_\mu(x)$. For each sample, define the two-point directional derivative estimate

$$Y_\ell := \frac{F(x+\mu u_\ell;\xi_\ell) - F(x-\mu u_\ell;\xi_\ell)}{2\mu}.$$

The scalar-clipped zeroth-order estimator is then

$$g(x) := \frac{d}{M}\sum_{\ell=1}^{M} \psi_\tau(Y_\ell)\,u_\ell,$$

where the pairs (u_ℓ, ξ_ℓ) are i.i.d., with $u_\ell \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and $\xi_\ell \sim \mathcal{P}$.

We decompose the estimator error as

$$g(x) - \nabla f_\mu(x) = \underbrace{g(x) - \mathbb{E}[g(x)]}_{A\text{: fluctuation}} + \underbrace{\mathbb{E}[g(x)] - \nabla f_\mu(x)}_{B\text{: clipping bias}}.$$

The term A captures the empirical fluctuation of the clipped average, while B captures the deterministic bias introduced by clipping. We control these terms separately in Lemma B.7 and Lemma B.5. Choosing the clipping level to balance the two terms, up to logarithmic factors, yields the following localized deviation bound for a single batched estimator at a fixed, non-random point x. The extension from this fixed-point estimate to the random trajectory is handled by localization and a union bound in the proof of Theorem 2.3.

Theorem 2.1 (Localized deviation with tuned clipping)**.** *Assume that Assumptions 1.1, 1.2, and 1.3 hold. Fix $\delta \in (0,1]$ and $T \geq 3$, set $\lambda := \log(T/\delta)$, and assume $M \geq \lambda$. Let $x \in \mathbb{R}^d$ satisfy $f(x) - f_* \leq 4\bar{\Delta}_0$, and choose $\tau := 8\,S_\mu\left(\frac{M}{\lambda}\right)^{1/p}$. Then, with probability at least $1-\delta/T$,*

$$\|g(x) - \nabla f_\mu(x)\|_2 \leq C_p\, d\, S_\mu \left(\frac{\lambda}{M}\right)^{\frac{p-1}{p}} \sqrt{1+\log(M/\lambda)},$$

where $C_p := 120 + \frac{2^{4-p}}{p-1}$.

All intermediate lemmas appear in Appendix B. These include the clipping-bias bound under weak-L_p control, directional weak-L_p bounds for the noiseless signal and the shared-randomness estimator Y, the clipped second-moment estimate, and the vector Bernstein inequality. The corresponding detailed appendix statement is Corollary B.9.

We turn the localized deviation bound of Theorem 2.1 into a descent estimate for the smoothed objective f_μ. Since the update uses the clipped estimator $g(x_t)$ in place of $\nabla f_\mu(x_t)$, the key step is to combine the smoothness inequality for f_μ with the high-probability control of $g(x_t) - \nabla f_\mu(x_t)$.

Proposition 2.2 (One-step descent inequality)**.** *Assume f is L-smooth. For $x \in \mathbb{R}^d$, let $x^+ := x - \alpha g(x)$ and $e(x) := g(x) - \nabla f_\mu(x)$. If $\alpha \leq \frac{1}{4L}$, then*

$$f_\mu(x^+) \leq f_\mu(x) - \frac{\alpha}{2}\|\nabla f_\mu(x)\|_2^2 + \frac{5\alpha}{4}\|e(x)\|_2^2.$$

For $M \in \mathbb{N}$, $\mu > 0$, $\delta \in (0,1)$, and $T \geq 3$, define

$$\eta_0(M, \mu, \delta, T) := C_p \, d \, S_\mu \left(\frac{\log(T/\delta)}{M} \right)^{\frac{p-1}{p}} \sqrt{1 + \log\left(\frac{M}{\log(T/\delta)} \right)},$$

and $\eta(M, \mu, \delta, T) := \eta_0(M, \mu, \delta, T) + L\mu$. Iterating Proposition 2.2 along the trajectory of Algorithm 1, applying the localized deviation bound of Theorem 2.1 on the event that the iterates remain in the region $f(x_t) - f_* \leq 4\bar{\Delta}_0$, and taking a union bound over T steps yields our main result. The localization argument is given in Appendix B.

Theorem 2.3 (High-probability convergence of base RSC-ZO)**.** *Assume that Assumptions 1.1, 1.2, and 1.3 hold. Let $\mu > 0$, $\Delta_0 := f(x_0) - f_*$, and $\bar{\Delta}_0 := \Delta_0 + \frac{L\mu^2}{2}$. Let $\epsilon > 0$, $\delta \in (0,1]$, $M \in \mathbb{N}$, and define*

$$T := \max\left\{ 3, \left\lceil \frac{32L\bar{\Delta}_0}{\epsilon^2} \right\rceil \right\}.$$

Run Algorithm 1 with $\beta = 0$ (no momentum), batch size M, stepsize $\alpha := \frac{1}{4L}$, and clipping threshold $\tau := 8\, S_\mu \left(\frac{M}{\log(T/\delta)} \right)^{1/p}$. Assume $M \geq \log(T/\delta)$, $\eta(M, \mu, \delta, T) \leq \frac{\epsilon}{4}$ and $\epsilon^2 \leq 32L\bar{\Delta}_0$. Then, with probability at least $1 - \delta$,

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2 \leq \epsilon^2.$$

Remark 2.4 (Query complexity)**.** The total number of stochastic function queries is $2MT$. Choose $\mu = \epsilon/(8Ld)$. Then $L\mu \leq \epsilon/8$, $dL\mu = O(\epsilon)$, and $\bar{\Delta}_0 = \Delta_0 + O(\epsilon^2/(Ld^2))$. Since

$$T = \max\left\{ 3, \left\lceil \frac{32L\bar{\Delta}_0}{\epsilon^2} \right\rceil \right\},$$

we have $T = O(1 + L\Delta_0/\epsilon^2)$. Up to logarithmic factors, the batch condition $\eta(M, \mu, \delta, T) \leq \epsilon/4$ requires $dS_\mu \left(\frac{\log(T/\delta)}{M} \right)^{\frac{p-1}{p}} \lesssim \epsilon$. Under the above smoothing choice,

$$dS_\mu = \sqrt{d}\,(\sqrt{L\bar{\Delta}_0} + \sigma) + dL\mu = \sqrt{d}\,(\sqrt{L\Delta_0} + \sigma) + O(\epsilon).$$

Thus the conditions $M \geq \log(T/\delta)$ and $\eta(M, \mu, \delta, T) \leq \epsilon/4$ can be satisfied with a batch size

$$M = \widetilde{O}\left(\max\left\{ 1, \left(\frac{\sqrt{d}\,(\sqrt{L\Delta_0} + \sigma)}{\epsilon} \right)^{\frac{p}{p-1}} \right\} \right).$$

Here $\widetilde{O}$ suppresses logarithmic factors and constants depending only on p. Consequently,

$$2MT = \widetilde{O}\left[\left(1 + \frac{L\Delta_0}{\epsilon^2} \right) \max\left\{ 1, \left(\frac{\sqrt{d}\,(\sqrt{L\Delta_0} + \sigma)}{\epsilon} \right)^{\frac{p}{p-1}} \right\} \right].$$

In particular, when $L\Delta_0 \geq \epsilon^2$, this simplifies to $2MT = \widetilde{O}\left(L\Delta_0 (\sqrt{L\Delta_0} + \sigma)^{\frac{p}{p-1}} \frac{d^{\frac{p}{2(p-1)}}}{\epsilon^{\frac{3p-2}{p-1}}} \right)$. At $p = 2$, the bound becomes

$$2MT = \widetilde{O}\left(L\Delta_0 (\sqrt{L\Delta_0} + \sigma)^2 \, d\, \epsilon^{-4} \right).$$

This recovers the classical $d\epsilon^{-4}$ dimension–accuracy dependence for stochastic ZO methods [Ghadimi and Lan, 2013, Chayti et al., 2025b], with a high-probability guarantee under the strictly weaker weak-L_2 tail condition. The ϵ-dependence also matches the general first-order heavy-tailed benchmark established by Zhang et al. [2020].

Remark 2.5 (Trivial regime)**.** The condition $\epsilon^2 \leq 32L\bar{\Delta}_0$ isolates the regime where the iteration bound above is needed. If $\epsilon^2 > 32L\bar{\Delta}_0$, then Lemma B.14 gives

$$\|\nabla f(x_0)\|_2^2 \leq 2L(f(x_0) - f_*) = 2L\Delta_0 \leq 2L\bar{\Delta}_0 < \epsilon^2.$$

Hence the desired stationarity conclusion already holds at the initial point.

3 High-Probability Analysis of Momentum RSC-ZO

We now study the momentum variant of RSC-ZO, corresponding to $\beta > 0$ in Algorithm 1. Averaging clipped directional estimates over time reduces the running batch size required per iteration; see [Chayti et al., 2025a, Section 2.2] for a discussion of this averaging effect. This reduction comes with a stepsize restriction of order $(1-\beta)/L$.

The method starts with a warm-start estimate

$$g_0 = G_\mu(x_0; M_0, \tau_0), \qquad m_0 = g_0, \qquad x_1 = x_0 - \alpha m_0.$$

For $t \geq 1$, it computes $g_t = G_\mu(x_t; M, \tau)$ using fresh independent samples and updates

$$m_t = \beta m_{t-1} + (1-\beta) g_t, \qquad x_{t+1} = x_t - \alpha m_t.$$

The following result gives explicit parameters for $\beta \in [1/2, 1)$.

Theorem 3.1 (High-probability convergence of momentum RSC-ZO)**.** *Suppose Assumptions 1.1, 1.2, and 1.3 hold. Let $\epsilon \in (0,1)$, $\delta \in (0,1]$, $\beta \in [1/2, 1)$, and $0 < \mu \leq \epsilon/(4Ld)$. Write*

$$\Delta_0 = f(x_0) - f_*, \qquad \bar{\Delta}_0 = \Delta_0 + \frac{L\mu^2}{2}, \qquad S_\mu = \frac{\sqrt{L\bar{\Delta}_0} + \sigma}{\sqrt{d}} + L\mu,$$

and assume $\epsilon^2 \leq 32 L \bar{\Delta}_0$. Set

$$C_p = 120 + \frac{2^{4-p}}{p-1}, \qquad A = \max\left\{ \mathrm{e},\ \frac{p}{p-1}\left(\frac{16 C_p d S_\mu}{\epsilon}\right)^2 \right\}, \qquad K = (A \log A)^{\frac{p}{2(p-1)}}.$$

Consider the parameters

$$\alpha = \frac{1-\beta}{16\sqrt{3}\, L}, \qquad T = \left\lceil \frac{512\sqrt{3}\, L \bar{\Delta}_0}{(1-\beta)\epsilon^2} \right\rceil + 2, \qquad \lambda = \log \frac{2T}{\delta}, \qquad \lambda_0 = \log \frac{2}{\delta}.$$

For Algorithm 1, take the batch sizes $M_0 = \lceil \lambda_0 K \rceil$, $M = \lceil (1-\beta)\lambda K \rceil$, with clipping thresholds $\tau_0 = 8 S_\mu \left(\frac{M_0}{\lambda_0}\right)^{1/p}$, $\tau = 8 S_\mu \left(\frac{M}{(1-\beta)\lambda}\right)^{1/p}$. Then, with probability at least $1-\delta$,

$$\frac{1}{T-2} \sum_{t=1}^{T-2} \|\nabla f(x_t)\|_2^2 \leq \epsilon^2.$$

The result follows from Corollary C.9; the supporting lemmas and proofs are given in Appendix C.

Remark 3.2 (Query complexity and momentum tradeoff)**.** The total number of stochastic function evaluations is $Q_{\mathrm{MOM}} = 2M_0 + 2M(T-1)$. The parameter choices above give

$$Q_{\mathrm{MOM}} = \widetilde{O}\left(K\left(1 + \frac{L\bar{\Delta}_0}{\epsilon^2}\right) + \frac{L\bar{\Delta}_0}{(1-\beta)\epsilon^2} \right),$$

where $\widetilde{O}$ suppresses logarithmic factors and constants depending only on p. The last term accounts for the fact that the running batch size cannot be smaller than one.

Compared with base RSC-ZO, momentum reduces the required running batch size by a factor of order $1-\beta$, while the stepsize decreases by the same factor. The resulting increase in the number of iterations is therefore offset by the smaller batch, up to logarithmic factors, until the batch reaches one. Taking β still closer to one then increases the query cost without reducing the batch further.

Batch size one. In particular, choose

$$\beta = 1 - \frac{1}{2K \log\left(\frac{2048\sqrt{3}\, L \bar{\Delta}_0 K}{\delta \epsilon^2}\right)}.$$

This choice lies in $[1/2, 1)$ and makes the running batch in Theorem 3.1 exactly $M = 1$. The warm-start batch remains $M_0 = \lceil \lambda_0 K \rceil$. The same high-probability guarantee holds, with total query complexity

$$Q_{\mathrm{MOM}} = \widetilde{O}\left[\left(1 + \frac{L\bar{\Delta}_0}{\epsilon^2}\right) \max\left\{ 1, \left(\frac{d S_\mu}{\epsilon}\right)^{\frac{p}{p-1}} \right\} \right].$$

Since $\mu \le \epsilon/(4Ld)$,

$$\bar{\Delta}_0 = \Delta_0 + O\left(\frac{\epsilon^2}{Ld^2}\right), \qquad dS_\mu = \sqrt{d}\big(\sqrt{L\Delta_0} + \sigma\big) + O(\epsilon).$$

Consequently, when $L\Delta_0 \ge \epsilon^2$,

$$Q_{\mathrm{MOM}} = \widetilde{O}\left[\frac{L\Delta_0}{\epsilon^2}\left(\frac{\sqrt{d}(\sqrt{L\Delta_0}+\sigma)}{\epsilon}\right)^{\frac{p}{p-1}}\right].$$

Thus momentum allows batch size one while retaining the query order of the base method in Theorem 2.3.

4 Zeroth-Order Fine-Tuning Experiments

4.1 Setup

We evaluate RSC-ZO on classification fine-tuning of RoBERTa-large (350M parameters) [Liu et al., 2019]. We consider two natural language inference tasks, SNLI [Bowman et al., 2015] and MNLI [Williams et al., 2018], and one topic classification task, TREC [Voorhees and Tice, 2000]. We follow the prompt-based fine-tuning setup of Malladi et al. [2023] for masked language models [Gao et al., 2021], using $k = 512$ training examples per task.

4.2 Results

We compare RSC-ZO against MeZO as the primary baseline. We also evaluate multi-direction variants of RSC-ZO, which average the zeroth-order estimate over $M \in \{1, 2, 4\}$ directions. Each update uses two forward passes per direction. We scale the data batch size inversely with M, keeping the total number of examples processed across these passes fixed.

Table 1 reports mean test accuracy and standard deviation over five seeds. RSC-ZO achieves substantially higher accuracy than MeZO on all three tasks. The gains are already strong for $M = 1$, showing the benefit of scalar clipping in the standard one-direction MeZO setting. Increasing M yields higher mean accuracy on MNLI and TREC, while SNLI performs best at $M = 1$. Since increasing M also reduces the data batch size, these results suggest a task-dependent tradeoff between directional averaging and data averaging under the matched evaluation budget.

We also compare LOZO [Chen et al., 2025], which uses low-rank perturbations, with its scalar-clipped variant using the same evaluation protocol. Scalar clipping improves LOZO on all three tasks.

Method	**SNLI**	**MNLI**	**TREC**
Zero-shot	50.4	49.2	32.0
MeZO	74.2 (1.5)	67.9 (2.8)	83.2 (4.3)
RSC-ZO ($M = 1$)	**83.0 (1.0)**	75.5 (0.9)	92.4 (1.2)
RSC-ZO ($M = 2$)	82.0 (0.2)	76.2 (1.5)	92.9 (1.2)
RSC-ZO ($M = 4$)	79.7 (0.7)	**76.9 (1.3)**	**93.5 (0.6)**
LOZO	80.0 (2.5)	73.2 (0.7)	89.8 (1.1)
Scalar-clipped LOZO	**83.7 (0.8)**	**76.6 (0.8)**	**94.6 (0.7)**

Table 1: Accuracy of RoBERTa-large fine-tuned with $k = 512$ examples. Results are mean accuracy with standard deviation over five seeds. Bold marks the best result within each method group.

All methods are trained for 100K steps with smoothing parameter $\mu = 10^{-3}$ and no weight decay. On SNLI, MNLI, and TREC, we use $\tau = 8$, 32, and 32 for RSC-ZO, and $\tau = 64$, 32, and 64 for scalar-clipped LOZO, respectively. Sensitivity to τ is reported in Table 3. Additional implementation details are given in Appendix E.

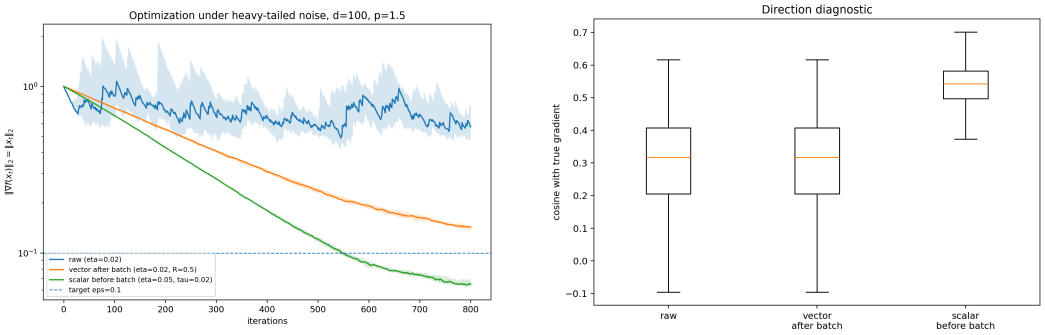

Figure 1: Controlled heavy-tailed quadratic experiment with $d = 100$, Pareto tail exponent $p = 1.5$, and $M = 256$. Left: optimization curves for $\|\nabla f(x_t)\|_2 = \|x_t\|_2$. Raw ZO is unstable, final-vector clipping stabilizes but plateaus above the target, and scalar clipping before aggregation reaches the target. Right: cosine similarity with the true gradient. Final-vector clipping has essentially the same alignment as the raw estimator because it preserves the raw direction, whereas scalar clipping substantially improves alignment.

5 Controlled experiment: clipping before versus after aggregation

We include a controlled experiment to isolate the estimator-level effect of scalar clipping. The goal is not to benchmark on a difficult objective, but to test the mechanism suggested by our theory: in zeroth-order estimation under heavy-tailed noise, it is beneficial to clip each scalar directional derivative before aggregation, rather than only clipping the final batched vector.

We consider $f(x) = \frac{1}{2}\|x\|_2^2$ and $F(x;\zeta) = \frac{1}{2}\|x\|_2^2 + \langle \zeta, x\rangle$, where the noise is sparse and heavy-tailed, $\zeta = sAe_J$. Here s is uniform on $\{-1,+1\}$, J is uniform on $\{1,\dots,d\}$, and A is Pareto with tail exponent p. Thus $\mathbb{E}[\zeta] = 0$, the population objective is f, and for $p \in (1,2)$ the stochastic gradient noise has finite first moment but infinite variance. For a two-point shared-randomness estimator,

$$Y_\ell = \frac{F(x+\mu u_\ell;\zeta_\ell) - F(x-\mu u_\ell;\zeta_\ell)}{2\mu} = \langle x, u_\ell\rangle + \langle \zeta_\ell, u_\ell\rangle.$$

We compare three natural estimators:

$$\begin{aligned} g_{\mathrm{raw}} &= \frac{d}{M}\sum_{\ell=1}^{M} Y_\ell u_\ell, && \text{unclipped ZO estimator,} \\ g_{\mathrm{vec}} &= \mathrm{clip}_{r_{\mathrm{vec}}}(g_{\mathrm{raw}}), && \text{vector clipping after aggregation,} \\ g_{\mathrm{sc}} &= \frac{d}{M}\sum_{\ell=1}^{M} \psi_\tau(Y_\ell) u_\ell, && \text{scalar clipping before aggregation.} \end{aligned}$$

The final-vector baseline is the direct analogue of standard gradient clipping in SGD: one first forms a stochastic gradient-like vector and then clips its norm. In contrast, scalar clipping acts earlier, at the level of each one-dimensional directional measurement.

Figure 1 shows that vector clipping after aggregation improves over the raw estimator, but remains substantially worse than scalar clipping before aggregation. The cosine diagnostic explains the gap: $\mathrm{clip}_{r_{\mathrm{vec}}}(g_{\mathrm{raw}})$ rescales the already-formed estimator and therefore preserves its direction, while scalar clipping prevents extreme directional derivatives from dominating the average. Additional dimension and tail-index sweeps, fair hyperparameter tuning, and outlier diagnostics are reported in Appendix D.

6 Discussion and Conclusion

We studied stochastic zeroth-order optimization for smooth nonconvex objectives under heavy-tailed sample-gradient noise. Under sample-wise smoothness and shared randomness, our analysis transfers weak-L_p control of this noise to scalar two-point directional estimates. This connects the assumption on unobserved stochastic gradients to estimators constructed from function values.

We proposed RSC-ZO, which clips each scalar directional estimate before aggregation, limiting the influence of extreme measurements. Its high-probability query complexity matches the ϵ-dependence of the general first-order heavy-tailed benchmark. At $p = 2$, it achieves the classical $d\epsilon^{-4}$ stochastic ZO dependence up to logarithmic factors, under the weaker weak-L_2 condition, which allows infinite variance. For an explicit choice of the momentum parameter, the running batch size can be reduced to one while preserving the total query order up to logarithmic factors.

In the controlled heavy-tailed quadratic experiments, clipping before aggregation improves alignment with the true gradient compared with raw ZO estimation and clipping the final batched vector. In RoBERTa-large fine-tuning, scalar clipping improves on both MeZO and LOZO, with the total number of examples processed across forward passes held fixed.

Limitations. Our guarantees require sample-wise smoothness, weak-L_p control of the sample-gradient noise, and shared randomness in the two-point estimator. They do not cover nonsmooth objectives or general function-value noise models with independent noise in the two evaluations. The theoretically prescribed clipping threshold also depends on problem-specific quantities; in our experiments, we tune it by grid search. Developing adaptive clipping rules with comparable high-probability guarantees is an important direction for future work.

The empirical evidence is limited to the tested noise model and fine-tuning tasks. A broader study across model scales, tasks, hyperparameter budgets, and robust ZO alternatives would help establish the practical scope of the method. Finally, although our bound matches the classical stochastic ZO dimension dependence at $p = 2$, its dependence on d remains a limitation in very high-dimensional problems. Combining scalar clipping with structured perturbations or low-dimensional subspace methods is a promising direction.

Acknowledgements and Disclosures

The authors gratefully acknowledge access to the computing resources of the Toubkal Supercomputer at UM6P (Morocco). The authors declare no competing interests. No specific funding supported this work.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015.

El Mahdi Chayti. RanSOM: Second-order momentum with randomized scaling for constrained and unconstrained optimization. *arXiv:2602.06824*, 2026. URL `https://arxiv.org/abs/2602.06824`.

El Mahdi Chayti, Nikita Doikov, and Martin Jaggi. Improving stochastic cubic newton with momentum. *arXiv: 2410.19644*, 2025a. URL `https://arxiv.org/abs/2410.19644`.

El Mahdi Chayti, Taha El Bakkali El Kadi, Omar Saadi, and Martin Jaggi. Stochastic optimization with random search. *arXiv preprint arXiv:2510.15610*, 2025b.

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 (AISec)*, pages 15–26, 2017.

Yiming Chen, Yuan Zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order fine-tuning for language models with low-rank structures. In *International Conference on Learning Representations (ICLR)*, 2025.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895, 2021.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

Michael C Fu. *Handbook of simulation optimization*, volume 216. Springer, 2015.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, 2021.

Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, J. Zico Kolter, Zachary C. Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

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

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *arXiv preprint arXiv:2410.13849*, 2024.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Nikita Kornilov, Alexander Gasnikov, Pavel Dvurechensky, and Darina Dvinskikh. Gradient-free methods for non-smooth convex stochastic optimization with heavy-tailed noise on convex compact. *Computational Management Science*, 20(1):37, 2023a.

Nikita Kornilov, Ohad Shamir, Aleksandr Lobanov, Darina Dvinskikh, Alexander Gasnikov, Innokentiy Shibaev, Eduard Gorbunov, and Samuel Horváth. Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023b.

Nikita Kornilov, Yuriy Dorn, Aleksandr Lobanov, Nikolay Kutuzov, Innokentiy Shibaev, Eduard Gorbunov, Alexander Gasnikov, and Alexander Nazin. Median clipping for zeroth-order non-smooth convex optimization and multi-armed bandit problem with heavy-tailed symmetric noise. In *arXiv preprint arXiv:2402.02461*, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations (ICLR)*, 2025.

Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *Proceedings of the 36th Annual Conference on Learning Theory (COLT)*, 2023.

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

Erfan Mirzaei, Andreas Maurer, Vladimir R Kostic, and Massimiliano Pontil. An empirical bernstein inequality for dependent data in hilbert spaces and applications. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 5158–5166. PMLR, 03–05 May 2025.

Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. *Advances in Neural Information Processing Systems*, 36:24191–24222, 2023.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: The case of unbounded variance. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Pierre Tarres and Yuan Yao. Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9): 5716–5735, 2014.

Ellen M Voorhees and Dawn M Tice. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J. Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

A Weak-L_2 Noise Can Have Infinite Variance

A natural question is whether the specialization $p = 2$ in Assumption 1.3 simply recovers the classical bounded-variance setting. Indeed, by Markov's inequality, any bounded-variance model satisfies Assumption 1.3 with $p = 2$. More precisely, if

$$\mathbb{E}_\xi \|\nabla F(x;\xi) - \nabla f(x)\|_2^2 \le \sigma^2 \quad \text{for all } x \in \mathbb{R}^d,$$

then

$$\mathbb{P}(\|\nabla F(x;\xi) - \nabla f(x)\|_2 > t) \le \frac{\sigma^2}{t^2} \quad \text{for all } x \in \mathbb{R}^d,\ t > 0.$$

Thus, the bounded-variance regime is contained in our $p = 2$ framework. The converse, however, is false: even at $p = 2$, Assumption 1.3 still allows stochastic models with infinite gradient-noise variance. Therefore, the $p = 2$ case of our theory is not a reformulation of bounded-variance results, but a genuine extension beyond them.

Proposition A.1 (The case $p = 2$ still allows infinite variance)**.** *Let $f_0 : \mathbb{R}^d \to \mathbb{R}$ be any L-smooth function, and let Z be a real-valued random variable with density*

$$\rho(z) = |z|^{-3}\mathbf{1}_{\{|z|\ge 1\}}.$$

Define the stochastic objective

$$F(x;\xi) := f_0(x) + Z(\xi)\, x_1,$$

where x_1 denotes the first coordinate of x, and let

$$f(x) := \mathbb{E}_\xi[F(x;\xi)].$$

Then the following hold:

1. *$f(x) = f_0(x)$, and for $\mathbb{P}$-almost every ξ, the map $x \mapsto F(x;\xi)$ is L-smooth.*
2. *The sample gradient noise satisfies*

   $$\nabla F(x;\xi) - \nabla f(x) = Z(\xi)e_1,$$

   where $e_1 = (1, 0, \dots, 0)^\top$.
3. *Assumption 1.3 holds with $p = 2$ and $\sigma = 1$, since for every $t > 0$,*

   $$\mathbb{P}(\|\nabla F(x;\xi) - \nabla f(x)\|_2 > t) = \mathbb{P}(|Z| > t) \le t^{-2}.$$

4. *The second moment is infinite:*

   $$\mathbb{E}_\xi \|\nabla F(x;\xi) - \nabla f(x)\|_2^2 = \mathbb{E}[Z^2] = \infty.$$

Proof of Proposition A.1. Since ρ is symmetric, Z has zero mean whenever the first moment exists. Moreover,

$$\mathbb{E}|Z| = \int_{\mathbb{R}} |z|\rho(z)\, dz = 2\int_1^\infty z \cdot z^{-3}\, dz = 2\int_1^\infty z^{-2}\, dz < \infty,$$

so $\mathbb{E}[Z] = 0$. Therefore,

$$f(x) = \mathbb{E}[F(x;\xi)] = f_0(x) + \mathbb{E}[Z]\, x_1 = f_0(x).$$

Also,

$$\nabla F(x;\xi) = \nabla f_0(x) + Z(\xi)e_1,$$

and hence

$$\nabla F(x;\xi) - \nabla f(x) = Z(\xi)e_1.$$

Because the added term $Z(\xi)x_1$ is linear in x, it does not affect smoothness; therefore $F(\cdot;\xi)$ is L-smooth whenever f_0 is L-smooth.

Since

$$\|\nabla F(x;\xi) - \nabla f(x)\|_2 = |Z(\xi)|,$$

for every $t \geq 1$,

$$\mathbb{P}(|Z| > t) = 2\int_t^\infty z^{-3}\,dz = t^{-2},$$

while for $0 < t < 1$ one trivially has $\mathbb{P}(|Z| > t) \leq 1 \leq t^{-2}$. Thus

$$\forall t > 0, \qquad \mathbb{P}(\|\nabla F(x;\xi) - \nabla f(x)\|_2 > t) \leq t^{-2},$$

so Assumption 1.3 holds with $p = 2$ and $\sigma = 1$.

Finally,

$$\mathbb{E}[Z^2] = \int_{\mathbb{R}} z^2\rho(z)\,dz = 2\int_1^\infty z^2 \cdot z^{-3}\,dz = 2\int_1^\infty z^{-1}\,dz = \infty.$$

Hence the gradient noise has infinite variance. □

Takeaway. The case $p = 2$ is not simply a restatement of bounded-variance theory. It strictly enlarges the noise model while preserving, up to logarithmic factors, the benchmark complexity

$$O\big(d\,\sigma^2\,\epsilon^{-4}\big)$$

known for stochastic zeroth-order methods in the *individual smoothness* setting under a bounded-variance assumption on the sample gradient noise [Ghadimi and Lan, 2013, Chayti et al., 2025b]. In particular, Theorem B.11 achieves the benchmark complexity bound $O(d\,\sigma^2\,\epsilon^{-4})$ up to logarithmic factors under the weaker polynomial-tail assumption (see Remark B.12).

B High-Probability Analysis of Base RSC-ZO

Notation. Define $f_* := \inf_{x\in\mathbb{R}^d} f(x)$, $\Delta_0 := f(x_0) - f_*$, and $\bar{\Delta}_0 := \Delta_0 + \frac{L\mu^2}{2}$. For brevity, write

$$S_\mu := \frac{\sqrt{L\bar{\Delta}_0} + \sigma}{\sqrt{d}} + L\mu.$$

Goal of this section. Our goal is to prove a high-probability bound on the estimator error

$$\|g(x) - \nabla f_\mu(x)\|_2,$$

where

$$Y_\ell := \frac{F(x+\mu u_\ell;\xi_\ell) - F(x-\mu u_\ell;\xi_\ell)}{2\mu} \quad \text{and} \quad g(x) := \frac{d}{M}\sum_{\ell=1}^M \psi_\tau(Y_\ell)u_\ell.$$

We decompose the estimator error as

$$g(x) - \nabla f_\mu(x) = \underbrace{(g(x) - \mathbb{E}[g(x)])}_{A} + \underbrace{(\mathbb{E}[g(x)] - \nabla f_\mu(x))}_{B}.$$

Here, A is the empirical fluctuation of the clipped sample average, while B is the bias introduced by clipping. Hence, to obtain a high-probability upper bound on $\|g(x) - \nabla f_\mu(x)\|_2$, it suffices to bound $\|A\|_2$ and $\|B\|_2$ separately. We first consider B.

Let $Y := \frac{F(x+\mu u;\xi) - F(x-\mu u;\xi)}{2\mu}$, where $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ is independent of ξ. Then

$$B = \mathbb{E}[g(x)] - \nabla f_\mu(x) = d\,\mathbb{E}\big[(\psi_\tau(Y) - Y)u\big].$$

Therefore,

$$\|B\|_2 = d\,\Big\|\mathbb{E}\big[(\psi_\tau(Y) - Y)u\big]\Big\|_2 \leq d\,\mathbb{E}\big[|\psi_\tau(Y) - Y|\big].$$

To control the quantity $\mathbb{E}[|\psi_\tau(Y) - Y|]$, we first establish the following abstract clipping-bias bound.

Lemma B.1 (Clipping bias under weak-L_p control)**.** *Let Z be a real-valued random variable such that, for some $S > 0$,*

$$\forall t \geq 4S, \qquad \mathbb{P}(|Z| > t) \leq 2^{2p+1}S^p t^{-p}.$$

Then, for every $\tau \geq 4S$,

$$\mathbb{E}\big[|Z - \psi_\tau(Z)|\big] \leq \frac{2^{2p+1}}{p-1}\,S^p\tau^{1-p}.$$

It remains to verify that the directional estimator Y satisfies the hypothesis of Lemma B.1. To this end, we use the decomposition

$$Y = \big(Y - D_\mu f(x,u)\big) + D_\mu f(x,u).$$

Thus, it is enough to control the tails of $Y - D_\mu f(x,u)$ and $D_\mu f(x,u)$. The first term measures the deviation of the stochastic directional estimator around its conditional mean, while the second is the directional signal itself. These are controlled in Lemmas B.2 and B.3, respectively.

Lemma B.2 (Directional weak-L_p bound)**.** *Assume that Assumptions 1.2 and 1.3 hold. Let $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, independent of ξ. Then, for every fixed $x \in \mathbb{R}^d$ and every $s \geq 1$,*

$$\mathbb{P}_{u,\xi}\left(\left|\frac{F(x+\mu u;\xi) - F(x-\mu u;\xi)}{2\mu} - D_\mu f(x,u)\right| > \left(\frac{\sigma}{\sqrt{d}} + L\mu\right) s\right) \leq s^{-p}.$$

Lemma B.3 (Localized weak-L_p bound for the noiseless directional signal)**.** *Assume that Assumptions 1.1 and 1.2 hold. Let $p \in (1,2]$. Let $x \in \mathbb{R}^d$ satisfy $f(x) - f_* \leq 4\bar{\Delta}_0$, and let $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. Then, for every $s \geq 1$,*

$$\mathbb{P}_u\left(|D_\mu f(x,u)| > \left(\frac{\sqrt{8L\bar{\Delta}_0}}{\sqrt{d}} + L\mu\right) s\right) \leq s^{-p}.$$

Lemma B.4 (Localized weak-L_p bound for the shared-randomness directional estimator)**.** *Assume that Assumptions 1.1, 1.2, and 1.3 hold. Let $x \in \mathbb{R}^d$ satisfy $f(x) - f_* \leq 4\bar{\Delta}_0$. Let $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, independent of ξ, and define $Y := \frac{F(x+\mu u;\xi) - F(x-\mu u;\xi)}{2\mu}$. Then, for every $s \geq 1$,*

$$\mathbb{P}(|Y| > 4S_\mu s) \leq 2s^{-p}.$$

Equivalently, for every $t \geq 4S_\mu$,

$$\mathbb{P}(|Y| > t) \leq 2^{2p+1} S_\mu^p t^{-p}.$$

We may now combine the preceding ingredients to obtain the following deterministic bound on the clipping bias term B.

Lemma B.5 (Bias induced by clipping)**.** *Assume that Assumptions 1.1, 1.2, and 1.3 hold. Let $x \in \mathbb{R}^d$ satisfy $f(x) - f_* \leq 4\bar{\Delta}_0$. Let $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, independent of ξ, and set $Y := \frac{F(x+\mu u;\xi) - F(x-\mu u;\xi)}{2\mu}$. Then, for every $\tau \geq 4S_\mu$,*

$$\Big\| d\,\mathbb{E}[\psi_\tau(Y)u] - \nabla f_\mu(x) \Big\|_2 \leq \frac{2^{2p+1}}{p-1}\, d\, S_\mu^p\, \tau^{1-p}.$$

Equivalently, if $B := \mathbb{E}[g(x)] - \nabla f_\mu(x)$, then

$$\|B\|_2 \leq \frac{2^{2p+1}}{p-1}\, d\, S_\mu^p\, \tau^{1-p}.$$

We now turn to the empirical fluctuation term

$$A = \frac{d}{M} \sum_{\ell=1}^{M} \Big(\psi_\tau(Y_\ell) u_\ell - \mathbb{E}[\psi_\tau(Y_\ell) u_\ell] \Big).$$

The summands are i.i.d. centered random vectors. We therefore control A using the vector Bernstein inequality of Lemma B.16. To apply that lemma, it remains to verify an almost-sure bound on each summand and a suitable second-moment bound. For each ℓ, define

$$X_\ell := \frac{d}{M} \Big(\psi_\tau(Y_\ell) u_\ell - \mathbb{E}[\psi_\tau(Y_\ell) u_\ell] \Big).$$

Since $\|u_\ell\|_2 = 1$ almost surely and $|\psi_\tau(Y_\ell)| \leq \tau$, we have

$$\|X_\ell\|_2 \leq \frac{d}{M} \Big(|\psi_\tau(Y_\ell)| + \|\mathbb{E}[\psi_\tau(Y_\ell) u_\ell]\|_2 \Big) \leq \frac{2d\tau}{M}.$$

This verifies the first condition in Lemma B.16.

To verify the second condition, we use the following clipped second-moment estimate.

Lemma B.6 (Clipped second moment under weak-L_p control)**.** *Let Z be a real-valued random variable such that, for some $S > 0$,*

$$\forall t \geq 4S, \quad \mathbb{P}(|Z| > t) \leq 2^{2p+1} S^p t^{-p}.$$

Then, for every $1 < p \leq 2$ and every $\tau \geq 4S$,

$$\mathbb{E}[\psi_\tau(Z)^2] \leq 64\, S^p \tau^{2-p} \big(1 + \log(\tau/S)\big).$$

We are now ready to combine the vector Bernstein inequality with the clipped second-moment estimate to control the empirical fluctuation term $A = g(x) - \mathbb{E}[g(x)]$. This yields the following high-probability bound.

Lemma B.7 (Empirical fluctuation of the clipped estimator)**.** *Assume that Assumptions 1.1, 1.2, and 1.3 hold. Let $x \in \mathbb{R}^d$ satisfy $f(x) - f_* \leq 4\Delta_0$. Let $\rho \in (0,1)$, and let $\tau \geq 4S_\mu$. For $\ell = 1, \dots, M$, let*

$$Y_\ell := \frac{F(x + \mu u_\ell; \xi_\ell) - F(x - \mu u_\ell; \xi_\ell)}{2\mu}, \qquad \widetilde{Y}_\ell := \psi_\tau(Y_\ell),$$

where the pairs (u_ℓ, ξ_ℓ) are independent across ℓ, with $u_\ell \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and $\xi_\ell \sim \mathbb{P}$. Define

$$A := \frac{d}{M} \sum_{\ell=1}^{M} \Big(\widetilde{Y}_\ell u_\ell - \mathbb{E}[\widetilde{Y}_\ell u_\ell] \Big).$$

Then, with probability at least $1 - \rho$,

$$\|A\|_2 \leq \frac{8d}{\sqrt{M}} S_\mu^{p/2} \tau^{1-\frac{p}{2}} \sqrt{1 + \log\left(\frac{\tau}{S_\mu}\right)} \left(1 + \sqrt{2\log(1/\rho)}\right) + \frac{8d\tau}{3M} \log(1/\rho).$$

Combining the fluctuation and bias bounds yields the following localized high-probability estimate for the full estimator error $g(x) - \nabla f_\mu(x)$.

Corollary B.8 (Localized high-probability deviation of the clipped estimator)**.** *Assume that Assumptions 1.1, 1.2, and 1.3 hold. Fix $\delta \in (0,1)$ and $T \geq 1$, and set $\lambda := \log \frac{T}{\delta}$. Let $x \in \mathbb{R}^d$ satisfy $f(x) - f_* \leq 4\bar{\Delta}_0$. Let $\tau \geq 4S_\mu$. For $\ell = 1, \dots, M$, let*

$$Y_\ell := \frac{F(x + \mu u_\ell; \xi_\ell) - F(x - \mu u_\ell; \xi_\ell)}{2\mu}, \qquad g(x) := \frac{d}{M} \sum_{\ell=1}^{M} \psi_\tau(Y_\ell) u_\ell,$$

where the pairs (u_ℓ, ξ_ℓ) are independent across ℓ, with $u_\ell \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and $\xi_\ell \sim \mathbb{P}$. Then, with probability at least $1 - \delta/T$,

$$\|g(x) - \nabla f_\mu(x)\|_2 \leq \frac{8d}{\sqrt{M}} S_\mu^{p/2} \tau^{1-\frac{p}{2}} \sqrt{1 + \log\left(\frac{\tau}{S_\mu}\right)} \left(1 + \sqrt{2\lambda}\right) + \frac{8d\tau}{3M}\lambda + \frac{2^{2p+1}}{p-1}\, d S_\mu^p \tau^{1-p}.$$

We now instantiate Corollary B.8 with the clipping level used in the algorithm. This choice balances the leading fluctuation and bias terms up to logarithmic factors and gives the following simplified localized deviation bound.

Corollary B.9 (Appendix form of Theorem 2.1)**.** *Assume that Assumptions 1.1, 1.2, and 1.3 hold. Fix $\delta \in (0,1]$ and $T \geq 3$, and set $\lambda := \log \frac{T}{\delta}$. Let $x \in \mathbb{R}^d$ satisfy $f(x) - f_* \leq 4\bar{\Delta}_0$. Suppose that $M \geq \lambda$, and choose $\tau := 8S_\mu \left(\frac{M}{\lambda}\right)^{1/p}$. For $\ell = 1, \dots, M$, let*

$$Y_\ell := \frac{F(x + \mu u_\ell; \xi_\ell) - F(x - \mu u_\ell; \xi_\ell)}{2\mu}, \qquad g(x) := \frac{d}{M} \sum_{\ell=1}^{M} \psi_\tau(Y_\ell) u_\ell.$$

Then, with probability at least $1 - \delta/T$,

$$\|g(x) - \nabla f_\mu(x)\|_2 \leq \left(120 + \frac{2^{4-p}}{p-1}\right) d S_\mu \left(\frac{\log(T/\delta)}{M}\right)^{\frac{p-1}{p}} \sqrt{1 + \log\left(\frac{M}{\log(T/\delta)}\right)}.$$

We now turn the localized deviation bound of Corollary B.9 into a descent estimate for the smoothed objective f_μ. Since the update uses the clipped estimator $g(x_t)$ in place of $\nabla f_\mu(x_t)$, the key step is to combine the smoothness inequality for f_μ with the high-probability control of $g(x_t) - \nabla f_\mu(x_t)$. The next proposition gives a deterministic one-step descent inequality.

Proposition B.10 (One-step descent inequality)**.** *Assume that f is L-smooth. Let $x \in \mathbb{R}^d$, let $x^+ := x - \alpha g(x)$, and define $e(x) := g(x) - \nabla f_\mu(x)$. If $\alpha \le \frac{1}{4L}$, then*

$$f_\mu(x^+) \le f_\mu(x) - \frac{\alpha}{2}\,\|\nabla f_\mu(x)\|_2^2 + \frac{5\alpha}{4}\,\|e(x)\|_2^2\,.$$

For $M \in \mathbb{N}$, $\mu > 0$, $\delta \in (0,1)$, and $T \ge 3$, define $C_p := 120 + \frac{2^{4-p}}{p-1}$,

$$\eta_0(M,\mu,\delta,T) := C_p\, dS_\mu \left(\frac{\log(T/\delta)}{M}\right)^{\frac{p-1}{p}} \sqrt{1 + \log\!\left(\frac{M}{\log(T/\delta)}\right)},$$

and $\eta(M,\mu,\delta,T) := \eta_0(M,\mu,\delta,T) + L\mu$.

Theorem B.11 (High-probability convergence of base RSC-ZO)**.** *Assume that Assumptions 1.1, 1.2, and 1.3 hold. Let $\mu > 0$, $\Delta_0 := f(x_0) - f_*$ and $\bar\Delta_0 := \Delta_0 + \frac{L\mu^2}{2}$. Let $\epsilon > 0$, $\delta \in (0,1]$, and $M \in \mathbb{N}$, and define $T := \max\left\{3, \left\lceil \frac{32L\bar\Delta_0}{\epsilon^2} \right\rceil\right\}$. Let $\{x_t\}_{t\ge 0}$ be generated by Algorithm 1 with $\beta = 0$, batch size M, stepsize $\alpha = 1/(4L)$, and threshold $\tau := 8S_\mu\Big(\frac{M}{\log(T/\delta)}\Big)^{1/p}$. Suppose that*

$$M \ge \log\frac{T}{\delta}, \qquad \eta(M,\mu,\delta,T) \le \frac{\epsilon}{4}, \qquad \epsilon^2 \le 32L\bar\Delta_0.$$

Then, with probability at least $1-\delta$,

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2 \le \epsilon^2.$$

In particular, on the same event,

$$\min_{0\le t<T} \|\nabla f(x_t)\|_2 \le \epsilon.$$

Remark B.12 (Complexity implied by Theorem B.11)**.** The total number of stochastic function queries is $2MT$. Choose $\mu = \epsilon/(8Ld)$. Then $L\mu \le \epsilon/8$, $dL\mu = O(\epsilon)$, and $\bar\Delta_0 = \Delta_0 + O\Big(\frac{\epsilon^2}{Ld^2}\Big)$. Since

$$T = \max\left\{3, \left\lceil \frac{32L\bar\Delta_0}{\epsilon^2} \right\rceil\right\},$$

we have $T = O(1 + L\Delta_0/\epsilon^2)$.

Up to logarithmic factors, the batch condition $\eta(M,\mu,\delta,T) \le \epsilon/4$ requires

$$dS_\mu \left(\frac{\log(T/\delta)}{M}\right)^{\frac{p-1}{p}} \lesssim \epsilon.$$

Under the above smoothing choice,

$$dS_\mu = \sqrt{d}\,(\sqrt{L\bar\Delta_0} + \sigma) + dL\mu = \sqrt{d}\,(\sqrt{L\Delta_0} + \sigma) + O(\epsilon).$$

Thus the conditions $M \ge \log(T/\delta)$ and $\eta(M,\mu,\delta,T) \le \epsilon/4$ can be satisfied with a batch size

$$M = \widetilde{\mathcal{O}}\left(\max\left\{1, \left(\frac{\sqrt{d}\,(\sqrt{L\Delta_0} + \sigma)}{\epsilon}\right)^{\frac{p}{p-1}}\right\}\right).$$

Here $\widetilde{\mathcal{O}}$ suppresses logarithmic factors and constants depending only on p. Consequently,

$$2MT = \widetilde{\mathcal{O}}\left[\left(1 + \frac{L\Delta_0}{\epsilon^2}\right)\max\left\{1, \left(\frac{\sqrt{d}\,(\sqrt{L\Delta_0} + \sigma)}{\epsilon}\right)^{\frac{p}{p-1}}\right\}\right].$$

In particular, when $L\Delta_0 \geq \epsilon^2$, this simplifies to

$$2MT = \widetilde{\mathcal{O}}\left(L\Delta_0(\sqrt{L\Delta_0} + \sigma)^{\frac{p}{p-1}} \frac{d^{\frac{p}{2(p-1)}}}{\epsilon^{\frac{3p-2}{p-1}}} \right).$$

At $p = 2$, in the same regime, the bound becomes

$$2MT = \widetilde{\mathcal{O}}\Big(L\Delta_0(\sqrt{L\Delta_0} + \sigma)^2 d\,\epsilon^{-4} \Big).$$

This recovers the classical $d\,\epsilon^{-4}$ dimension–accuracy dependence up to logarithmic and problem-dependent factors, with a high-probability guarantee under weak-L_2 noise.

Remark B.13 (Trivial regime)**.** If $\epsilon^2 > 32L\bar{\Delta}_0$, the initial point already satisfies the desired stationarity condition. Indeed, Lemma B.14 gives

$$\|\nabla f(x_0)\|_2^2 \leq 2L(f(x_0) - f_*) = 2L\Delta_0 \leq 2L\bar{\Delta}_0 < \epsilon^2.$$

Thus no optimization steps are needed in this regime.

The proofs of the lemmas, corollaries, and theorem in this section are collected in Appendix B.1.

B.1 Proofs

Lemma B.14 (Gradient bound from suboptimality)**.** *Assume that f is L-smooth and bounded below by f_*. Then, for every $x \in \mathbb{R}^d$,*

$$\|\nabla f(x)\|_2^2 \leq 2L\big(f(x) - f_*\big).$$

Proof. For any $x \in \mathbb{R}^d$, L-smoothness gives

$$f\Big(x - \frac{1}{L}\nabla f(x)\Big) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2.$$

Since $f_* \leq f(y)$ for every y, we obtain

$$f_* \leq f\Big(x - \frac{1}{L}\nabla f(x)\Big) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2,$$

which rearranges to the claim. □

Theorem B.15 ([Mirzaei et al., 2025, Theorem 5])**.** *Suppose that the X_i are m independent mean-zero random variables with values in a Hilbert space H, satisfying $\|X_i\| \leq c$. Then for $\delta \in (0, 1)$,*

$$\Pr\left\{ \left\| \sum_i X_i \right\| > \sqrt{\sum_i \mathbb{E}\|X_i\|^2} \left(1 + \sqrt{2\ln(1/\delta)}\right) + \frac{4c}{3}\ln(1/\delta) \right\} < \delta.$$

The following lemma is the finite-dimensional form used in our analysis.

Lemma B.16 (Vector Bernstein inequality with variance proxy)**.** *Let $X_1, \ldots, X_M$ be independent mean-zero random vectors in $\mathbb{R}^d$. Assume that*

$$\|X_\ell\|_2 \leq R \quad \textit{almost surely for all } \ell,$$

and

$$\sum_{\ell=1}^{M} \mathbb{E}\,\|X_\ell\|_2^2 \leq V.$$

Then, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\left\| \sum_{\ell=1}^{M} X_\ell \right\|_2 \leq \sqrt{V}\left(1 + \sqrt{2\log(1/\delta)}\right) + \frac{4R}{3}\log(1/\delta).$$

Proof. We can apply Theorem B.15 with $H = \mathbb{R}^d$, $m = M$, and $c = R$. The theorem gives

$$\left\| \sum_{\ell=1}^{M} X_\ell \right\|_2 \le \sqrt{\sum_{\ell=1}^{M} \mathbb{E}\,\|X_\ell\|_2^2} \left(1 + \sqrt{2\log(1/\delta)}\right) + \frac{4R}{3}\log(1/\delta)$$

with probability at least $1 - \delta$. Since $\sum_{\ell=1}^{M} \mathbb{E}\,\|X_\ell\|_2^2 \le V$, the result follows. □

Lemma B.17 (Smoothing bias)**.** *Assume that f is L-smooth. Then, for every $x \in \mathbb{R}^d$ and every $\mu > 0$,*

$$|f_\mu(x) - f(x)| \le \frac{L\mu^2}{2} \quad \textit{and} \quad \|\nabla f_\mu(x) - \nabla f(x)\|_2 \le L\mu.$$

Proof. By L-smoothness, for every $v \in \mathbb{B}^d$,

$$|f(x + \mu v) - f(x) - \langle \nabla f(x), \mu v\rangle| \le \frac{L\mu^2}{2}\|v\|^2 \le \frac{L\mu^2}{2}.$$

Taking expectation over $v \sim \mathrm{Unif}(\mathbb{B}^d)$ and using $\mathbb{E}[v] = 0$, we obtain

$$|f_\mu(x) - f(x)| \le \frac{L\mu^2}{2}.$$

Next, since f is L-smooth,

$$\nabla f_\mu(x) = \mathbb{E}_{v\sim\mathrm{Unif}(\mathbb{B}^d)}[\nabla f(x + \mu v)].$$

Therefore,

$$\|\nabla f_\mu(x) - \nabla f(x)\|_2 \le \mathbb{E}_{v\sim\mathrm{Unif}(\mathbb{B}^d)} \|\nabla f(x + \mu v) - \nabla f(x)\|_2 \le L\mu\,\mathbb{E}\,\|v\|_2 \le L\mu,$$

since $\|v\|_2 \le 1$ almost surely for $v \in \mathbb{B}^d$. □

Proof of Lemma B.1. We have $\psi_\tau(z) = \mathrm{sgn}(z)\min\{|z|, \tau\}$. Therefore

$$|Z - \psi_\tau(Z)| = \big(|Z| - \tau\big)_+,$$

where $a_+ := \max\{a, 0\}$.

For any nonnegative random variable X and any $a > 0$,

$$\mathbb{E}[(X - a)_+] = \mathbb{E}\left[\int_a^\infty 1_{\{X>t\}}\, dt\right] = \int_a^\infty \mathbb{P}(X > t)\, dt.$$

Applying this with $X = |Z|$ and $a = \tau$, we obtain

$$\mathbb{E}\big[|Z - \psi_\tau(Z)|\big] = \int_\tau^\infty \mathbb{P}(|Z| > t)\, dt.$$

Since $\tau \ge 4S$, the assumed tail bound is valid for every $t \ge \tau$, so

$$\mathbb{E}\big[|Z - \psi_\tau(Z)|\big] \le 2^{2p+1} S^p \int_\tau^\infty t^{-p}\, dt.$$

Since $p > 1$, we have $\int_\tau^\infty t^{-p}\, dt = \frac{\tau^{1-p}}{p-1}$. Hence

$$\mathbb{E}\big[|Z - \psi_\tau(Z)|\big] \le \frac{2^{2p+1}}{p-1} S^p \tau^{1-p}.$$

□

Proof of Lemma B.2. Fix x, and condition on x. By the fundamental theorem of calculus,

$$\frac{F(x+\mu u;\xi)-F(x-\mu u;\xi)}{2\mu}=\frac{1}{2}\int_{-1}^{1}\big\langle\nabla F(x+\mu su;\xi),u\big\rangle\,ds.$$

Similarly,

$$\frac{f(x+\mu u)-f(x-\mu u)}{2\mu}=\frac{1}{2}\int_{-1}^{1}\big\langle\nabla f(x+\mu su),u\big\rangle\,ds.$$

Hence

$$\frac{F(x+\mu u;\xi)-F(x-\mu u;\xi)}{2\mu}-D_\mu f(x,u)=\big\langle\nabla F(x;\xi)-\nabla f(x),u\big\rangle+R(x,u,\xi),$$

where

$$\begin{aligned}R(x,u,\xi):=&\frac{1}{2}\int_{-1}^{1}\big\langle\nabla F(x+\mu su;\xi)-\nabla F(x;\xi),u\big\rangle\,ds\\&-\frac{1}{2}\int_{-1}^{1}\big\langle\nabla f(x+\mu su)-\nabla f(x),u\big\rangle\,ds.\end{aligned}$$

By Assumption 1.2 and Remark 1.5, both $F(\cdot;\xi)$ and f are L-smooth. Therefore

$$|R(x,u,\xi)|\le\frac{1}{2}\int_{-1}^{1}L\mu|s|\,ds+\frac{1}{2}\int_{-1}^{1}L\mu|s|\,ds=L\mu.$$

Thus

$$\left|\frac{F(x+\mu u;\xi)-F(x-\mu u;\xi)}{2\mu}-D_\mu f(x,u)\right|\le|\langle\zeta,u\rangle|+L\mu\quad\text{where}\quad\zeta:=\nabla F(x;\xi)-\nabla f(x).$$

Define

$$v:=\begin{cases}\zeta/\|\zeta\|_2, & \zeta\neq0,\\ e_1, & \zeta=0,\end{cases}$$

where e_1 is the first canonical basis vector, and let

$$V:=\sqrt{d}\,|U_1|,\quad\text{where}\quad U\sim\mathrm{Unif}(\mathbb{S}^{d-1}),$$

is chosen independently of ξ. Since $u\sim\mathrm{Unif}(\mathbb{S}^{d-1})$, by rotational invariance we have, conditionally on x and ξ,

$$\sqrt{d}\,|\langle v,u\rangle|\overset{d}{=}V.$$

Since $|\langle\zeta,u\rangle|=\frac{\|\zeta\|_2}{\sqrt{d}}\sqrt{d}\,|\langle v,u\rangle|$, it follows that, conditionally on x and ξ,

$$|\langle\zeta,u\rangle|\overset{d}{=}\frac{\|\zeta\|_2}{\sqrt{d}}\,V.$$

Moreover, since $1<p\le2$, $\mathbb{E}[V^p]=d^{p/2}\mathbb{E}[|U_1|^p]\le d^{p/2}\big(\mathbb{E}[U_1^2]\big)^{p/2}=1$. Therefore, for every $r>0$,

$$\begin{aligned}\Pr(|\langle\zeta,u\rangle|>r\mid x)&=\mathbb{E}[\Pr(|\langle\zeta,u\rangle|>r\mid x,\xi)|\,x]\\&=\mathbb{E}\left[\Pr\left(\frac{\|\zeta\|_2}{\sqrt{d}}V>r\;\middle|\;x,\xi\right)\middle|\,x\right]\\&=\mathbb{E}\left[\Pr\left(\|\zeta\|_2>\frac{r\sqrt{d}}{V}\;\middle|\;x,V\right)\middle|\,x\right].\end{aligned}$$

Since V is independent of ξ, Assumption 1.3 yields

$$\Pr\left(\|\zeta\|_2>\frac{r\sqrt{d}}{V}\;\middle|\;x,V\right)\le\left(\frac{\sigma V}{r\sqrt{d}}\right)^p.$$

Hence

$$\Pr(|\langle \zeta, u\rangle| > r \mid x) \le \mathbb{E}\left[\left(\frac{\sigma V}{r\sqrt{d}}\right)^p \,\middle|\, x\right] = \left(\frac{\sigma}{r\sqrt{d}}\right)^p \mathbb{E}[V^p] \le \frac{\sigma^p}{r^p d^{p/2}}.$$

Finally, for every $s \ge 1$, since $L\mu \le L\mu s$, we have

$$\begin{aligned}
&\Pr\left(\left|\frac{F(x+\mu u;\xi) - F(x-\mu u;\xi)}{2\mu} - D_\mu f(x,u)\right| > \left(\frac{\sigma}{\sqrt{d}} + L\mu\right) s \,\middle|\, x\right) \\
&\le \Pr\left(|\langle \zeta, u\rangle| + L\mu > \left(\frac{\sigma}{\sqrt{d}} + L\mu\right) s \,\middle|\, x\right) \\
&\le \Pr\left(|\langle \zeta, u\rangle| > \frac{\sigma}{\sqrt{d}}\, s \,\middle|\, x\right) \\
&\le \left(\frac{\sigma}{(\sigma/\sqrt{d}) s\,\sqrt{d}}\right)^p = s^{-p}.
\end{aligned}$$

□

Proof of Lemma B.3. Fix $x \in \mathbb{R}^d$ such that $f(x) - f_* \le 4\bar{\Delta}_0$, and let $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. By the fundamental theorem of calculus,

$$D_\mu f(x,u) = \frac{1}{2}\int_{-1}^{1} \big\langle \nabla f(x+\mu s u), u\big\rangle\, ds = \langle \nabla f(x), u\rangle + R_\mu(x,u),$$

where

$$R_\mu(x,u) := \frac{1}{2}\int_{-1}^{1} \big\langle \nabla f(x+\mu s u) - \nabla f(x), u\big\rangle\, ds.$$

Since f is L-smooth by Remark 1.5,

$$|R_\mu(x,u)| \le \frac{1}{2}\int_{-1}^{1} L\mu|s|\, ds = \frac{L\mu}{2}.$$

Set $g := \nabla f(x)$. By Lemma B.14 and the localization $f(x) - f_* \le 4\bar{\Delta}_0$,

$$\|g\|_2^2 \le 2L(f(x) - f_*) \le 8L\bar{\Delta}_0.$$

If $g = 0$, then $|D_\mu f(x,u)| \le \frac{L\mu}{2} \le L\mu s$ for every $s \ge 1$, so the claim is immediate.

Assume now that $g \ne 0$, and define $v := \frac{g}{\|g\|_2}$, where $Z := \sqrt{d}\,|\langle v, u\rangle|$. By rotational invariance,

$$Z \overset{d}{=} \sqrt{d}\,|U_1|, \quad \text{where} \quad U \sim \mathrm{Unif}(\mathbb{S}^{d-1}).$$

Hence, since $1 < p \le 2$,

$$\mathbb{E}[Z^p] = d^{p/2}\mathbb{E}[|U_1|^p] \le d^{p/2}\big(\mathbb{E}[U_1^2]\big)^{p/2} = 1.$$

Therefore, for every $r > 0$, Markov's inequality gives

$$\mathbb{P}_u(|\langle g, u\rangle| > r) = \mathbb{P}_u\left(Z > \frac{r\sqrt{d}}{\|g\|_2}\right) \le \left(\frac{\|g\|_2}{r\sqrt{d}}\right)^p \mathbb{E}[Z^p] \le \left(\frac{\|g\|_2}{r\sqrt{d}}\right)^p.$$

Now let $s \ge 1$. Since $\|g\|_2 \le \sqrt{8L\bar{\Delta}_0}$, we obtain

$$\mathbb{P}_u\left(|\langle g, u\rangle| > \frac{\sqrt{8L\bar{\Delta}_0}}{\sqrt{d}}\, s\right) \le s^{-p}.$$

Using $|D_\mu f(x,u)| \le |\langle g,u\rangle| + \frac{L\mu}{2}$ and the fact that $\frac{L\mu}{2} \le L\mu s$ for $s \ge 1$, we conclude

$$\mathbb{P}_u\left(|D_\mu f(x,u)| > \left(\frac{\sqrt{8L\bar{\Delta}_0}}{\sqrt{d}} + L\mu\right)s\right) \le \mathbb{P}_u\left(|\langle g,u\rangle| > \frac{\sqrt{8L\bar{\Delta}_0}}{\sqrt{d}}\,s\right) \le s^{-p}.$$

□

Proof of Lemma B.4. By Lemma B.3, for every $s \ge 1$,

$$\mathbb{P}\bigg(|D_\mu f(x,u)| > 3\bigg(\frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu\bigg)s\bigg) \le s^{-p},$$

because

$$\frac{\sqrt{8L\bar{\Delta}_0}}{\sqrt{d}} + L\mu \le 3\left(\frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu\right).$$

By Lemma B.2, for every $s \ge 1$,

$$\mathbb{P}(|Y - D_\mu f(x,u)| > \bigg(\frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu\bigg)s) \le s^{-p}.$$

Therefore, for every $s \ge 1$,

$$\begin{aligned}
&\mathbb{P}\bigg(|Y| > 4\bigg(\frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu\bigg)s\bigg) \\
&\quad\le \mathbb{P}\bigg(|D_\mu f(x,u)| > 3\bigg(\frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu\bigg)s\bigg) \\
&\qquad + \mathbb{P}\bigg(|Y - D_\mu f(x,u)| > \bigg(\frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu\bigg)s\bigg) \\
&\quad\le 2s^{-p}.
\end{aligned}$$

Let $S := \frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu$. If $t \ge 4S$, then taking $s := \frac{t}{4S} \ge 1$ in the preceding bound gives

$$\mathbb{P}(|Y| > t) \le 2s^{-p} = 2\left(\frac{4S}{t}\right)^p = 2^{2p+1}S^p t^{-p}.$$

□

Proof of Lemma B.5. Let $S := \frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu$. Since

$$\mathbb{E}[Y \mid u] = \frac{f(x+\mu u) - f(x-\mu u)}{2\mu} = D_\mu f(x,u),$$

we have

$$d\,\mathbb{E}[Yu] = d\,\mathbb{E}[D_\mu f(x,u)u] = \nabla f_\mu(x).$$

Therefore,

$$d\,\mathbb{E}[\psi_\tau(Y)u] - \nabla f_\mu(x) = d\,\mathbb{E}[(\psi_\tau(Y) - Y)u].$$

It holds that

$$\Big\|d\,\mathbb{E}[\psi_\tau(Y)u] - \nabla f_\mu(x)\Big\|_2 \le d\,\mathbb{E}\big[|\psi_\tau(Y) - Y|\big].$$

By Lemma B.4, the random variable Y satisfies

$$\forall t \ge 4S, \quad \mathbb{P}(|Y| > t) \le 2^{2p+1}S^p t^{-p}.$$

Since the lemma assumes $\tau \ge 4S$, Lemma B.1 applies and yields

$$\mathbb{E}[|Y - \psi_\tau(Y)|] \le \frac{2^{2p+1}}{p-1}S^p\tau^{1-p}.$$

Substituting this into the previous bound gives

$$\left\| d\,\mathbb{E}[\psi_\tau(Y)u] - \nabla f_\mu(x) \right\|_2 \le \frac{2^{2p+1}}{p-1}\, d\, S^p \tau^{1-p}.$$

Recalling the definition of S, this is exactly

$$\left\| d\,\mathbb{E}[\psi_\tau(Y)u] - \nabla f_\mu(x) \right\|_2 \le \frac{2^{2p+1}}{p-1}\, d \left(\frac{\sqrt{L\bar{\Delta}_0} + \sigma}{\sqrt{d}} + L\mu \right)^p \tau^{1-p}.$$

For the equivalent form, let $g(x) := \frac{d}{M} \sum_{\ell=1}^{M} \psi_\tau(Y_\ell) u_\ell$ where (Y_ℓ, u_ℓ) are i.i.d. copies of (Y, u). Then

$$\mathbb{E}[g(x)] = d\,\mathbb{E}[\psi_\tau(Y)u],$$

and therefore

$$\mathbb{E}[g(x)] - \nabla f_\mu(x) = d\,\mathbb{E}[\psi_\tau(Y)u] - \nabla f_\mu(x).$$

Thus the same bound holds for $B := \mathbb{E}[g(x)] - \nabla f_\mu(x)$. □

Proof of Lemma B.6. We have $\psi_\tau(z) = \operatorname{sgn}(z) \min\{|z|, \tau\}$, so $\psi_\tau(Z)^2 = \min\{Z^2, \tau^2\}$. Using the layer-cake formula,

$$\mathbb{E}[\psi_\tau(Z)^2] = 2 \int_0^\infty t\, \mathbb{P}(|\psi_\tau(Z)| > t)\, dt = 2 \int_0^\tau t\, \mathbb{P}(|Z| > t)\, dt.$$

We split the integral at $4S$:

$$\mathbb{E}[\psi_\tau(Z)^2] = 2 \int_0^{4S} t\, \mathbb{P}(|Z| > t)\, dt + 2 \int_{4S}^\tau t\, \mathbb{P}(|Z| > t)\, dt.$$

Since $\mathbb{P}(|Z| > t) \le 1$, the first term is bounded by $2 \int_0^{4S} t\, dt = 16S^2$. For the second term, use the assumed tail bound:

$$2 \int_{4S}^\tau t\, \mathbb{P}(|Z| > t)\, dt \le 2^{2p+2} S^p \int_{4S}^\tau t^{1-p}\, dt.$$

Now, for $4S \le t \le \tau$ and $1 < p \le 2$,

$$t^{1-p} = t^{-1} t^{2-p} \le t^{-1} \tau^{2-p},$$

hence

$$\int_{4S}^\tau t^{1-p}\, dt \le \tau^{2-p} \int_{4S}^\tau \frac{dt}{t} = \tau^{2-p} \log\Big(\frac{\tau}{4S}\Big).$$

Therefore,

$$\mathbb{E}[\psi_\tau(Z)^2] \le 16S^2 + 2^{2p+2} S^p \tau^{2-p} \log\Big(\frac{\tau}{4S}\Big).$$

Since $\tau \ge 4S$ and $1 < p \le 2$, $S^2 \le S^p \tau^{2-p}$, so

$$16S^2 \le 16 S^p \tau^{2-p}.$$

Also,

$$\log\Big(\frac{\tau}{4S}\Big) \le \log\Big(\frac{\tau}{S}\Big).$$

Thus

$$\mathbb{E}[\psi_\tau(Z)^2] \le \Big(16 + 2^{2p+2} \log(\tau/S)\Big) S^p \tau^{2-p}.$$

Finally, since $1 < p \le 2$, we have $2^{2p+2} \le 64$. Hence

$$\mathbb{E}[\psi_\tau(Z)^2] \le 64\, S^p \tau^{2-p} \big(1 + \log(\tau/S)\big).$$

□

Proof of Lemma B.7. Let $S := \frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu$. For each ℓ, define

$$X_\ell := \frac{d}{M}\Big(\widetilde{Y}_\ell u_\ell - \mathbb{E}[\widetilde{Y}_\ell u_\ell]\Big).$$

Then $A = \sum_{\ell=1}^M X_\ell$ and $X_1, \dots, X_M$ are independent mean-zero random vectors.

Since $\|u_\ell\|_2 = 1$ almost surely and $|\widetilde{Y}_\ell| \le \tau$, we have

$$\|X_\ell\|_2 \le \frac{d}{M}\Big(|\widetilde{Y}_\ell| + \|\mathbb{E}[\widetilde{Y}_\ell u_\ell]\|_2\Big) \le \frac{2d\tau}{M}.$$

Thus the almost-sure bound in Lemma B.16 holds with $R := \frac{2d\tau}{M}$. Next, using $\mathbb{E}\|Z - \mathbb{E}Z\|_2^2 \le \mathbb{E}\|Z\|_2^2$, we obtain

$$\sum_{\ell=1}^M \mathbb{E}\|X_\ell\|_2^2 \le \frac{d^2}{M}\mathbb{E}[\widetilde{Y}^2],$$

where $\widetilde{Y} = \psi_\tau(Y)$, and Y has the same distribution as in Lemma B.4. By Lemma B.4,

$$\forall t \ge 4S, \quad \mathbb{P}(|Y| > t) \le 2^{2p+1} S^p t^{-p}.$$

Since $\tau \ge 4S$, Lemma B.6 gives

$$\mathbb{E}[\widetilde{Y}^2] \le 64 S^p \tau^{2-p}\left(1 + \log(\tau/S)\right).$$

Therefore

$$V := \sum_{\ell=1}^M \mathbb{E}\|X_\ell\|_2^2 \le \frac{64d^2}{M} S^p \tau^{2-p}\left(1 + \log(\tau/S)\right).$$

Applying Lemma B.16 with failure probability ρ, we obtain, with probability at least $1-\rho$,

$$\|A\|_2 \le \sqrt{V}\left(1 + \sqrt{2\log(1/\rho)}\right) + \frac{4R}{3}\log(1/\rho).$$

Using the bounds on V and R, this gives

$$\|A\|_2 \le \frac{8d}{\sqrt{M}} S^{p/2}\tau^{1-\frac{p}{2}}\sqrt{1+\log(\tau/S)}\left(1 + \sqrt{2\log(1/\rho)}\right) + \frac{8d\tau}{3M}\log(1/\rho).$$

□

Proof of Corollary B.8. We have

$$g(x) - \nabla f_\mu(x) = \underbrace{g(x) - \mathbb{E}[g(x)]}_{A} + \underbrace{\mathbb{E}[g(x)] - \nabla f_\mu(x)}_{B}.$$

Apply Lemma B.7 with failure probability $\rho := \frac{\delta}{T}$. Since $\log(1/\rho) = \log(T/\delta) = \lambda$, we obtain, with probability at least $1 - \delta/T$,

$$\|A\|_2 \le \frac{8d}{\sqrt{M}}\left(\frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu\right)^{p/2}\tau^{1-\frac{p}{2}} \times \sqrt{1 + \log\left(\frac{\tau}{\frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu}\right)}\left(1 + \sqrt{2\lambda}\right) + \frac{8d\tau}{3M}\lambda.$$

On the other hand, Lemma B.5 gives deterministically

$$\|B\|_2 \le \frac{2^{2p+1}}{p-1}\, d\left(\frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu\right)^p \tau^{1-p}.$$

Combining the two bounds by the triangle inequality gives the result. □

Proof of Corollary B.9 (Appendix form of Theorem 2.1). Let $S := \frac{\sqrt{L\bar{\Delta}_0}+\sigma}{\sqrt{d}} + L\mu$. Since $T \geq 3$ and $\delta \in (0,1]$, we have $\lambda = \log(T/\delta) \geq \log 3 > 1$. Moreover, since $M \geq \lambda$,

$$\tau = 8S\left(\frac{M}{\lambda}\right)^{1/p} \geq 8S \geq 4S,$$

so Corollary B.8 applies. From Corollary B.8, with probability at least $1 - \delta/T$,

$$\begin{aligned}\|g(x) - \nabla f_\mu(x)\|_2 \leq \frac{8d}{\sqrt{M}} S^{p/2}\tau^{1-\frac{p}{2}}\sqrt{1+\log(\tau/S)}\left(1+\sqrt{2\lambda}\right) \\ + \frac{8d\tau}{3M}\lambda + \frac{2^{2p+1}}{p-1}dS^p\tau^{1-p}.\end{aligned}$$

We bound the three terms separately.

For the first term, substituting $\tau = 8S\left(\frac{M}{\lambda}\right)^{1/p}$ gives $\frac{8d}{\sqrt{M}}S^{p/2}\tau^{1-\frac{p}{2}} = 8^{2-\frac{p}{2}}dSM^{-\frac{p-1}{p}}\lambda^{-\frac{2-p}{2p}}$. Since $\lambda \geq 1$, $1+\sqrt{2\lambda} \leq (1+\sqrt{2})\sqrt{\lambda}$. Therefore,

$$\begin{aligned}&\frac{8d}{\sqrt{M}}S^{p/2}\tau^{1-\frac{p}{2}}\sqrt{1+\log(\tau/S)}\left(1+\sqrt{2\lambda}\right) \\ &\leq 8^{2-\frac{p}{2}}(1+\sqrt{2})dS\left(\frac{\lambda}{M}\right)^{\frac{p-1}{p}}\sqrt{1+\log\left(8\left(\frac{M}{\lambda}\right)^{1/p}\right)}.\end{aligned}$$

Since $1 < p \leq 2$, we have $8^{2-p/2} \leq 8^{3/2} < 23$. Also, because $M \geq \lambda$,

$$1+\log\left(8\left(\frac{M}{\lambda}\right)^{1/p}\right) \leq (1+\log 8)\left(1+\log\left(\frac{M}{\lambda}\right)\right).$$

Thus the first term is bounded by

$$98\,dS\left(\frac{\lambda}{M}\right)^{\frac{p-1}{p}}\sqrt{1+\log\left(\frac{M}{\lambda}\right)}.$$

For thelinear term,

$$\frac{8d\tau}{3M}\lambda = \frac{64}{3}dS\left(\frac{\lambda}{M}\right)^{\frac{p-1}{p}} \leq 22\,dS\left(\frac{\lambda}{M}\right)^{\frac{p-1}{p}}\sqrt{1+\log\left(\frac{M}{\lambda}\right)}.$$

For the clipping-bias term,

$$\frac{2^{2p+1}}{p-1}dS^p\tau^{1-p} = \frac{2^{4-p}}{p-1}dS\left(\frac{\lambda}{M}\right)^{\frac{p-1}{p}}.$$

Since $\sqrt{1+\log(M/\lambda)} \geq 1$, this is at most

$$\frac{2^{4-p}}{p-1}dS\left(\frac{\lambda}{M}\right)^{\frac{p-1}{p}}\sqrt{1+\log\left(\frac{M}{\lambda}\right)}.$$

Combining the three bounds gives

$$\|g(x) - \nabla f_\mu(x)\|_2 \leq \left(120 + \frac{2^{4-p}}{p-1}\right)dS\left(\frac{\lambda}{M}\right)^{\frac{p-1}{p}}\sqrt{1+\log\left(\frac{M}{\lambda}\right)}.$$

□

Proof of Proposition B.10. Since f is L-smooth, its spherical smoothing f_μ is also L-smooth. Hence

$$f_\mu(x^+) \le f_\mu(x) + \langle \nabla f_\mu(x), x^+ - x \rangle + \frac{L}{2} \left\| x^+ - x \right\|_2^2 .$$

Using $x^+ - x = -\alpha g(x)$, we obtain

$$f_\mu(x^+) \le f_\mu(x) - \alpha \langle \nabla f_\mu(x), g(x) \rangle + \frac{L\alpha^2}{2} \|g(x)\|_2^2 .$$

Now write

$$g(x) = \nabla f_\mu(x) + e(x),$$

so that

$$\langle \nabla f_\mu(x), g(x) \rangle = \|\nabla f_\mu(x)\|_2^2 + \langle \nabla f_\mu(x), e(x) \rangle .$$

Moreover,

$$\|g(x)\|_2^2 = \|\nabla f_\mu(x) + e(x)\|_2^2 \le 2 \|\nabla f_\mu(x)\|_2^2 + 2 \|e(x)\|_2^2 ,$$

and, by Young's inequality,

$$|\langle \nabla f_\mu(x), e(x) \rangle| \le \|\nabla f_\mu(x)\|_2 \, \|e(x)\|_2 \le \frac{1}{4} \|\nabla f_\mu(x)\|_2^2 + \|e(x)\|_2^2 .$$

Substituting these bounds into the previous inequality yields

$$f_\mu(x^+) \le f_\mu(x) - \alpha\Big(\frac{3}{4} - L\alpha\Big) \|\nabla f_\mu(x)\|_2^2 + \alpha(1 + L\alpha) \|e(x)\|_2^2 .$$

If $\alpha \le \frac{1}{4L}$, then

$$\frac{3}{4} - L\alpha \ge \frac{1}{2} \qquad \text{and} \qquad 1 + L\alpha \le \frac{5}{4}.$$

Therefore,

$$f_\mu(x^+) \le f_\mu(x) - \frac{\alpha}{2} \|\nabla f_\mu(x)\|_2^2 + \frac{5\alpha}{4} \|e(x)\|_2^2 .$$

□

Proof of Theorem B.11. Set $\lambda := \log \frac{T}{\delta}$. Since $T \ge 3$ and $\delta \in (0, 1]$, we have $\lambda \ge \log 3 > 1$.

We first define the filtration generated by the algorithm. Let $\mathcal{F}_{-1} := \sigma(x_0)$, and for $t \ge 0$, define

$$\mathcal{F}_t := \sigma(x_0, \{u_{s,\ell}, \xi_{s,\ell} : 0 \le s \le t,\ 1 \le \ell \le M\}) .$$

Then x_t is $\mathcal{F}_{t-1}$-measurable, while g_t is $\mathcal{F}_t$-measurable. Moreover, conditionally on $\mathcal{F}_{t-1}$, the fresh samples

$$\{u_{t,\ell}, \xi_{t,\ell} : 1 \le \ell \le M\}$$

are independent of the past and have the same distribution.

For each $t = 0, \dots, T-1$, define the localization event

$$\mathcal{L}_t := \Big\{ f(x_s) - f_* \le 4\bar{\Delta}_0 \text{ for all } s = 0, \dots, t \Big\}.$$

Since $x_0, \dots, x_t$ are $\mathcal{F}_{t-1}$-measurable, we have $\mathcal{L}_t \in \mathcal{F}_{t-1}$. Also define the estimator-good event

$$\mathcal{E}_t := \Big\{ \|g_t - \nabla f_\mu(x_t)\|_2 \le \eta_0(M, \mu, \delta, T) \Big\}.$$

On the event $\mathcal{L}_t$, the point x_t satisfies $f(x_t) - f_* \le 4\bar{\Delta}_0$. Therefore, conditionally on $\mathcal{F}_{t-1}$, we may apply Corollary B.9 to the fixed point x_t, using the fresh samples at iteration t. Hence, almost surely,

$$\mathbf{1}_{\mathcal{L}_t} \mathbb{P}(\mathcal{E}_t^c \mid \mathcal{F}_{t-1}) \le \mathbf{1}_{\mathcal{L}_t} \frac{\delta}{T}.$$

It follows that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_t^c \cap \mathcal{L}_t) &= \mathbb{E}\big[\mathbf{1}_{\mathcal{E}_t^c}\mathbf{1}_{\mathcal{L}_t}\big] \\ &= \mathbb{E}\big[\mathbb{E}\big[\mathbf{1}_{\mathcal{E}_t^c}\mathbf{1}_{\mathcal{L}_t} \mid \mathcal{F}_{t-1}\big]\big] \\ &= \mathbb{E}[\mathbf{1}_{\mathcal{L}_t}\mathbb{P}(\mathcal{E}_t^c \mid \mathcal{F}_{t-1})] \\ &\le \mathbb{E}\left[\mathbf{1}_{\mathcal{L}_t}\frac{\delta}{T}\right] \\ &\le \frac{\delta}{T}.\end{aligned}$$

By the union bound,

$$\mathbb{P}\left(\bigcup_{t=0}^{T-1}(\mathcal{E}_t^c \cap \mathcal{L}_t)\right) \le \delta.$$

Define $\mathcal{G} := \bigcap_{t=0}^{T-1}(\mathcal{L}_t^c \cup \mathcal{E}_t)$. Since $\mathcal{G}^c = \bigcup_{t=0}^{T-1}(\mathcal{E}_t^c \cap \mathcal{L}_t)$, it holds that

$$\mathbb{P}(\mathcal{G}) \ge 1 - \delta.$$

We work on the good event $\mathcal{G}$ for the rest of the proof. We first prove by induction that the iterates remain localized. Clearly

$$f(x_0) - f_* = \Delta_0 \le \bar{\Delta}_0 \le 4\bar{\Delta}_0,$$

so $\mathcal{L}_0$ holds. Suppose that $\mathcal{L}_t$ holds for some $t \le T-2$. Then, on the good event, $\mathcal{E}_s$ holds for every $s = 0, \dots, t$, since $\mathcal{L}_t$ implies $\mathcal{L}_s$ for all $s \le t$. Proposition B.10 gives, for every $s = 0, \dots, t$,

$$f_\mu(x_{s+1}) \le f_\mu(x_s) - \frac{\alpha}{2}\|\nabla f_\mu(x_s)\|_2^2 + \frac{5\alpha}{4}\eta_0(M,\mu,\delta,T)^2.$$

Dropping the negative term and summing from $s = 0$ to t, we obtain

$$f_\mu(x_{t+1}) - f_* \le f_\mu(x_0) - f_* + \frac{5\alpha}{4}(t+1)\eta_0(M,\mu,\delta,T)^2.$$

We have $t+1 \le T$, and using Lemma B.17, $f_\mu(x_0) - f_* \le f(x_0) - f_* + \frac{L\mu^2}{2} = \bar{\Delta}_0$, then

$$f_\mu(x_{t+1}) - f_* \le \bar{\Delta}_0 + \frac{5\alpha T}{4}\eta_0(M,\mu,\delta,T)^2.$$

Now $\alpha = \frac{1}{4L}$, $\eta_0 \le \eta \le \epsilon/4$, and

$$T = \max\left\{3, \left\lceil \frac{32L\bar{\Delta}_0}{\epsilon^2}\right\rceil\right\} \le \frac{96L\bar{\Delta}_0}{\epsilon^2}.$$

Indeed, if $a := 32L\bar{\Delta}_0/\epsilon^2$, then the assumption $\epsilon^2 \le 32L\bar{\Delta}_0$ gives $a \ge 1$, and hence

$$\max\{3, \lceil a\rceil\} \le 3a = \frac{96L\bar{\Delta}_0}{\epsilon^2}.$$

Therefore, $\frac{5\alpha T}{4}\eta_0^2 \le \frac{5}{16L}\cdot\frac{96L\bar{\Delta}_0}{\epsilon^2}\cdot\frac{\epsilon^2}{16} = \frac{15}{8}\bar{\Delta}_0$. Consequently,

$$f_\mu(x_{t+1}) - f_* \le \frac{23}{8}\bar{\Delta}_0.$$

Using Lemma B.17,

$$f(x_{t+1}) - f_* \le f_\mu(x_{t+1}) - f_* + \frac{L\mu^2}{2} \le \frac{23}{8}\bar{\Delta}_0 + \bar{\Delta}_0 = \frac{31}{8}\bar{\Delta}_0 < 4\bar{\Delta}_0.$$

Thus $\mathcal{L}_{t+1}$ holds. By induction, $\mathcal{L}_t$ holds for all $t = 0, \dots, T-1$. Hence, on the good event, $\mathcal{E}_t$ holds for all $t = 0, \dots, T-1$.

We now prove the stationarity bound. Summing the one-step descent inequality over $t = 0, \dots, T-1$, and using $\|e_t\|_2 \le \eta_0(M,\mu,\delta,T)$, gives

$$\frac{\alpha}{2}\sum_{t=0}^{T-1}\|\nabla f_\mu(x_t)\|_2^2 \le f_\mu(x_0) - f_* + \frac{5\alpha T}{4}\eta_0(M,\mu,\delta,T)^2.$$

Since $f_\mu(x_0) - f_* \le \bar{\Delta}_0$, it holds that

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f_\mu(x_t)\|_2^2 \le \frac{2\bar{\Delta}_0}{\alpha T} + \frac{5}{2}\eta_0(M,\mu,\delta,T)^2.$$

Using $\alpha = \frac{1}{4L}$ and $T \ge \frac{32L\bar{\Delta}_0}{\epsilon^2}$, we get

$$\frac{2\bar{\Delta}_0}{\alpha T} = \frac{8L\bar{\Delta}_0}{T} \le \frac{\epsilon^2}{4}.$$

Therefore,

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f_\mu(x_t)\|_2^2 \le \frac{\epsilon^2}{4} + \frac{5}{2}\eta_0(M,\mu,\delta,T)^2.$$

Using Lemma B.17, $\|\nabla f(x_t)\|_2 \le \|\nabla f_\mu(x_t)\|_2 + L\mu$. Hence

$$\|\nabla f(x_t)\|_2^2 \le 2\|\nabla f_\mu(x_t)\|_2^2 + 2L^2\mu^2.$$

Averaging over $t = 0, \dots, T-1$, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2 &\le 2\left(\frac{\epsilon^2}{4} + \frac{5}{2}\eta_0^2\right) + 2L^2\mu^2 \\ &\le \frac{\epsilon^2}{2} + 5\eta_0^2 + 2L^2\mu^2.\end{aligned}$$

Since $\eta = \eta_0 + L\mu$, we have

$$5\eta_0^2 + 2L^2\mu^2 \le 5(\eta_0 + L\mu)^2 = 5\eta^2.$$

Using the assumption $\eta \le \epsilon/4$, we conclude that

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2 \le \frac{\epsilon^2}{2} + 5\frac{\epsilon^2}{16} = \frac{13}{16}\epsilon^2 \le \epsilon^2.$$

□

C High-Probability Analysis of Momentum RSC-ZO

Momentum averages clipped estimates from previous iterations. This reduces the stochastic fluctuation, but the older estimates also introduce a tracking error. We bound these two terms separately and then combine them in a descent argument.

Throughout this section, Assumptions 1.1, 1.2, and 1.3 hold. As in the analysis of the base method, write

$$\Delta_0 = f(x_0) - f_*, \qquad \bar{\Delta}_0 = \Delta_0 + \frac{L\mu^2}{2}, \qquad S_\mu = \frac{\sqrt{L\bar{\Delta}_0} + \sigma}{\sqrt{d}} + L\mu, \qquad C_p = 120 + \frac{2^{4-p}}{p-1}.$$

Let $\beta \in [1/2, 1)$, $T \ge 4$, and $\delta \in (0, 1]$. Set

$$\lambda = \log\frac{2T}{\delta}, \qquad \lambda_0 = \log\frac{2}{\delta}.$$

Choose integer batch sizes $M \ge (1-\beta)\lambda$ and $M_0 \ge \lambda_0$, and use the thresholds

$$\tau = 8S_\mu\left(\frac{M}{(1-\beta)\lambda}\right)^{1/p}, \qquad \tau_0 = 8S_\mu\left(\frac{M_0}{\lambda_0}\right)^{1/p}.$$

Define the error bounds

$$\eta = C_p d S_\mu \left(\frac{(1-\beta)\lambda}{M}\right)^{\frac{p-1}{p}} \sqrt{1 + \log\frac{M}{(1-\beta)\lambda}},$$

$$\eta_0 = C_p d S_\mu \left(\frac{\lambda_0}{M_0}\right)^{\frac{p-1}{p}} \sqrt{1 + \log\frac{M_0}{\lambda_0}}.$$

Here η_0 controls the warm start, while η controls both the running fluctuation and the clipping bias.

The algorithm starts with

$$g_0 = G_\mu(x_0; M_0, \tau_0), \qquad m_0 = g_0, \qquad x_1 = x_0 - \alpha g_0,$$

and, for $t \geq 1$, uses

$$g_t = G_\mu(x_t; M, \tau), \qquad m_t = \beta m_{t-1} + (1-\beta) g_t, \qquad x_{t+1} = x_t - \alpha m_t.$$

Each estimator uses fresh independent pairs $(u_{t,\ell}, \xi_{t,\ell})$, with $u_{t,\ell} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ independent of $\xi_{t,\ell} \sim \mathcal{P}$. Within each pair, the two function evaluations share the same $\xi_{t,\ell}$.

Let $\mathcal{F}_t$ contain all samples drawn through iteration t, including the warm-start batch, and let $\mathcal{F}_{-1} = \sigma(x_0)$. Then x_t is $\mathcal{F}_{t-1}$-measurable. Define

$$\mathcal{L}_t = \{f(x_s) - f_* \leq 4\bar{\Delta}_0 \text{ for } s = 0, \dots, t\}, \qquad t \geq 1.$$

This event is known before drawing batch t.

For the proof, separate each estimator into its centered fluctuation and its bias:

$$a_t = g_t - \mathbb{E}[g_t \mid \mathcal{F}_{t-1}], \qquad b_t = \mathbb{E}[g_t \mid \mathcal{F}_{t-1}] - \nabla f_\mu(x_t).$$

The accumulated fluctuation and the remaining tracking error are

$$z_t = (1-\beta) \sum_{s=1}^{t} \beta^{t-s} a_s, \qquad e_t = m_t - \nabla f_\mu(x_t) - z_t.$$

We use the Lyapunov function

$$\Phi_t = f_\mu(x_t) + \frac{2\alpha}{1-\beta} \|e_t\|_2^2.$$

C.1 Lemmas and convergence guarantees

Lemma C.1 (Clipped-estimator bounds)**.** *Let x be measurable with respect to a sigma-field $\mathcal{H}$ and satisfy $f(x) - f_* \leq 4\bar{\Delta}_0$. Conditional on $\mathcal{H}$, draw M fresh independent pairs (u_ℓ, ξ_ℓ) with the distribution specified above. Write*

$$Y_\ell = \frac{F(x + \mu u_\ell; \xi_\ell) - F(x - \mu u_\ell; \xi_\ell)}{2\mu}, \qquad g = \frac{d}{M} \sum_{\ell=1}^{M} \psi_\tau(Y_\ell) u_\ell.$$

For every $\tau \geq 8 S_\mu$,

$$\|\mathbb{E}[g \mid \mathcal{H}] - \nabla f_\mu(x)\|_2 \leq \frac{2^{2p+1}}{p-1} d S_\mu^p \tau^{1-p}.$$

The centered summands

$$W_\ell = \frac{d}{M} \left(\psi_\tau(Y_\ell) u_\ell - \mathbb{E}[\psi_\tau(Y_\ell) u_\ell \mid \mathcal{H}]\right)$$

satisfy

$$\|W_\ell\|_2 \leq \frac{2d\tau}{M}, \qquad \sum_{\ell=1}^{M} \mathbb{E}[\|W_\ell\|_2^2 \mid \mathcal{H}] \leq \frac{64 d^2}{M} S_\mu^p \tau^{2-p} \left(1 + \log \frac{\tau}{S_\mu}\right).$$

In particular, for the running threshold chosen above, $\|b_t\|_2 \leq \eta$ on $\mathcal{L}_t$.

Lemma C.2 (Vector Freedman)**.** *Let $(X_i)_{i=1}^n$ be a Hilbert-space martingale-difference sequence adapted to $(\mathcal{H}_i)_{i=0}^n$. Suppose that, almost surely,*

$$\|X_i\|_2 \leq R, \qquad \sum_{i=1}^{n} \mathbb{E}[\|X_i\|_2^2 \mid \mathcal{H}_{i-1}] \leq V.$$

For every $\ell \geq \log 2$, with probability at least $1 - e^{-\ell}$,

$$\left\| \sum_{i=1}^{n} X_i \right\|_2 \leq 2\sqrt{V\ell} + \frac{4}{3} R\ell.$$

Lemma C.3 (Concentration along the trajectory)**.** *With probability at least $1-\delta$, the following statements hold simultaneously:*

$$\|g_0 - \nabla f_\mu(x_0)\|_2 \le \eta_0,$$

and, for every $t = 1, \dots, T-1$,

$$\mathcal{L}_t \quad \Longrightarrow \quad \|z_t\|_2 \le \eta.$$

We denote this event by $\mathcal{E}$. The warm-start bound alone holds with probability at least $1-\delta/2$.

Lemma C.4 (Tracking error)**.** *For $t \ge 2$,*

$$e_t = \beta e_{t-1} + (1-\beta) b_t + \beta\big(\nabla f_\mu(x_{t-1}) - \nabla f_\mu(x_t)\big).$$

On $\mathcal{L}_t$,

$$\|e_t\|_2^2 \le \frac{1+\beta}{2}\|e_{t-1}\|_2^2 + 2(1-\beta)\eta^2 + \frac{3\beta^2 L^2 \alpha^2}{1-\beta}\|m_{t-1}\|_2^2.$$

Lemma C.5 (Descent)**.** *If $\alpha \le 1/(2L)$, then, for every $t \ge 1$,*

$$f_\mu(x_{t+1}) \le f_\mu(x_t) - \frac{\alpha}{2}\|\nabla f_\mu(x_t)\|_2^2 - \frac{\alpha}{4}\|m_t\|_2^2 + \alpha\|e_t\|_2^2 + \alpha\|z_t\|_2^2.$$

On $\mathcal{E} \cap \mathcal{L}_t$, for $t \le T-1$, the last term is at most $\alpha\eta^2$.

Theorem C.6 (Lyapunov inequality)**.** *Assume $\alpha \le (1-\beta)/(16L)$. For $t = 1, \dots, T-2$, on $\mathcal{E} \cap \mathcal{L}_{t+1}$,*

$$\Phi_{t+1} \le \Phi_t - \frac{\alpha}{2}\|\nabla f_\mu(x_t)\|_2^2 - \frac{\alpha}{8}\|m_t\|_2^2 + 5\alpha\eta^2.$$

Lemma C.7 (Initialization)**.** *Assume $\alpha \le (1-\beta)/(16L)$ and*

$$\max\{\eta_0, \eta\} \le \frac{1}{2}\sqrt{L\bar{\Delta}_0}.$$

Whenever $\|g_0 - \nabla f_\mu(x_0)\|_2 \le \eta_0$, one has

$$f(x_1) - f_* \le 4\bar{\Delta}_0, \qquad \Phi_1 - f_* \le 2\bar{\Delta}_0.$$

Theorem C.8 (High-probability convergence)**.** *Assume the hypotheses of Lemma C.7 and*

$$5(T-1)\alpha\eta^2 \le \bar{\Delta}_0.$$

Then, with probability at least $1-\delta$,

$$f(x_t) - f_* \le 4\bar{\Delta}_0 \qquad \text{for } t = 1, \dots, T-1,$$

and

$$\frac{1}{T-2}\sum_{t=1}^{T-2}\|\nabla f(x_t)\|_2^2 \le \frac{8\bar{\Delta}_0}{\alpha(T-2)} + 20\eta^2 + 2L^2\mu^2.$$

Corollary C.9 (Explicit parameters and query complexity)**.** *Let $\epsilon \in (0,1)$, $0 < \mu \le \epsilon/(4Ld)$, and $\epsilon^2 \le 32L\bar{\Delta}_0$. Choose*

$$\alpha = \frac{1-\beta}{16\sqrt{3}\,L}, \qquad T = \left\lceil \frac{512\sqrt{3}\,L\bar{\Delta}_0}{(1-\beta)\epsilon^2} \right\rceil + 2.$$

If $\eta_0, \eta \le \epsilon/16$, then, with probability at least $1-\delta$,

$$\frac{1}{T-2}\sum_{t=1}^{T-2}\|\nabla f(x_t)\|_2^2 \le \epsilon^2.$$

The following explicit batch sizes ensure these error bounds. Set

$$A = \max\left\{\mathrm{e},\ \frac{p}{p-1}\left(\frac{16C_p d S_\mu}{\epsilon}\right)^2\right\}, \qquad K = (A\log A)^{\frac{p}{2(p-1)}},$$

and take

$$M_0 = \lceil \lambda_0 K \rceil, \qquad M = \lceil (1-\beta)\lambda K \rceil.$$

The number of stochastic function evaluations is $Q_{\mathrm{MOM}} = 2M_0 + 2M(T-1)$, and it satisfies the explicit bound

$$Q_{\mathrm{MOM}} \le 2\lambda_0 K + 2 + 2\big((1-\beta)\lambda K + 1\big)\left(\frac{512\sqrt{3}\,L\bar{\Delta}_0}{(1-\beta)\epsilon^2} + 2\right).$$

In particular,

$$Q_{\mathrm{MOM}} = \widetilde{O}\left(K\left(1 + \frac{L\bar{\Delta}_0}{\epsilon^2}\right) + \frac{L\bar{\Delta}_0}{(1-\beta)\epsilon^2}\right).$$

An explicit choice giving $M = 1$ is

$$\beta = 1 - \frac{1}{2K\log\!\left(\frac{2048\sqrt{3}\,L\bar{\Delta}_0 K}{\delta\epsilon^2}\right)}.$$

For this choice, the total query complexity becomes

$$Q_{\mathrm{MOM}} = \widetilde{O}\left[\left(1 + \frac{L\bar{\Delta}_0}{\epsilon^2}\right)\max\left\{1, \left(\frac{dS_\mu}{\epsilon}\right)^{\frac{p}{p-1}}\right\}\right].$$

The notation $\widetilde{O}$ suppresses logarithmic factors and constants depending only on p.

The concentration bound depends on $(1-\beta)\lambda/M$. Thus momentum has the same averaging effect as increasing the running batch size by $1/(1-\beta)$. The smaller batch is balanced by the stepsize restriction $\alpha \le (1-\beta)/(16L)$. This preserves the total query order down to batch size one. Choosing β even closer to one increases the number of iterations without reducing the batch further.

Finally, the smoothing choice gives

$$\bar{\Delta}_0 = \Delta_0 + O\left(\frac{\epsilon^2}{Ld^2}\right), \qquad dS_\mu = \sqrt{d}\big(\sqrt{L\Delta_0} + \sigma\big) + O(\epsilon).$$

Hence, when $L\Delta_0 \ge \epsilon^2$, the batch-one choice gives

$$Q_{\mathrm{MOM}} = \widetilde{O}\left[\frac{L\Delta_0}{\epsilon^2}\left(\frac{\sqrt{d}(\sqrt{L\Delta_0} + \sigma)}{\epsilon}\right)^{\frac{p}{p-1}}\right].$$

C.2 Proofs

Proof of Lemma C.1. Condition on $\mathcal{H}$. The directional tail bound from the base-method analysis gives

$$\Pr(|Y_\ell| > r \mid \mathcal{H}) \le 2^{2p+1}\left(\frac{S_\mu}{r}\right)^p, \qquad r \ge S_\mu.$$

For $r < 4S_\mu$, this bound is trivial; for $r \ge 4S_\mu$, it is the localized weak-L_p estimate already proved for the base method. Since $d\,\mathbb{E}[Y_\ell u_\ell \mid \mathcal{H}] = \nabla f_\mu(x)$,

$$\begin{aligned}\|\mathbb{E}[g \mid \mathcal{H}] - \nabla f_\mu(x)\|_2 &\le d\int_\tau^\infty \Pr(|Y_\ell| > r \mid \mathcal{H})\,dr \\ &\le \frac{2^{2p+1}}{p-1} dS_\mu^p \tau^{1-p}.\end{aligned}$$

Also, $|\psi_\tau(Y_\ell)| \le \tau$, which gives the almost-sure bound on W_ℓ. For its variance, use

$$\begin{aligned}\mathbb{E}[\psi_\tau(Y_\ell)^2 \mid \mathcal{H}] &= \int_0^\tau 2r\,\Pr(|Y_\ell| > r \mid \mathcal{H})\,dr \\ &\le S_\mu^2 + 2^{2p+2} S_\mu^p \int_{S_\mu}^\tau r^{1-p}\,dr \\ &\le 64 S_\mu^p \tau^{2-p}\left(1 + \log\frac{\tau}{S_\mu}\right).\end{aligned}$$

The last inequality follows from $p \le 2$, $r^{1-p} \le \tau^{2-p}/r$, and $2^{2p+2} \le 64$. Centering cannot increase the variance, so multiplying by d^2/M proves the claim.

Substituting the running threshold in the bias bound gives

$$\|b_t\|_2 \le \frac{2^{4-p}}{p-1} dS_\mu \left(\frac{(1-\beta)\lambda}{M}\right)^{\frac{p-1}{p}} \le \eta.$$

□

Proof of Lemma C.2. The Hilbert-space Bernstein inequality states that

$$\Pr\left(\left\|\sum_{i=1}^n X_i\right\|_2 \ge r\right) \le 2\exp\left(-\frac{r^2}{2(V + Rr/3)}\right).$$

See [Tarres and Yao, 2014, Corollary A.2]. For $s = \ell + \log 2$, take $r = \sqrt{2Vs} + 2Rs/3$. The failure probability is at most $2e^{-s} = e^{-\ell}$. Since $s \le 2\ell$, the stated bound follows. □

Proof of Lemma C.3. Fix $t \in \{1, \dots, T-1\}$. To apply concentration before localization has been proved, consider the stopped sum

$$\widetilde{z}_t = (1-\beta)\sum_{s=1}^t \beta^{t-s} \mathbf{1}_{\mathcal{L}_s} a_s.$$

It equals z_t on $\mathcal{L}_t$. Write $a_s = \sum_{\ell=1}^M W_{s,\ell}$, with the centered summands from Lemma C.1. Reveal the samples one at a time within each batch. The increments

$$(1-\beta)\beta^{t-s}\mathbf{1}_{\mathcal{L}_s} W_{s,\ell}$$

are martingale differences: $\mathcal{L}_s$ is known before batch s, and each sample is conditionally independent of the previously revealed samples.

Their norms are bounded by

$$R = \frac{2d\tau(1-\beta)}{M}.$$

Using $(1-\beta)^2 \sum_{j\ge 0} \beta^{2j} \le 1-\beta$, their total conditional variance is at most

$$V = \frac{64d^2(1-\beta)}{M} S_\mu^p \tau^{2-p}\left(1 + \log\frac{\tau}{S_\mu}\right).$$

Lemma C.2, with $\ell = \lambda$, therefore bounds $\|\widetilde{z}_t\|_2$ by $2\sqrt{V\lambda} + 4R\lambda/3$, except on an event of probability $\delta/(2T)$.

For this calculation only, write $q = (1-\beta)\lambda/M$. The threshold is $\tau = 8S_\mu q^{-1/p}$, so

$$\begin{aligned} R\lambda &= 16 dS_\mu q^{\frac{p-1}{p}}, \\ 2\sqrt{V\lambda} &= 16\, 8^{1-p/2} dS_\mu q^{\frac{p-1}{p}} \sqrt{1 + \log 8 + \frac{1}{p}\log(1/q)} \\ &\le 96 dS_\mu q^{\frac{p-1}{p}} \sqrt{1 + \log(1/q)}. \end{aligned}$$

Here we used $8^{1-p/2} \le 3$ and $1 + \log 8 + p^{-1}\log(1/q) \le 4(1 + \log(1/q))$. Since $96 + 64/3 < 120$, the fluctuation is at most

$$120 dS_\mu q^{\frac{p-1}{p}} \sqrt{1 + \log(1/q)} \le \eta.$$

A union bound over $t = 1, \dots, T-1$ costs at most $\delta/2$.

The warm-start calculation is identical, without geometric weights, with $q = \lambda_0/M_0$ and failure probability $\delta/2$. Adding its clipping bias, whose coefficient is $2^{4-p}/(p-1)$, gives exactly the bound η_0. Combining the two events proves the lemma. □

Proof of Lemma C.4. Substitute $g_t = \nabla f_\mu(x_t) + a_t + b_t$ in the momentum update, and subtract $z_t = \beta z_{t-1} + (1-\beta)a_t$. This gives the recurrence for e_t.

On $\mathcal{L}_t$, convexity of the squared norm gives

$$\|\beta e_{t-1} + (1-\beta)b_t\|_2^2 \le \beta\|e_{t-1}\|_2^2 + (1-\beta)\eta^2.$$

Apply Young's inequality with parameter $(1-\beta)/(2\beta)$ and use $\beta \ge 1/2$. The result is

$$\begin{aligned}\|e_t\|_2^2 \le \frac{1+\beta}{2}\|e_{t-1}\|_2^2 + 2(1-\beta)\eta^2 \\ + \frac{3\beta}{1-\beta}\|\beta(\nabla f_\mu(x_{t-1}) - \nabla f_\mu(x_t))\|_2^2.\end{aligned}$$

Smoothness bounds the final squared norm by $\beta^2 L^2\alpha^2\|m_{t-1}\|_2^2$. Using $\beta^3 \le \beta^2$ proves the claim. □

Proof of Lemma C.5. Smoothness and $2\langle a, b\rangle = \|a\|_2^2 + \|b\|_2^2 - \|a-b\|_2^2$ give

$$\begin{aligned}f_\mu(x_{t+1}) \le f_\mu(x_t) - \frac{\alpha}{2}\|\nabla f_\mu(x_t)\|_2^2 - \frac{\alpha}{2}(1-L\alpha)\|m_t\|_2^2 \\ + \frac{\alpha}{2}\|m_t - \nabla f_\mu(x_t)\|_2^2.\end{aligned}$$

Now use $L\alpha \le 1/2$ and $\|m_t - \nabla f_\mu(x_t)\|_2^2 = \|e_t + z_t\|_2^2 \le 2\|e_t\|_2^2 + 2\|z_t\|_2^2$. On $\mathcal{E} \cap \mathcal{L}_t$, the concentration lemma gives $\|z_t\|_2 \le \eta$. □

Proof of Theorem C.6. Multiply the tracking-error inequality at time $t+1$ by $2\alpha/(1-\beta)$ and add it to the descent inequality. The terms in $\|e_t\|_2^2$ cancel, giving

$$\begin{aligned}\Phi_{t+1} \le \Phi_t - \frac{\alpha}{2}\|\nabla f_\mu(x_t)\|_2^2 \\ - \alpha\left(\frac{1}{4} - \frac{6\beta^2L^2\alpha^2}{(1-\beta)^2}\right)\|m_t\|_2^2 + 5\alpha\eta^2.\end{aligned}$$

Since $\alpha \le (1-\beta)/(16L)$, the term $6\beta^2L^2\alpha^2/(1-\beta)^2$ is at most $6/256 < 1/8$. This proves the stated inequality. □

Proof of Lemma C.7. Write $D = \bar{\Delta}_0$ for this proof. Applying the same smoothness calculation as in Lemma C.5 to the initial step gives

$$\begin{aligned}f_\mu(x_1) - f_* \le f_\mu(x_0) - f_* + \frac{\alpha}{2}\|g_0 - \nabla f_\mu(x_0)\|_2^2 \\ \le D + \frac{\alpha}{2}\eta_0^2 \le \left(1 + \frac{1}{256}\right)D,\end{aligned}$$

because $\alpha \le 1/(32L)$ and $\eta_0^2 \le LD/4$. Since $|f_\mu - f| \le L\mu^2/2 \le D$, this puts x_1 inside the localized region. We can therefore use $\|b_1\|_2 \le \eta$.

The gradient and smoothing bounds imply

$$\|\nabla f_\mu(x_0)\|_2 \le \sqrt{2L\Delta_0} + L\mu \le 2\sqrt{LD}.$$

Also,

$$e_1 = \beta\big(g_0 - \nabla f_\mu(x_1)\big) + (1-\beta)b_1.$$

Using $L\alpha \le 1/32$, $1-\beta \le 1/2$, and $\eta_0, \eta \le \sqrt{LD}/2$, we obtain

$$\begin{aligned}\|e_1\|_2 \le (1+L\alpha)\eta_0 + 2L\alpha\sqrt{LD} + (1-\beta)\eta \\ \le \left(\frac{33}{64} + \frac{1}{16} + \frac{1}{4}\right)\sqrt{LD} = \frac{53}{64}\sqrt{LD} \le \sqrt{LD}.\end{aligned}$$

Consequently,

$$\Phi_1 - f_* \le \left(1 + \frac{1}{256}\right)D + \frac{2\alpha}{1-\beta}LD \le \left(1 + \frac{1}{256} + \frac{1}{8}\right)D < 2D.$$

□

Proof of Theorem C.8. Work on $\mathcal{E}$, whose probability is at least $1-\delta$. Initialization gives $\mathcal{L}_1$ and $\Phi_1 - f_* \le 2\bar{\Delta}_0$.

Suppose that the trajectory leaves the localized region before time T, and let r be the first exit time. Then $r \ge 2$. Summing the Lyapunov inequality through time $r-2$ gives

$$\Phi_{r-1} \le f_* + 2\bar{\Delta}_0 + 5(r-2)\alpha\eta^2.$$

The descent inequality still applies at time $r-1$, so

$$f_\mu(x_r) \le f_\mu(x_{r-1}) + \alpha\|e_{r-1}\|_2^2 + \alpha\eta^2 \le \Phi_{r-1} + \alpha\eta^2.$$

It follows that

$$f(x_r) - f_* \le 2\bar{\Delta}_0 + 5(T-1)\alpha\eta^2 + \frac{L\mu^2}{2} \le 4\bar{\Delta}_0,$$

contradicting the definition of r. Thus all iterates through $T-1$ remain localized.

We may now sum the Lyapunov inequality over $t = 1, \dots, T-2$. Since $\Phi_{T-1} \ge f_*$,

$$\frac{\alpha}{2}\sum_{t=1}^{T-2} \|\nabla f_\mu(x_t)\|_2^2 \le 2\bar{\Delta}_0 + 5(T-2)\alpha\eta^2.$$

Finally, use $\|\nabla f(x_t)\|_2^2 \le 2\|\nabla f_\mu(x_t)\|_2^2 + 2L^2\mu^2$ and divide by $T-2$. □

Proof of Corollary C.9. First, $\eta_0, \eta \le \epsilon/16$ and $\epsilon^2 \le 32L\bar{\Delta}_0$ imply $\max\{\eta_0, \eta\} \le \sqrt{L\bar{\Delta}_0}/2$. Thus the initialization lemma applies. The choice of T also gives

$$\alpha(T-1) \le \frac{32\bar{\Delta}_0}{\epsilon^2} + 2\alpha \le \frac{34\bar{\Delta}_0}{\epsilon^2}.$$

Hence

$$5(T-1)\alpha\eta^2 \le \frac{170}{256}\bar{\Delta}_0 < \bar{\Delta}_0.$$

The convergence theorem therefore yields

$$\begin{aligned}\frac{1}{T-2}\sum_{t=1}^{T-2} \|\nabla f(x_t)\|_2^2 &\le \frac{\epsilon^2}{4} + \frac{20\epsilon^2}{256} + \frac{\epsilon^2}{8} \\ &= \frac{29}{64}\epsilon^2 < \epsilon^2.\end{aligned}$$

It remains to verify the explicit batch sizes. For this calculation, write $\gamma = (p-1)/p$ and $a = 16C_p dS_\mu/\epsilon$. Then

$$A = \max\{\mathrm{e}, a^2/\gamma\}, \qquad K = (A\log A)^{1/(2\gamma)}.$$

Since $A \ge \mathrm{e}$, $\gamma \le 1/2$, and $\log\log A \le \log A - 1$,

$$\begin{aligned}a^2(1+\log K) &\le \gamma A + \frac{A}{2}(\log A + \log\log A) \\ &\le A\log A = K^{2\gamma}.\end{aligned}$$

Moreover, $K \ge \exp(1/(2\gamma))$, so the function $q^\gamma\sqrt{1+\log(1/q)}$ is increasing for $0 < q \le 1/K$. The rounded batch sizes satisfy

$$\frac{\lambda_0}{M_0} \le \frac{1}{K}, \qquad \frac{(1-\beta)\lambda}{M} \le \frac{1}{K}.$$

Therefore both error bounds are at most

$$C_p dS_\mu K^{-\gamma}\sqrt{1+\log K} \le \frac{\epsilon}{16}.$$

The query bound follows directly from $M_0 \le \lambda_0 K + 1$, $M \le (1-\beta)\lambda K + 1$, and

$$T - 1 \le \frac{512\sqrt{3}\,L\bar{\Delta}_0}{(1-\beta)\epsilon^2} + 2.$$

Finally, consider the proposed batch-one choice. For this verification, write

$$C = \frac{512\sqrt{3}\, L\bar{\Delta}_0}{\epsilon^2}, \qquad \Lambda = \log \frac{4CK}{\delta}.$$

The prescribed momentum satisfies $1-\beta = 1/(2K\Lambda)$. Since $C \geq 16\sqrt{3}$, we have $\beta \in [1/2, 1)$ and $T \leq 2C/(1-\beta)$. It follows that

$$\lambda \leq \log \frac{4C}{\delta(1-\beta)} = \Lambda + \log(2\Lambda) \leq 2\Lambda.$$

Thus $(1-\beta)\lambda K \leq 1$, and the rounded running batch is exactly $M = 1$. The explicit formula for K gives

$$K = \widetilde{O}\left(\max\left\{1, \left(\frac{dS_\mu}{\epsilon}\right)^{\frac{p}{p-1}}\right\}\right).$$

Together with $1/(1-\beta) = 2K\Lambda$, this proves the final complexity bound. □

D Experiments: Clipping Before versus After Aggregation

We now provide a controlled experiment designed to isolate the mechanism behind scalar clipping in stochastic zeroth-order optimization. The goal is not to benchmark on a difficult objective, but to test a specific prediction of our analysis: under heavy-tailed noise, clipping each scalar directional derivative before aggregation can be substantially more effective than clipping only the final batched vector estimator.

The latter baseline is important because it is the most direct analogue of standard gradient clipping in SGD. In first-order stochastic optimization, one observes a stochastic gradient vector and clips its norm before taking a step. The corresponding zeroth-order analogue is therefore to first construct the batched gradient proxy

$$g_{\mathrm{raw}} = \frac{d}{M}\sum_{\ell=1}^{M} Y_\ell u_\ell$$

and then apply vector clipping,

$$g_{\mathrm{vec}} = \mathrm{clip}_{r_{\mathrm{vec}}}(g_{\mathrm{raw}}).$$

Our method clips earlier, at the level of each scalar directional derivative Y_ℓ, before averaging. We also include the unclipped estimator g_{raw}, which corresponds to the standard MeZO-style zeroth-order update without clipping. Thus, the experiment compares three natural choices: no clipping, clipping after constructing a gradient-like vector as in standard SGD gradient clipping, and clipping each zeroth-order directional measurement before aggregation.

Experimental setup. We consider the quadratic objective

$$f(x) = \frac{1}{2}\|x\|_2^2,$$

observed through the stochastic zeroth-order oracle

$$F(x;\zeta) = \frac{1}{2}\|x\|_2^2 + \langle \zeta, x\rangle.$$

The noise is sparse and heavy-tailed:

$$\zeta = sAe_J,$$

where s is uniform on $\{-1,+1\}$, J is uniform on $\{1,\dots,d\}$, and A follows a Pareto distribution with tail exponent p. Thus

$$\nabla F(x;\zeta) - \nabla f(x) = \zeta.$$

For $p \in (1,2)$, this noise has finite first moment but infinite variance. The random sign ensures that $\mathbb{E}[\zeta] = 0$, so the population objective is exactly $f(x) = \frac{1}{2}\|x\|_2^2$. Since the perturbation $\langle \zeta, x\rangle$ is linear in x, every sample objective remains 1-smooth.

For a direction u_ℓ, the shared-randomness two-point estimator satisfies

$$Y_\ell = \frac{F(x+\mu u_\ell;\zeta_\ell) - F(x-\mu u_\ell;\zeta_\ell)}{2\mu} = \langle x, u_\ell\rangle + \langle \zeta_\ell, u_\ell\rangle.$$

We compare three estimators:

$$\begin{cases} g_{\mathrm{raw}} = \dfrac{d}{M}\displaystyle\sum_{\ell=1}^{M} Y_\ell u_\ell, & \text{raw / MeZO-style estimator,} \\ g_{\mathrm{vec}} = \mathrm{clip}_{r_{\mathrm{vec}}}(g_{\mathrm{raw}}), & \text{vector clipping after aggregation,} \\ g_{\mathrm{sc}} = \dfrac{d}{M}\displaystyle\sum_{\ell=1}^{M} \psi_\tau(Y_\ell) u_\ell, & \text{scalar clipping before aggregation.} \end{cases}$$

Why this comparison matters. The distinction between g_{vec} and g_{sc} is central. Final-vector clipping controls only the norm of the already-aggregated vector. Indeed, when $\|g_{\mathrm{raw}}\|_2 > r_{\mathrm{vec}}$,

$$\mathrm{clip}_{r_{\mathrm{vec}}}(g_{\mathrm{raw}}) = r_{\mathrm{vec}} \frac{g_{\mathrm{raw}}}{\|g_{\mathrm{raw}}\|_2},$$

so the direction of g_{raw} is preserved. Therefore, if a few extreme directional derivatives Y_ℓ dominate the sum, final-vector clipping can make the step smaller but cannot repair the corrupted direction.

In contrast, scalar clipping acts before aggregation:

$$Y_\ell \mapsto \psi_\tau(Y_\ell).$$

Thus each directional contribution is bounded before it enters the average. This prevents a single heavy-tailed directional derivative from dominating the estimated direction.

We emphasize that clipping each individual vector $dY_\ell u_\ell$ before averaging would be essentially equivalent to scalar clipping, since

$$\|dY_\ell u_\ell\|_2 = d|Y_\ell|.$$

The meaningful baseline is therefore clipping the final batched estimator, which is the natural analogue of standard gradient clipping.

Fair tuning. For each method, we tune the stepsize using 6 validation seeds. For final-vector clipping, we additionally tune the vector clipping radius r_{vec}, while for scalar clipping we tune the scalar threshold τ. We then fix the selected hyperparameters and evaluate all methods on 20 independent seeds. The reported curves and tables are computed only from these evaluation seeds, not from the validation seeds used for tuning.

Representative optimization behavior. Figure 3 shows the optimization trajectory for $d = 100$, $p = 1.5$, $M = 256$, and target accuracy $\epsilon = 0.1$. The raw estimator is unstable under the infinite-variance noise. Final-vector clipping stabilizes the iterates, but the method plateaus above the target accuracy. In contrast, scalar clipping before aggregation continues to decrease the stationarity measure and reaches the target reliably.

The corresponding median final stationarity values are:

method	median final $\|\nabla f(x_T)\|_2$	success rate
raw	0.662	0%
vector after batch	0.142	0%
scalar before batch	0.065	100%

Thus final-vector clipping helps significantly relative to the raw estimator, but scalar clipping is the only method that reliably reaches the target in this setting.

Direction diagnostic. To verify the mechanism, we measure the cosine similarity between each estimator and the true gradient:

$$\frac{\langle g, \nabla f(x)\rangle}{\|g\|_2 \|\nabla f(x)\|_2}.$$

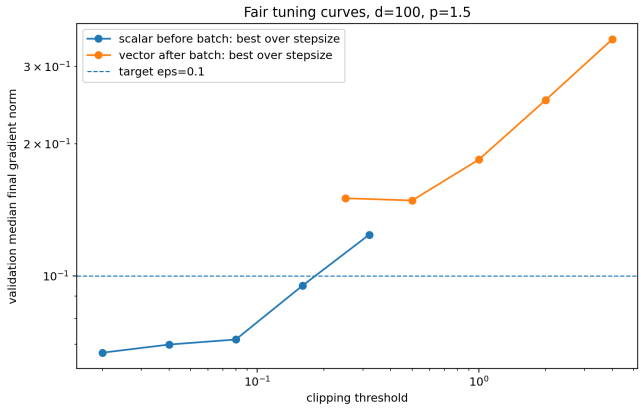

Figure 2: Tuning curves for the representative setting $d = 100$, $p = 1.5$. For each scalar threshold τ and vector clipping radius r_{vec}, we report the best validation performance over stepsizes. Even after tuning, scalar clipping before aggregation achieves a lower final stationarity measure than final-vector clipping after aggregation.

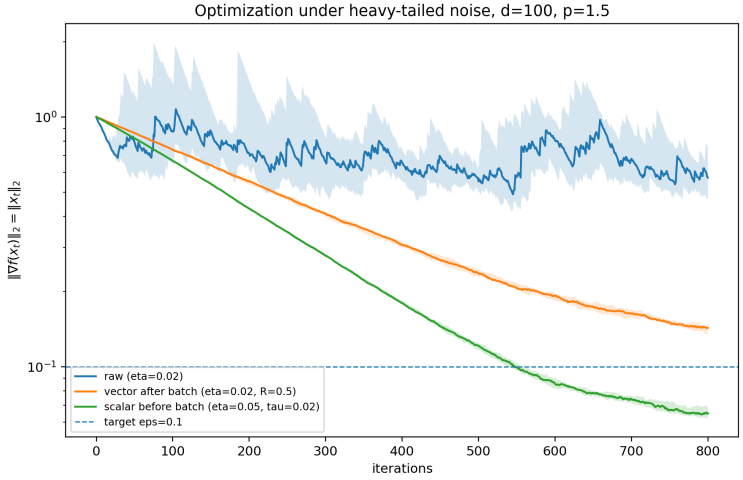

Figure 3: Representative optimization curves for $d = 100$, $p = 1.5$. We plot $\|\nabla f(x_t)\|_2 = \|x_t\|_2$ versus the number of iterations. Raw zeroth-order optimization is unstable under heavy-tailed noise. Final-vector clipping stabilizes the method but plateaus above the target. Scalar clipping before aggregation reaches the target and obtains the lowest final stationarity measure.

Figure 4 shows that the raw and final-vector-clipped estimators have nearly identical cosine distributions. This is expected because final-vector clipping only rescales g_{raw} and therefore preserves its direction. By contrast, scalar clipping substantially improves alignment with the true gradient, increasing the median cosine from approximately 0.317 to approximately 0.543.

This confirms that the advantage of scalar clipping is not merely norm control. The improvement comes from changing the aggregation itself: extreme directional derivatives are controlled before they can determine the direction of the batched estimator.

Dimension sweep. We next vary the ambient dimension d while fixing the tail exponent to $p = 1.5$. Figure 5 reports the median final stationarity and the success rate. The raw estimator degrades rapidly with dimension. Final-vector clipping improves stability but fails to reach the target for larger dimensions. Scalar clipping remains below the target across all tested dimensions.

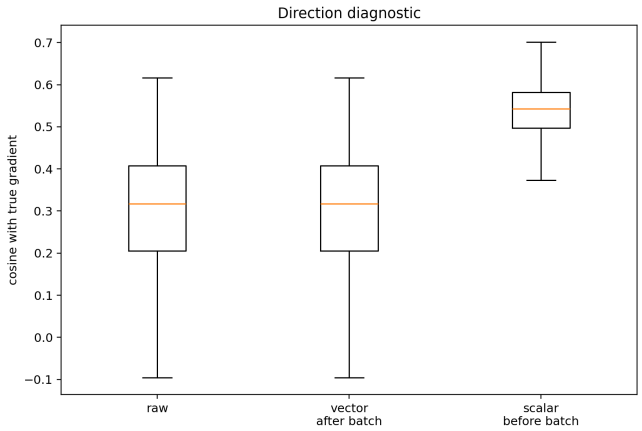

Figure 4: Direction diagnostic for $d = 100$, $p = 1.5$. We report the cosine similarity between each stochastic estimator and the true gradient. Final-vector clipping has essentially the same alignment as the raw estimator because it preserves the raw direction. Scalar clipping before aggregation yields much better alignment, explaining its improved optimization performance.

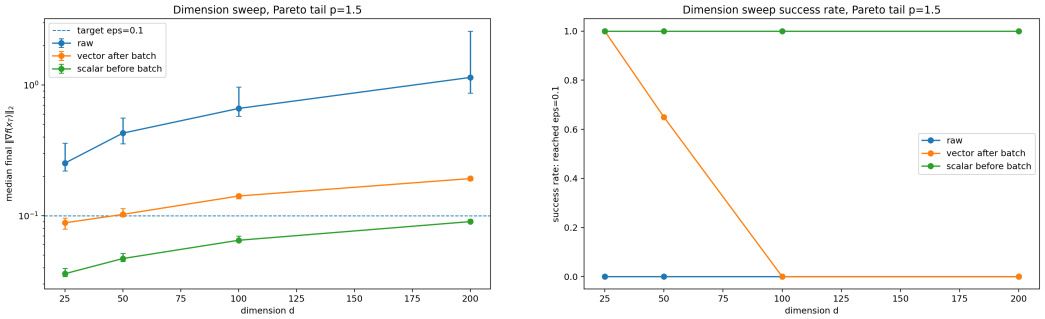

Figure 5: Dimension sweep for Pareto tail exponent $p = 1.5$. Left: median final stationarity $\|\nabla f(x_T)\|_2$. Right: success rate for reaching $\epsilon = 0.1$. Scalar clipping before aggregation remains robust as the dimension grows, while raw zeroth-order optimization and final-vector clipping degrade.

The median final stationarity values are:

d	raw	vector after batch	scalar before batch
25	0.253	0.088	0.036
50	0.429	0.102	0.047
100	0.662	0.142	0.065
200	1.144	0.193	0.090

The success rates show the same trend: scalar clipping succeeds on all tested dimensions, while vector clipping succeeds only at smaller dimensions and raw zeroth-order optimization fails throughout.

Tail-index sweep. Finally, we vary the Pareto tail exponent p while fixing $d = 100$. Smaller p corresponds to heavier tails. Figure 6 shows that the advantage of scalar clipping is strongest in the heavy-tailed regime and decreases as the noise becomes lighter-tailed.

The median final stationarity values are:

p	raw	vector after batch	scalar before batch
1.2	4.120	0.249	0.076
1.5	0.662	0.142	0.065
1.8	0.329	0.097	0.063
2.5	0.132	0.075	0.060

For $p = 1.2$, neither the raw estimator nor final-vector clipping reaches the target. Scalar clipping remains stable and achieves a substantially lower stationarity measure. As the Pareto tail exponent

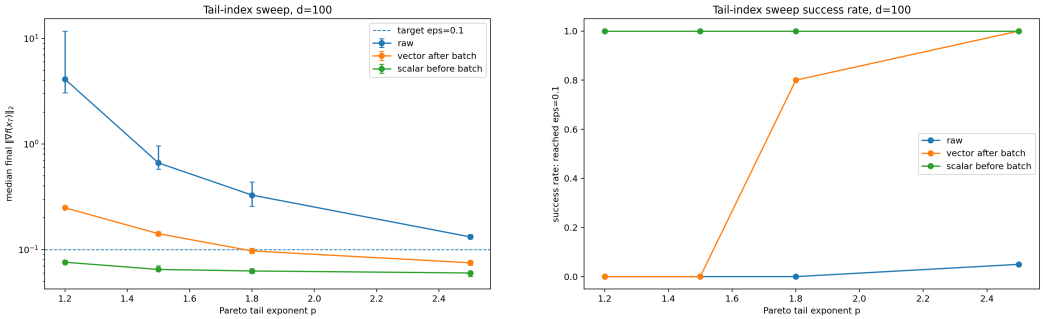

Figure 6: Tail-index sweep for $d = 100$. Smaller p corresponds to heavier-tailed noise. Scalar clipping gives the largest improvement in the infinite-variance regime $p < 2$. We also include $p = 2.5$ as a finite-variance sanity check, where the gap becomes smaller but scalar clipping remains competitive.

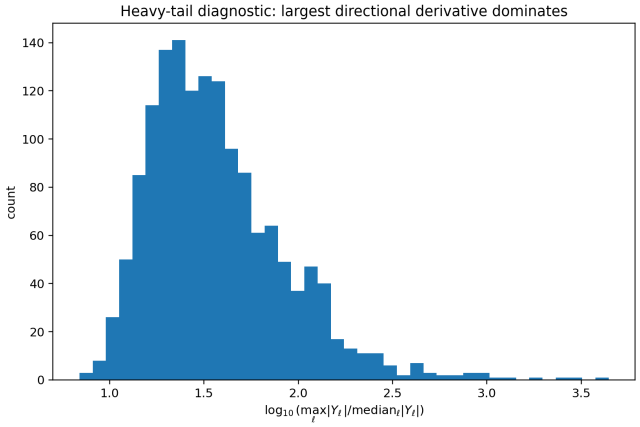

Figure 7: Outlier dominance diagnostic for the representative setting. We plot the distribution of $\log_{10}(\max_\ell |Y_\ell| / \text{median}_\ell |Y_\ell|)$. Large values indicate that a small number of directional derivatives can dominate the raw batched estimator. Scalar clipping prevents these outliers from controlling the aggregate direction.

increases and the tails become lighter, the gap narrows. The case with Pareto tail exponent 2.5 has finite variance and satisfies the weak-L_2 assumption, so it is covered by our theory with $p = 2$. We include it as a finite-variance comparison.

Outlier dominance. We also record the ratio between the largest directional derivative and the median directional derivative in each batch. Figure 7 shows that this ratio is often large, confirming that individual directional samples can dominate the raw average under heavy-tailed noise. This supports the need to clip before aggregation.

Summary. These experiments support the estimator-level mechanism underlying our theory. Final-vector clipping is useful because it controls the norm of the stochastic zeroth-order estimate, and indeed it improves over the raw estimator. However, it preserves the direction of the already-aggregated heavy-tailed estimate. Scalar clipping acts earlier: it clips each noisy directional derivative before aggregation. This yields better alignment with the true gradient, lower final stationarity, and higher success rates across dimensions and tail indices.

E Additional Details for Zeroth-Order Fine-Tuning Experiments

This appendix gives the implementation details for the RoBERTa-large fine-tuning experiments in Section 4. We follow the prompt-based MeZO setup of Malladi et al. [2023]. All methods are trained for 100K steps on $k = 512$ training examples per task.

For the multi-direction RSC-ZO variants, the estimator averages over $M \in \{1, 2, 4\}$ random directions, with data batch sizes 16, 8, and 4, respectively. Each direction requires two forward passes. Thus, every update processes $2M(16/M) = 32$ examples across all forward passes, counting an example each time it is evaluated. Increasing M therefore uses more directions but fewer examples per direction. MeZO uses one direction and a data batch size of 16, giving the same total number of examples processed per update.

Table 2 lists the hyperparameter grids for all methods.

Experiment	**Hyperparameters**	**Values**
MeZO / RSC-ZO	Training steps	100K
	Batch size ($M = 1, 2, 4$)	$16, 8, 4$
	Learning rate	$\{10^{-6}, 10^{-5}\}$
	μ	10^{-3}
	Weight decay	0
LOZO / Scalar-clipped LOZO	Training steps	100K
	Batch size	16
	Learning rate	2×10^{-7}
	Rank (r)	$\{4, 8\}$
	Interval (ν)	$\{50, 100\}$
	μ	10^{-3}
	Weight decay	0

Table 2: Hyperparameter grids for RoBERTa-large fine-tuning with $k = 512$. For RSC-ZO, each update uses $2M$ forward passes with data batch size $16/M$, keeping the total number of examples processed across these passes fixed at 32. MeZO uses $M = 1$.

Table 3 reports the sensitivity of RSC-ZO and scalar-clipped LOZO to the clipping threshold. Both methods are stable around their best-performing thresholds. For RSC-ZO, $\tau = 8$ and 16 perform similarly on SNLI, as do $\tau = 16$ and 32 on MNLI and TREC. On the other hand, scalar-clipped LOZO favors larger thresholds. In our main experiments on SNLI, MNLI, and TREC, we use $\tau = 8$, 32, and 32 for RSC-ZO, and $\tau = 64$, 32, and 64 for scalar-clipped LOZO, respectively.

		Clipping threshold τ						
Method	**Task**	2	4	8	16	32	64	128
RSC-ZO	SNLI	73.9	79.7	**83.4**	83.1	79.0	76.8	–
	MNLI	60.3	65.6	72.1	**76.1**	75.0	72.1	–
	TREC	61.6	78.2	89.2	**94.4**	94.0	90.4	–
Scalar-clipped LOZO	SNLI	65.7	72.4	78.4	82.6	82.9	**83.1**	81.0
	MNLI	54.9	59.4	64.3	72.3	**77.7**	76.0	72.7
	TREC	47.0	58.2	75.8	88.4	93.8	**94.2**	92.6

Table 3: Test accuracy of RSC-ZO and scalar-clipped LOZO on each task for each clipping threshold τ.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main theoretical claims, the weak-L_p noise assumption, the high-probability nature of the guarantee, and the scope of the zeroth-order setting.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The discussion section describes the main limitations: the analysis relies on sample-wise smoothness and weak-L_p control of the sample-gradient noise, and the empirical evaluation is limited to a small set of zeroth-order fine-tuning tasks and controlled synthetic experiments.

   Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions are stated in Section 1.1. The main theorems are stated in Sections 2 and 3, and the detailed proofs are provided in Appendices B and C.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper reports the model, datasets, number of training examples, number of training steps, batch size, learning rate, smoothing parameter, clipping-threshold grid, selected thresholds, and number of seeds. Additional implementation details are provided in the appendix. We provide also the code.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the full code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 and Appendix E describe the fine-tuning setup, datasets, model, number of examples, training steps, shared hyperparameters, clipping-threshold grid, and selected thresholds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The RoBERTa-large fine-tuning results report mean accuracy and standard deviation over five seeds. The paper explicitly states that the values in parentheses are standard deviations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper evaluates algorithmic efficiency purely in terms of exact function queries (oracle complexity) rather than wall-clock compute time, so hardware specifications do not affect the claims.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical and experimental research on optimization algorithms using standard public benchmarks. We are not aware of any aspect of the work that violates the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper is strictly focused on foundational mathematical optimization algorithms and has no direct path to negative societal impacts requiring mitigation discussion.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new pretrained model, generative model, scraped dataset, or other high-risk asset. The empirical evaluation uses existing public datasets and an existing pretrained model.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: The paper cites the original sources for the model, datasets, and codebase used in the experiments. Our implementation builds on the publicly available MeZO codebase of Malladi et al. [2023], which is released under the MIT License, and we preserve the original copyright and license notice in our code release. The README explicitly credits MeZO and states that our modifications are released under the MIT License unless otherwise specified. The datasets and pretrained model are accessed through their standard public releases; we do not redistribute dataset contents or pretrained model weights.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release code implementing RSC-ZO, based on the MeZO codebase. The released repository includes a README with installation instructions, dataset preparation, commands for reproducing the experiments, and license information. We do not release new datasets or pretrained model weights.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a methodological component in the algorithmic development of this work. Only for writing and editing

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.