ZERO-ORDER PARAMETER-FREE OPTIMIZATION FOR LMO-BASED METHODS: NOVEL APPROACH FOR EFFICIENT FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning large language models (LLMs) has become a central application of modern optimization, enabling pretrained models to adapt to diverse downstream tasks and domain-specific data. A major obstacle in large-scale fine-tuning is the memory overhead of backpropagation, which requires storing activations, gradients, and optimizer states. Zeroth-order (ZO) optimization offers a memory-efficient alternative, but its performance is highly sensitive to the stepsize and smoothing parameter, often requiring costly task-specific tuning. Parameter-free (PF) optimization addresses this issue by adapting algorithmic parameters without prior knowledge of problem-dependent constants. Moreover, large-scale fine-tuning can benefit from geometry-aware updates that account for the heterogeneous structure of parameter blocks, which can be modeled through methods that exploit linear minimization oracle (LMO). In this work, we study PF adaptation for LMO-based ZO optimization and introduce `AdaNAGED`, a method that unifies gradient-free training, adaptive tuning, and non-Euclidean update geometry. We establish convergence guarantees and validate the method on large-scale LLM fine-tuning task with $\texttt{OPT} - 1.3\text{B}$ model.

1 INTRODUCTION

Optimization is a fundamental component of modern machine learning, and its role is particularly prominent in the fine-tuning of large language models (LLMs) (Howard & Ruder, 2018; Lester et al., 2021). Fine-tuning offers an effective and flexible way to adapt pretrained models to diverse downstream tasks while leveraging the knowledge acquired during pretraining (Gururangan et al., 2020; Zhang et al., 2020; Wu et al., 2025). Formally, we consider the following optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x), \tag{1}$$

where x denotes the model parameters, and f is the loss function. Training is typically performed using gradient-based optimization methods, such as `SGD` (Robbins & Monro, 1951) or `Adam` (Kingma & Ba, 2014). These methods rely on gradients computed via backpropagation (Rojas, 1996), which provides the first-order information required for parameter updates. This combination has become the standard training pipeline in modern machine learning, offering strong empirical performance and reliable optimization behavior across a wide range of tasks.

Despite its strong empirical performance, backpropagation becomes a major memory bottleneck in large-scale LLM fine-tuning, as it requires storing intermediate activations, gradients, and optimizer states throughout training (Rajbhandari et al., 2020; Malladi et al., 2023). Since fine-tuning adapts an already pretrained model to a narrower target task or domain, this overhead motivates optimization methods that reduce or avoid the memory costs of first-order training. Existing approaches address this challenge through block-wise descent methods (Richtárik & Takáč, 2014; Luo et al., 2024), quantization (Dettmers et al., 2021; Novikov et al., 2023), and parameter-efficient fine-tuning methods (Sun et al., 2022; Wang et al., 2024). However, these methods still rely on backpropagation and therefore *cannot fully remove the memory burden* associated with gradient-based optimization. A fundamentally different optimization paradigm is provided by zeroth-order optimization (ZO) (Ghadimi & Lan, 2013). Instead of relying on gradients, ZO methods assume

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

access only to function-value evaluations of the objective. Formally, ZO replaces gradients with finite-difference estimates:

$$\nabla f(x) \approx \frac{f(x+\tau e) - f(x)}{\tau} e, \tag{2}$$

where e is a random sample vector and τ is a smoothing parameter. Zeroth-order optimization is a mature area of optimization, comprising a broad class of algorithms with established convergence guarantees (Zhang & Ying, 2024; Lin et al., 2025; Bar & Giryes, 2025). This paradigm directly addresses the memory bottleneck: instead of storing the intermediate quantities required for back-propagation, ZO methods rely on function-value queries along sampled perturbation directions. In particular, after the finite-difference estimator is constructed, the vector e need not be stored explicitly; following Malladi et al. (2023), it can be represented by the seed of a random generator, introducing essentially no additional memory overhead.

The ZO setting introduces several challenges, including the need for careful hyperparameter selection. In particular, the performance of ZO methods can be highly sensitive to the choice of the stepsize and the smoothing parameter. Although theoretically optimal choices exist, computing them typically requires prior access to problem-dependent quantities, such as Lipschitz constant, or the distance to the optimal solution (Robbins & Monro, 1951; Stich, 2019; Hendrikx et al., 2020), which are *unavailable* in practical applications. As a result, implementations often rely on time-consuming grid search, which must be repeated for each new task. A natural way to address this issue is through parameter-free optimization (PF). Algorithms in this paradigm do not require prior knowledge of problem-specific constants. Instead, they estimate such quantities adaptively from the local optimization landscape and the statistics accumulated along the trajectory. In first-order optimization, this line of work is well developed and includes a variety of schemes for estimating problem-dependent constants while preserving strong convergence guarantees (Carmon & Hinder, 2022; Deng et al., 2024; Kreisler et al., 2024).

Compared with the first-order setting, parameter-free zeroth-order optimization remains considerably less explored. Extending PF principles to ZO is challenging because hyperparameters must be adapted from finite-difference information rather than exact gradients. This is especially delicate for the smoothing parameter τ: large values increase the bias of the gradient approximation, while overly small values can lead to instability due to the division by τ (Ma & Huang, 2025). As a result, ZO-PF methods must adapt not only the stepsize but also the smoothing parameter τ without access to problem-dependent constants, making this direction challenging.

Beyond hyperparameter choice, ZO fine-tuning also raises the question of how to adapt updates to the geometry of large-scale models. Since different parameter groups may naturally correspond to different vector or matrix structures, Euclidean updates may not always capture the most suitable local geometry. This motivates optimization methods defined with respect to non-Euclidean norms. A prominent example of this principle is `Muon` (Jordan et al., 2024), which can be viewed as an `SGD`-type method that exploits spectral-norm geometry for matrix-shaped parameters. By approximately orthogonalizing matrix updates, `Muon` balances update directions according to the structure of matrix-valued weights. This idea can be formulated more generally using linear minimization oracles (LMOs) over norm balls (Bernstein & Newhouse, 2024; Pethick et al., 2025). Given a norm-induced domain $\mathcal{D}$, an LMO is defined as

$$\mathrm{lmo}_{\mathcal{D}}(x) = \arg\min_{v \in \mathcal{D}} \langle x, v \rangle. \tag{3}$$

Specific choices of the underlying norm recover different update rules: vector norms yield methods related to `Normalized SGD` with momentum (Cutkosky & Mehta, 2020) and `Sign-SGD` with momentum (Sun et al., 2023), whereas matrix norms capture methods such as `Muon` (Jordan et al., 2024) and `Gluon` (Riabinin et al., 2025). Such geometry-aware updates are especially attractive for large-scale fine-tuning, as they can improve the conditioning of parameter updates and provide more stable optimization across heterogeneous vector- and matrix-shaped blocks.

Parameter-free optimization and LMO-based updates address two essential aspects of ZO fine-tuning. The former reduces the reliance on costly task-specific hyperparameter tuning, while the latter enables optimization under structured non-Euclidean geometries through linear minimization oracles. Despite their individual relevance, it remains unclear *how parameter-free ideas can be carried out in the zeroth-order setting when the update direction is defined by an LMO*. In this work, we provide a comprehensive study of this question.

1.1 Related Work

Zeroth-order methods. Early foundations of gradient-free optimization include simultaneous perturbation methods, which estimate the descent directions from function-value evaluations along randomly perturbed points (Spall, 1998). A modern analysis of stochastic zeroth-order methods for nonconvex optimization was developed in (Ghadimi & Lan, 2013), where the gradient-based updates are replaced by estimators constructed from zeroth-order oracle queries. Since then, a wide range of estimator designs and sampling strategies has been studied (Lin et al., 2025). Classic choices include random directions on the unit sphere (Flaxman et al., 2004), coordinate-wise perturbations based on the canonical basis (Duchi et al., 2015), and Gaussian smoothing estimators (Ghadimi & Lan, 2013; Malladi et al., 2023). These estimators are typically designed to approximate the gradient of a smoothed objective using only function values.

More recent work has explored more structured perturbation schemes and algorithmic improvements. For example, `HiZOO` uses Hessian-informed zeroth-order updates to improve the convergence of LLM fine-tuning (Zhao et al., 2024). Another line of work connects randomized finite-difference estimators with policy-gradient methods, interpreting ZO updates from a reinforcement-learning perspective and relating them to `REINFORCE`-type estimators (Williams, 1992; Qiu et al., 2025; Seung et al., 2026). Orthogonal advances include momentum-based ZO methods (Chen et al., 2019), and variance-reduced estimators (Liu et al., 2018; Ji et al., 2019; Gautam et al., 2024).

ZO methods have been successfully applied to deep neural networks, demonstrating the practical relevance of this paradigm (Chen et al., 2017; Zhang et al., 2024). Moreover, they have emerged as a promising approach to alleviating the memory constraints of LLM fine-tuning (Malladi et al., 2023; Lin et al., 2025).

Parameter-free optimization. A large part of the theory of adaptive and parameter-free optimization has been developed for non-smooth convex objectives. In this setting, the optimal stepsize for gradient-based methods (Nemirovskij & Yudin, 1983) is typically chosen as

$$\gamma = \frac{\|x^0 - x^*\|}{M\sqrt{T}}. \tag{4}$$

where M is the Lipschitz constant of the objective ($|f(x) - f(y)| \leq M\|x - y\| \; \forall x, y \in \mathbb{R}^d$). In practice, however, neither M nor the initial distance to the solution $\|x^0 - x^*\|$ is available. The core idea of parameter-free optimization is to estimate such quantities adaptively and recover convergence guarantees without requiring them as input.

Early adaptive methods, including `AdaGrad` (Duchi et al., 2011), `AdaDelta` (Zeiler, 2012), and `Adam` (Kingma & Ba, 2014), were among the first widely used approaches to reduce the dependence on prior knowledge of problem constants. These methods exploit the statistics of already computed gradients to rescale the updates and construct an adaptive stepsize schedule throughout training. However, some of them introduce additional memory overhead due to the stored statistics, and they generally do not explicitly adapt to the initial distance to the optimum. Parameter-free ideas have also been extensively studied in online learning, although under a different problem formulation (Orabona, 2019). This line of work introduced coin-betting, reward-doubling, and related schemes (Orabona, 2013; McMahan & Orabona, 2014; Orabona & Pál, 2016; Cutkosky & Orabona, 2018). These methods eliminate several forms of manual tuning, but their guarantees are usually tied to online-learning assumptions, such as bounded stochastic oracles.

A subsequent milestone in parameter-free optimization was the development of methods that adapt to the initial distance $\|x^0 - x^*\|$ in (4). One of the first parameter-free methods with such adaptation was proposed in Carmon & Hinder (2022). However, the construction relies on an additional search procedure, which limits its practical efficiency. The `DoG` algorithm (Ivgi et al., 2023) was later introduced as an iterative scheme that adaptively estimates both the Lipschitz constant and the distance to the solution, although its distance estimate can be unbounded. Subsequent works (Defazio & Mishchenko, 2023; Mishchenko & Defazio, 2023) addressed this issue by introducing more stable distance estimators. However, some algorithms introduced in these papers still require prior knowledge of the Lipschitz constant and lack stochastic analysis. Recent work has further extended the parameter-free paradigm by incorporating momentum (Schaipp et al., 2023), weighting mechanisms (Kreisler et al., 2024), and sharper theoretical analyses (Attia & Koren, 2023; 2024).

Parameter-free methods have also been studied beyond the non-smooth setting. In the smooth case, DoWG (Khaled et al., 2023) provides an analysis of an AdaGrad−Norm stepsize. A complementary line of work builds on the adaptive stepsize idea from Malitsky & Mishchenko (2019). In this direction, the work (Mishkin et al., 2024) proposes a parameter-free method based on the local smoothness estimates computed along the optimization trajectory. Later, Medyakov et al. (2026) utilized this concept and applied it to the analysis of Sign−SGD in the PF setting.

LMO-based methods. Linear minimization oracles (LMOs) first appeared in the context of Frank−Wolfe-type methods for constrained optimization. This line of work studies optimization over structured feasible sets, including convex polyhedra, nuclear-norm balls, and flow polytopes (Bomze et al., 2021; Braun et al., 2022). The same technique was later recognized as useful for unconstrained problems formulated with respect to abstract norms.

The foundation of non-Euclidean analysis in this context can be traced to the Sign−SGD algorithm (Bernstein et al., 2018; Karimireddy et al., 2019). This method has been shown to outperform SGD when analyzed through the ℓ_1-norm geometry (Balles & Hennig, 2018; Balles et al., 2020). Later, Muon was introduced as a matrix-aware algorithm and analyzed under the spectral norm (Jordan et al., 2024). This sparked a line of related work on quantization (Gupta et al., 2025), faster randomized SVD decompositions (Huang et al., 2025), and layer-wise optimization (Riabinin et al., 2025), further establishing non-Euclidean analysis as a useful perspective.

Subsequently, (Bernstein & Newhouse, 2024) and (Pethick et al., 2025) proposed analyzing optimizer convergence under abstract norms using LMOs. This framework generalizes several existing methods, including Normalized SGD with momentum (Cutkosky & Mehta, 2020), Sign−SGD with momentum (Sun et al., 2023), and Muon (Jordan et al., 2024).

Combined approaches. In recent years, parameter-free ideas have also been extended to zeroth-order optimization. The work (Ren & Luo, 2025) initiated this direction with the POEM algorithm. Its stepsize adaptation follows the DoG-type rule (Ivgi et al., 2023), but it uses ZO gradient estimates and retains the Euclidean update geometry. However, their analysis requires a prior estimate of the Lipschitz constant, which weakens the PF nature of the method. Moreover, the scheme requires to store the original model weights, which undermines the memory efficiency. This line was later extended to decentralized optimization by D−POEM (Chen & Rogozin, 2026), and to variance-reduced schemes (Peng et al.). Nevertheless, these methods require $\mathcal{O}(d)$ function evaluations for ZO gradient estimator construction, which is infeasible for LLM fine-tuning. Moreover, their empirical validation is limited to dimensions below those relevant for large-scale models.

Recent work has also explored parameter-free extensions of LMO-based and non-Euclidean methods. In particular, several variants of Muon incorporate Adam-style normalization or adaptive scaling into matrix-aware updates (Si et al., 2025; Li et al., 2025). Parameter-free concepts have also been studied in LMO-based constrained optimization, mostly in the Frank{Wolfe-based methods (Sahu et al., 2019; Carderera et al., 2021; Ito et al., 2023). However, these methods do not address the issue of memory-efficient LLM fine-tuning.

LMO-based ideas have also appeared in zeroth-order optimization. Recent works exploit the matrix structure in the optimization variables to design more geometry-aware ZO methods (Veprikov et al., 2024; Chen et al., 2024; Petrov et al., 2025). However, this line of work does not consider parameter-free challenges.

1.2 Contribution

In light of the related work discussed above, our main contributions are as follows.

- We propose a novel approach for *constructing parameter-free stepsize* in the zeroth-order setting. Our method estimates local smoothness using differences between ZO gradient approximations at neighboring points along the optimization trajectory. This *preserves the memory-efficient nature of ZO optimization*, as it only requires storing a generator seed corresponding to the perturbation direction of ZO gradient.
- Based on the proposed parameter-free estimators, we introduce AdaNAGED, an algorithm that exploits LMO-based updates. As a result, *our method combines three desirable proper-*

ties: memory-efficient zeroth-order optimization, reduced dependence on hyperparameter tuning, and geometry-aware updates.

- We provide *convergence guarantees* for the proposed method considering Lipschitz continuous, smooth and convex objectives. The resulting rate is asymptotically optimal.
- We conduct experiments by fine-tuning the `OPT-1.3B` model on the `SST-2` task. The performance of our method matches the tuned `Sign-SGD`, `Muon` and `AdaMM`.

2 PRELIMINARIES

In this section, we present the preliminary concepts required for constructing and analyzing the desired method. To start with, we define some notation used in the paper. We denote $\|\cdot\|$ and $\|\cdot\|_*$ as an abstract norm and its dual conjugate, respectively, $\mathbb{E}[\cdot]$ represents the mathematical expectation. Additionally, for generality, we introduce the equivalence factor within the chosen norm and the 2-norm: $c_2\|v\| \leq \|v\|_2 \leq C_2\|v\|, c_{2*}\|v\|_* \leq \|v\|_2 \leq C_{2*}\|v\|_* \; \forall v \in \mathbb{R}^d$.

LMO. In this work, we use LMOs over norm balls of arbitrary radius, which allows the update geometry and the update scale to be controlled separately (Pethick et al., 2025). For a given abstract norm $\|\cdot\|$, we define:

$$\mathrm{lmo}(x) = \arg\min_{\|v\|\leq 1} \langle x, v\rangle.$$

This formulation covers a broad class of non-Euclidean update rules by varying the underlying norm. For example, for the Euclidean norm, $\mathrm{lmo}(x) = -\frac{x}{\|x\|_2}$, while for the ℓ_∞-norm, $\mathrm{lmo}(x) = -\cdot sign(x)$.

Zero-order optimization. In the ZO setting, gradients are replaced by finite-difference approximations computed from function-value queries. Given a perturbation direction e and a smoothing parameter $\tau > 0$, we define the ZO gradient estimator as

$$g_\tau(x, e) = \frac{f(x + \tau e) - f(x)}{\tau} e.$$

For the analysis, it is standard to introduce the corresponding smoothed objective $f_\tau(x) = \mathbb{E}[f(x + \tau e)]$, where the expectation is taken over the random perturbation direction e (Flaxman et al., 2004; Nesterov & Spokoiny, 2017). This allows the behavior of ZO methods to be studied in terms of the gradients of the smoothed function rather than the original objective directly.

Next, we list the assumptions regarding the objective.

Assumption 2.1. *The objective f is L-smooth, i.e., for any $x, y \in \mathbb{R}^d$ it implies*

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|.$$

Lipschitzness of the gradient is an inherent aspect of neural network optimization due to their almost-everywhere smooth behavior (Li et al., 2018). Moreover, smoothness is one of the foundational assumptions in optimization theory (Robbins & Monro, 1951; Stich, 2019).

Assumption 2.2. *The objective f is convex, i.e., for any $x, y \in \mathbb{R}^d$ it implies*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle.$$

Although neural networks, and LLMs in particular, are non-convex, they are shown to establish convex-like properties (Kleinberg et al., 2018; Zhou et al., 2019; Liu et al., 2022). This motivates analyzing the convex setting as a principled baseline. Moreover, convex optimization serves as a theoretical foundation for the design of optimization algorithms. For example, momentum (Nesterov et al., 2018) and `AdaGrad` (Duchi et al., 2011) were initially developed and analyzed for convex problems.

Assumption 2.3. *The objective f is M-Lipschitz, i.e., for any $x, y \in \mathbb{R}^d$ it implies*

$$|f(x) - f(y)| \leq M\|x - y\|, \quad \textit{where} \quad M > 0.$$

ZO methods are often analyzed under Lipschitz continuity assumptions, since function-value queries alone require control over how the objective changes under random perturbations. This condition makes it possible to relate finite-difference estimators to the gradient of a smoothed objective and to control the smoothing bias introduced by the perturbation radius τ (Flaxman et al., 2004; Nesterov & Spokoiny, 2017; Duchi et al., 2015).

3 Main Results

Motivation. We begin by motivating how a parameter-free stepsize should be constructed in our setting. To this end, we first examine the stepsize that would be optimal for a ZO LMO-based method if the relevant problem constants were known. As discussed in Section 2, the standard ZO analysis is naturally carried out for the smoothed objective f_τ. Applying the basic descent argument to an LMO-based update with a fixed stepsize γ and smoothing parameter τ yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f_\tau(x^t)\|_* \le \frac{\Delta}{\gamma T} + \frac{\gamma L}{2}.$$

Here, $\Delta = f_\tau(x^0) - f_\tau(x^*)$. Balancing the two terms gives the optimal fixed stepsize

$$\gamma = \sqrt{\frac{\Delta}{LT}}, \quad \text{which leads to the convergence rate } \mathcal{O}\left(\sqrt{\frac{\Delta L}{T}}\right). \tag{5}$$

However, both Δ and L are problem-dependent quantities and are not available in practice. A parameter-free method should therefore avoid requiring them as inputs. Following the standard principle behind parameter-free algorithms, we construct iterative estimates of these constants along the optimization trajectory and use them directly to set the stepsize. This allows the method to adapt within a single run without an external grid search. The central question is thus how to build valid ZO-compatible approximations of Δ and L that preserve the desired convergence guarantees.

Our estimators. We first construct an approximation of the smoothness constant L. Our parameter-free objective is closely related to the approach of Mishkin et al. (2024); Medyakov et al. (2026), who use local smoothness estimates computed along the optimization trajectory. In their setting, this estimate takes the form $L^t = {\|\nabla f^t(x^{t+1}) - \nabla f^t(x^t)\|_*}/{\|x^{t+1} - x^t\|}$. This quantity is natural because smoothness is only invoked between two adjacent iterates; thus, it is sufficient to estimate curvature along the realized update direction. In our ZO setting, however, exact gradient oracles are unavailable. We therefore replace the gradients in the local smoothness estimate with their ZO approximations. Specifically, we define

$$L^t = \frac{\|g^t(x^{t+1}, e^t) - g^t(x^t, e^t)\|_*}{\|x^{t+1} - x^t\|}, \quad \text{where } g^t(x, e) = g_{\tau^t}(x, e). \tag{6}$$

Note that to improve the stability of this estimate, we use the same perturbation direction e^t at both points. Our construction also preserves the memory-efficient nature of ZO optimization: it only requires the corresponding loss differences and the perturbation direction, which can be stored using the seed of a random generator (Malladi et al., 2023).

For the convergence analysis, and also for the numerical stability of the method, such local estimates must remain bounded. In the approach of Mishkin et al. (2024); Medyakov et al. (2026), this follows directly from Assumption 2.1. In contrast, boundedness is not obvious for our estimate (6). We therefore establish the following lemma, which shows that the proposed estimator remains controlled under our assumptions.

Lemma 3.1. *Suppose Assumptions 2.1, 2.2, and 2.3 hold. Then the estimator (6) is bounded:*

$$L^t \le {2L}/{(c_{2*}c_2)},$$

where c_{2*}, c_2 are norm equivalence factors, defined in Section 2.

Theoretically, we need to approximate the factor $\sqrt{LT}$ appearing in the denominator of the optimal stepsize. We therefore aggregate the local smoothness estimates (6) in the manner of `AdaGrad-Norm` (Duchi et al., 2011), leading to the following denominator for the stepsize:

$$S^t = S^{t-1} + L^t, \ \gamma^t \sim {1}/{\sqrt{S^t}}.$$

For the approximation of Δ, a precise estimate is typically not necessary. Instead, it is often sufficient to use a lower bound on the objective. Namely, assume that there exists a known constant $\widetilde{f}$ such that $f(x) \geq \widetilde{f}$ holds for all $x \in \mathbb{R}^d$. Then

$$\Delta = f(x_0) - f(x_*) \leq f(x_0) - \widetilde{f}.$$

Such a bound is readily available in many practical settings: for empirical risk minimization with nonnegative losses, one can take $\widetilde{f} = 0$; similarly, feasibility-type formulations often satisfy $f(x^*) = 0$ (Boyd et al., 2003), making the same lower bound valid. Since our analysis is carried out for the smoothed objective, we also note that the same lower bound is preserved for f_τ. Indeed, $f_\tau(x) = \mathbb{E}[f(x + \tau e)] \geq \mathbb{E}[\widetilde{f}] = \widetilde{f}$. Therefore, the quantity $f(x^0) - \widetilde{f}$ provides a valid upper estimate of the initial gap. Consequently, we use the following stepsize for training:

$$\gamma^t = \frac{\sqrt{f(x^0) - \widetilde{f}}}{\sqrt{\sum_{i=-1}^{t-1} L^i}}.$$

Next, we turn to the choice of the smoothing parameter τ. Rather than introducing an additional tuning rule, we choose τ^t directly from the stability requirement that appears in the proof of Lemma 3.1. In particular, the analysis requires controlling the relation between the stepsize γ^t and the smoothing radius τ^t. The key bound takes the form (see Lemma C.2 for details)

$$L^t \leq L \frac{(\gamma^t)^2 + (\tau^t/c_{2*})^2}{\tau^t \gamma^t},$$

To make this bound uniform, the two terms in the numerator should be balanced, which naturally leads to a linear scaling between τ^t and γ^t. We therefore set

$$\tau^t = \gamma^t c_{2*} = \frac{c_{2*}\sqrt{f(x^0) - \widetilde{f}}}{\sqrt{\sum_{i=0}^{t-1} L^i}}.$$

Algorithm 1 `AdaNAGED`

Input: x^0, $S^{-1} = L^{-1}, c_2, \gamma^0 = \frac{\sqrt{f(x^0)-\widetilde{f}}}{\sqrt{S^{-1}}}, \tau^0 = c_2\gamma^0$.

1: **for** $t = 0, \ldots, T-1$ **do**
2: $e^t \sim S^{d-1}_{\|\cdot\|_2}$ // Sample random seed
3: $g^t(x^t, e^t) = (f(x^t + \tau^t e^t) - f(x^t))\frac{e^t}{\tau^t}$ // Compute ZO-gradient
4: $v^t = \mathrm{lmo}(g^t(x^t, e^t))$ // Compute LMO
5: $x^{t+1} = x^t + \gamma^t v^t$ // Provide model update
6: $g^t(x^{t+1}, e^t) = (f(x^{t+1} + \tau^t e^t) - f(x^{t+1}))\frac{e^t}{\tau^t}$ // Compute ZO-gradient in new point
7: $L^t = \frac{\|g^t(x^{t+1},e^t) - g^t(x^t,e^t)\|_*}{\|x^{t+1}-x^t\|}$ // Approximate L
8: $S^t = S^{t-1} + L^t, \gamma^{t+1} = \frac{\sqrt{f(x^0)-\widetilde{f}}}{\sqrt{S^t}}, \tau^{t+1} = c_{2*}\gamma^{t+1}$ // Update parameters
9: **end for**

Algorithm and theoretical analysis. We present the main algorithmic contribution of the paper: `AdaNAGED` (`ADAptive Norm-Agnostic Gradient-frEe Descent`). Using the estimators constructed above, Algorithm 1 samples a random direction e^t (Line 2), computes the corresponding ZO gradient (Line 3), provides a model update (Lines 4, 5), and computes a ZO approximation in a new point (Line 6) to update γ, τ quantities (Line 8). The initial estimate L^{-1} is calculated from two random points as in Line 7. Now we present the main theoretical result.

Theorem 3.2. *Suppose Assumptions 2.1, 2.2, 2.3 hold. Then Algorithm 1 to achieve ε-accuracy needs*

$$T = \mathcal{O}\left(\frac{\Delta L^3}{c_2^3 c_{2*}^3 \varepsilon^2}\left(1 + \frac{M^2 c_2^2}{\Delta L^{-1}}\right)\mathbb{E}\left[\frac{1}{L^{-1}}\right]^2 \mathbb{E}\left[\log\frac{TL}{L^{-1}}\right]^2 + MC_{2*}\right) \quad \textit{iterations,}$$

where $\varepsilon = \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|\nabla f^t(x^t)\|_*\right]$, *and* $\Delta = f(x^0) - \widetilde{f}$.

Discussion. Although the objective is convex, we formulate the guarantee as convergence to an ε-stationary point of the smoothed objective, which is the standard form of convergence for LMO-based methods (Pethick et al., 2025). The bound preserves the optimal asymptotic dependence $\mathcal{O}(1/\varepsilon^2)$ (Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017), while incurring an additional factor $\mathcal{O}\left(\left(M^2L^2/\Delta(L^{-1})^3\right)\right)$ from using parameter-free estimates instead of the optimal stepsize. The remaining non-vanishing term comes from the stochasticity of randomized finite-difference directions, which is standard in LMO analysis (Bernstein et al., 2018; Veprikov et al., 2024).

3.1 Results for different optimizers

Now we demonstrate how the choice of a particular norm leads to different optimizers. By choosing $\|\cdot\| = \|\cdot\|_\infty, \|\cdot\|_* = \|\cdot\|_1$, Algorithm 1 transforms into a `Sign-SGD`-like method that follows both the ZO and PF paradigms. We can derive exact values for equivalence factors: $c_2 = 1, C_2 = \sqrt{d}, c_{2*} = \frac{1}{\sqrt{d}}, C_{2*} = 1$. Notably, here the optimal choice for ZO smoothing is $\tau^t = \gamma^t$.

Corollary 3.3. *Suppose Assumptions 2.1, 2.2, 2.3 hold. Then Algorithm 1 to achieve ε-accuracy needs*

$$T = \mathcal{O}\left(\frac{\Delta L^3 d^{3/2}}{\varepsilon^2}\left(1 + \frac{M^2}{\Delta L^{-1}}\right)\mathbb{E}\left[\frac{1}{L^{-1}}\right]^2 \mathbb{E}\left[\log\frac{TL}{L^{-1}}\right]^2 + M\right) \quad \textit{iterations,}$$

where $\varepsilon = \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|\nabla f^t(x^t)\|_1\right]$, *and* $\Delta = f^0(x^0) - \widetilde{f}$.

This rate is the same as in Theorem 3.2, but it explicitly clarifies the constants induced by the choice of norm. In this case, the norm-dependent factors yield a $d^{3/2}$ dependence in the leading term. Such dimensional dependence is intrinsic to ZO optimization (Ghadimi & Lan, 2013). Moreover, a comparable and even worse dependence is observed in the ZO LMO analysis in (Veprikov et al., 2024), indicating that the dimensionality in our bound is consistent with existing results.

Next, we consider the matrix domain and select $\|\cdot\| = \|\cdot\|_{sp}, \|\cdot\|_* = \|\cdot\|_{nuc}$, the spectral matrix norm, and the nuclear matrix norm, respectively, as we aim to recover a `Muon`-like method. However, the LMO in the chosen norm results in a polar operator derived from SVD:

$$\mathrm{lmo}(X) = -\texttt{Polar}(X) = -UV^T,$$

where $X = U\Sigma V^T$ is SVD decomposition. This creates a practical implementation challenge due to the cost of computing the SVD. To address this issue, Newton–Schulz steps (N-SS) have been proposed as an efficient alternative (Kim & Oh, 2026). At their core, N-SS apply a matrix polynomial transformation that converges to the polar operator. Since this procedure avoids matrix inversion, it substantially reduces computational cost. Moreover, the approximation error becomes negligible after only a few iterations due to its double-exponential convergence. Accordingly, we modify Algorithm 1 by replacing exact LMO operator (Line 4) with q steps of Newton-Schultz algorithm. Thus, we obtain Algorithm 2. The full description of the algorithm can be found in the appendix D.

The algorithm 2 converges according to the following Theorem. The proof can be found in the appendix D.1.

Theorem 3.4. *Suppose Assumptions 2.1, 2.2, 2.3 hold. Then Algorithm 2 to achieve ε-accuracy needs*

$$T = \mathcal{O}\left(\chi_q^5\left(\frac{\Delta L^3}{c_{2*}^3\varepsilon^2}\left(1 + \frac{M^2}{\Delta L^{-1}}\right)\mathbb{E}\left[\frac{1}{L^{-1}}\right]^2 \mathbb{E}\left[\log\frac{TL}{L^{-1}}\right]^2\right) + \chi_q^2 M\right),$$

iterations, where $\varepsilon = \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|\nabla f^t(x^t)\|_*\right]$, $\Delta = f(x^0) - \widetilde{f}$, *and* $\chi_q = \frac{1}{1-\varepsilon_q}$.

The convergence guarantee mirrors the Theorem 3.2 only introducing extra χ_q^5 factor. As previously discussed, this factor is insignificant in practice.

4 EXPERIMENTS

Setup. We evaluate the proposed parameter-free ZO methods with large-scale language-model. We use LoRA (Hu et al., 2022) to fine-tune the pretrained OPT-1.3B model (Zhang et al., 2022) on the SST-2 sentiment classification task (Socher et al., 2013), which is a standard benchmark for evaluating downstream adaptation of language models. More details are provided in Appendix A.1.

Baselines. We compare against tuned zeroth-order (`ZO-AdaMM` (Chen et al., 2019), `ZO-SignSGD` (Liu et al., 2019), and `ZO-Muon` (Petrov et al., 2025)) and first-order baselines (`Adam`, `Muon`), and PF method in ZO setting (`POEM` (Ren & Luo, 2025)). For our method `AdaNAGED`, we evaluate two instantiations. First, we use the ℓ_∞ geometry, achieving a SignSGD-like LMO update. Second, the spectral setting yields in a Muon-like variant based on the Newton–Schulz approximation of the polar operator.

4.1 EMPIRICAL RESULTS

In Table 1 we report the best accuracy achieved, average memory usage during training, and approximate total training time.

Table 1: Results for fine-tuning OPT-1.3B on SST-2 task.

Method	Best Accuracy	Memory usage (MB)	Wall-clock time (hours)
`Adam`	0.921	9476	35
`ZO-AdaMM`	0.912	1258	60
`ZO-SignSGD`	0.910	1248	60
`ZO-Muon`	0.921	1240	60
`POEM`	0.772	1189	35
`AdaNAGED`$_\infty$	0.924	1183	35
`AdaNAGED`$_{sp}$	0.922	1243	35

Accuracy. The results show that proposed parameter-free methods has the slight edge over the tuned ZO optimizers in terms of the final accuracy. This indicates that the proposed parameter-free adaptation can recover the performance of the baselines.

Training time. Table 1 demonstrates end-to-end training time, including the preliminary runs for the parameter tuning. It is apparent that PF approaches (`AdaNAGED`, `POEM`) have nearly $\times 2$ acceleration in terms of total time compared to the baselines.

Memory usage. To observe the advantages of ZO optimization, we measure the GPU consumption for all the methods, without accounting for the total model size (5025 MB). Table 1 demonstrates the approximate $\times 8$ improvement of ZO methods compared to backpropagation.

Overall, considering the trade-offs between accuracy, GPU usage and running time, the two versions of `AdaNAGED` show the best performance across all the baselines, thus supporting our claims.

5 CONCLUSION

In this paper, we propose the first method for LLM fine-tuning in the intersection of three areas: Zeroth Order, Parameter-Free and Linear Minimization Oracle optimization. Our approach exhibits the benefits of all paradigms, namely, costly gradient computations are avoided, no knowledge about problem constants is assumed, and adaptation to the landscape can be achieved due to the flexible choice of norm. Moreover, we present a modification of the algorithm with Newton-Schultz steps as a more practical variation.

We provide rigorous convergence guarantees for the original method and the modification. Our experiments validate the practical performance of the proposed methods on the established benchmark, showing metrics consistent with the baselines.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We have reviewed all AI-assisted work by proof-reading the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our theoretical analysis is presented in Appendix C. The proposed method is described in detail in Section 3. The details of experimental validation are listed in Appendix A.1, including the hardware and hyperparameter choices. The dataset and the benchmark task are open-sourced.

REFERENCES

Amit Attia and Tomer Koren. Sgd with adagrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning*, pp. 1147–1171. PMLR, 2023.

Amit Attia and Tomer Koren. How free is parameter-free stochastic optimization? *arXiv preprint arXiv:2402.03126*, 2024.

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pp. 404–413. PMLR, 2018.

Lukas Balles, Fabian Pedregosa, and Nicolas Le Roux. The geometry of sign gradient descent. *arXiv preprint arXiv:2002.08056*, 2020.

Noga Bar and Raja Giryes. Zoqo: Zero-order quantized optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pp. 560–569. PMLR, 2018.

Immanuel M Bomze, Francesco Rinaldi, and Damiano Zeffiro. Frank–wolfe and friends: a journey into projection-free first-order optimization methods. *4OR*, 19(3):313–345, 2021.

Stephen Boyd, Lin Xiao, and Almir Mutapcic. Subgradient methods. *lecture notes of EE392o, Stanford University, Autumn Quarter*, 2004(01):175, 2003.

Gábor Braun, Alejandro Carderera, Cyrille W Combettes, Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Sebastian Pokutta. Conditional gradient methods. *arXiv preprint arXiv:2211.14103*, 2022.

Alejandro Carderera, Jelena Diakonikolas, Cheuk Yin Lin, and Sebastian Pokutta. Parameter-free locally accelerated conditional gradients. In *International Conference on Machine Learning*, pp. 1283–1293. PMLR, 2021.

Yair Carmon and Oliver Hinder. Making sgd parameter-free. In *Conference on learning theory*, pp. 2360–2389. PMLR, 2022.

Jiawei Chen and Alexander Rogozin. A parameter-free zeroth-order algorithm for decentralized stochastic convex optimization. *arXiv preprint arXiv:2603.15219*, 2026.

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 *ACM AISec*, 2017.

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. *Advances in neural information processing systems*, 32, 2019.

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

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International conference on machine learning*, pp. 2260–2268. PMLR, 2020.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pp. 1493–1529. PMLR, 2018.

Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by d-adaptation. In *International conference on machine learning*, pp. 7449–7479. PMLR, 2023.

Qi Deng, Guanghui Lan, and Zhenwei Lin. Uniformly optimal and parameter-free first-order methods for convex and function-constrained optimization. *arXiv preprint arXiv:2412.06319*, 2024.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*, 2021.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

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

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.

Tanmay Gautam, Youngsuk Park, Hao Zhou, Parameswaran Raman, and Wooseok Ha. Variance-reduced zeroth-order methods for fine-tuning language models. *arXiv preprint arXiv:2404.08080*, 2024.

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

Aman Gupta, Rafael Celente, Abhishek Shivanna, DT Braithwaite, Gregory Dexter, Shao Tang, Hiroto Udagawa, Daniel Silva, Rohan Ramanath, and S Sathiya Keerthi. Effective quantization of muon optimizer states. *arXiv preprint arXiv:2509.23106*, 2025.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8342–8360, 2020.

Hadrien Hendrikx, Lin Xiao, Sebastien Bubeck, Francis Bach, and Laurent Massoulie. Statistically preconditioned accelerated gradient method for distributed optimization. In *International conference on machine learning*, pp. 4203–4227. PMLR, 2020.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339, 2018.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Feihu Huang, Yuning Luo, and Songcan Chen. Limuon: Light and fast muon optimizer for large models. *arXiv preprint arXiv:2509.14562*, 2025.

Masaru Ito, Zhaosong Lu, and Chuan He. A parameter-free conditional gradient method for composite minimization under hölder condition. *Journal of Machine Learning Research*, 24(166): 1–34, 2023.

Maor Ivgi, Oliver Hinder, and Yair Carmon. Dog is sgd's best friend: A parameter-free dynamic step size schedule. In *International conference on machine learning*, pp. 14465–14499. PMLR, 2023.

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

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International conference on machine learning*, pp. 3252–3261. PMLR, 2019.

Ahmed Khaled, Konstantin Mishchenko, and Chi Jin. Dowg unleashed: An efficient universal parameter-free gradient descent method. *Advances in Neural Information Processing Systems*, 36:6748–6769, 2023.

Gyu Yeol Kim and Min-hwan Oh. Convergence of muon with newton-schulz. *arXiv preprint arXiv:2601.19156*, 2026.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International conference on machine learning*, pp. 2698–2707. PMLR, 2018.

Itai Kreisler, Maor Ivgi, Oliver Hinder, and Yair Carmon. Accelerated parameter-free stochastic optimization. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 3257–3324. PMLR, 2024.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3045–3059, 2021.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025.

Liting Lin, Hansong Ma, Junxiao Wang, and Shiyu Yang. A survey on zeroth-order optimization for machine learning. In *International Conference on Web Information Systems and Applications*, pp. 481–497. Springer, 2025.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

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

Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signsgd via zeroth-order oracle. In *International conference on learning representations*, 2019.

Qijun Luo, Hengxu Yu, and Xiao Li. Badam: A memory efficient full parameter optimization method for large language models. *Advances in Neural Information Processing Systems*, 37: 24926–24958, 2024.

Shaocong Ma and Heng Huang. Revisiting zeroth-order optimization: Minimum-variance two-point estimators and directionally aligned perturbations. *arXiv preprint arXiv:2510.19975*, 2025.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. *arXiv preprint arXiv:1910.09529*, 2019.

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

H Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory*, pp. 1020–1039. PMLR, 2014.

Daniil Medyakov, Stanko Sergey, Gleb Molodtsov, Philip Zmushko, Evseev Grigoriy, Egor Petrov, and Aleksandr Beznosikov. Sign-sgd via parameter-free optimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*, 2023.

Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert M Gower. Directional smoothness and gradient methods: Convergence and adaptivity. *Advances in Neural Information Processing Systems*, 37:14810–14848, 2024.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

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

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

Georgii Sergeevich Novikov, Daniel Bershatsky, Julia Gusak, Alex Shonenkov, Denis Valerievich Dimitrov, and Ivan Oseledets. Few-bit backward: Quantized gradients of activation functions for memory footprint reduction. In *International Conference on Machine Learning*, pp. 26363–26381. PMLR, 2023.

Francesco Orabona. Dimension-free exponentiated gradient. *Advances in Neural Information Processing Systems*, 26, 2013.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Yuxing Peng, Yuanyuan Liu, Fanhua Shang, and Hongying Liu. Parameter-free variance reduced zeroth-order optimization for non-convex problems.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025.

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

Junbin Qiu, Zhengpeng Xie, Xiangda Yan, Yongjie Yang, and Yao Shu. Zeroth-order optimization is secretly single-step policy optimization. *arXiv preprint arXiv:2506.14460*, 2025.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pp. 1–16. IEEE, 2020.

Kunjie Ren and Luo Luo. A parameter-free and near-optimal zeroth-order algorithm for stochastic convex optimization. *arXiv preprint arXiv:2502.05600*, 2025.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making muon & scion great again!(bridging theory and practice of lmo-based optimizers for llms). *arXiv preprint arXiv:2505.13416*, 2025.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1):1–38, 2014.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Raul Rojas. The backpropagation algorithm. In *Neural networks: a systematic introduction*, pp. 149–182. Springer, 1996.

Anit Kumar Sahu, Manzil Zaheer, and Soummya Kar. Towards gradient free and projection free stochastic optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3468–3477. PMLR, 2019.

Fabian Schaipp, Ruben Ohana, Michael Eickenberg, Aaron Defazio, and Robert M Gower. Momo: Momentum models for adaptive learning rates. *arXiv preprint arXiv:2305.07583*, 2023.

Hyunseok Seung, Jaewoo Lee, and Hyunsuk Ko. Low-rank curvature for zeroth-order optimization in llm fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25235–25242, 2026.

Chongjie Si, Debing Zhang, and Wei Shen. Adamuon: Adaptive muon optimizer. *arXiv preprint arXiv:2507.11005*, 2025.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

James C Spall. An overview of the simultaneous perturbation method for efficient optimization. *Johns Hopkins apl technical digest*, 19(4):482–492, 1998.

Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of signsgd under weaker assumptions. In *International Conference on Machine Learning*, pp. 33077–33099. PMLR, 2023.

Zehua Sun, Huanqi Yang, Kai Liu, Zhimeng Yin, Zhenjiang Li, and Weitao Xu. Recent advances in lora: A comprehensive survey. *ACM Transactions on Sensor Networks*, 18(4):1–44, 2022.

Andrey Veprikov, Alexander Bogdanov, Vladislav Minashkin, and Aleksandr Beznosikov. New aspects of black box conditional gradient: Variance reduction and one point feedback. *Chaos, Solitons & Fractals*, 189:115654, 2024.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. Lora-pro: Are low-rank adapters properly optimized? *arXiv preprint arXiv:2407.18242*, 2024.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Xiao-Kun Wu, Min Chen, Wanyi Li, Rui Wang, Limeng Lu, Jia Liu, Kai Hwang, Yixue Hao, Yanru Pan, Qingguo Meng, et al. Llm fine-tuning: Concepts, opportunities, and challenges. *Big Data and Cognitive Computing*, 9(4):87, 2025.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Qining Zhang and Lei Ying. Zeroth-order policy gradient for reinforcement learning from human feedback without reward inference. *arXiv preprint arXiv:2409.17401*, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pp. 270–278, 2020.

Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor W Tsang. Second-order fine-tuning without pain for llms: A hessian informed zeroth-order optimizer. *arXiv preprint arXiv:2402.15173*, 2024.

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. Sgd converges to global minimum in deep learning via star-convex path. *arXiv preprint arXiv:1901.00451*, 2019.

APPENDIX

A ADDITIONAL EXPERIMENTS

A.1 EXPERIMENTAL DETAILS

The experiments are implemented in Python using the PyTorch library (Paszke et al., 2019), leveraging a single CPU (Intel Xeon 2.20 GHz) and a single GPU (NVIDIA A100) for computation.

We report the parameters for the tuned baselines in Table 2. For the baselines, random sampling is performed from the normal distribution $N(0, I)$. In Muon-like methods, 5 Newton-Schultz iterations were performed to approximate LMO.

Table 2: Tuned hyperparameters for all the methods.

Method	Stepsize	ZO smoothing	Momentum
`Adam`	1e-6	1e-3	(0.9, 0.999)
`ZO-AdaMM`	5e-5	1e-3	(0.9, 0.999)
`ZO-SignSGD`	1e-5	1e-3	0.9
`ZO-Muon`	5e-5	1e-3	0.9

A.2 ABLATION STUDIES

Batchsize. We present additional experiments to ensure robustness of the proposed algorithms. First, we explore the varying batchsize for `AdaNAGED` in ℓ_∞-norm. Table 3 demonstrates the natural trade-off: the time-consuming calculations with increased batchsize yield a better performance.

Table 3: Batchsize ablation study.

Batchsize	Final accuracy	Running time (hours)
16	0.657	15
32	0.844	22
64	0.924	28
128	0.929	35

Initial stepsize. Further, we study the influence of the choice of initial stepsize. For this experiment, we run `AdaNAGED` in spectral norm with various starting values of γ. The results are presented in Figure 1. After $\approx$4000 iterations, the stepsize convergences to the similar values, close to the tuned stepsize of `ZO-Muon` (Table 2). This signifies robustness of our stepsize update procedure.

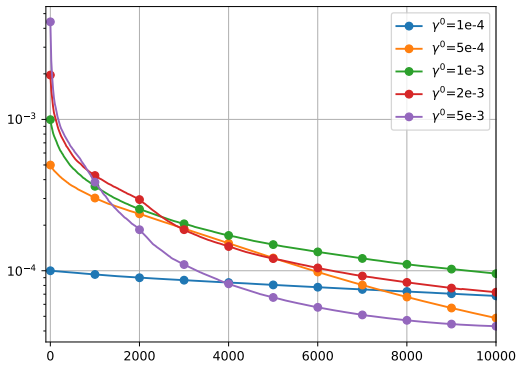

Figure 1: The stepsize decrease across the runs.

B Notation and Useful Equations

In this section we rigorously introduce the notation for the proofs. $\|\cdot\|$ and $\|\cdot\|_*$ denote an abstract norm and its dual conjugate, respectively. We define the following equivalence factors of the chosen norm and the 2-norm:

$$c_2\|v\| \leq \|v\|_2 \leq C_2\|v\|, c_{2*}\|v\|_* \leq \|v\|_2 \leq C_{2*}\|v\|_* \; \forall v \in \mathbb{R}^d. \tag{7}$$

The LMO operator is defined as follows,

$$\text{lmo}(x) = \arg\min_{v\in\mathbb{R}^d, \|v\|\leq 1} \langle x, v\rangle.$$

We define $f_\tau(x) = \mathbb{E}[f(x+\tau e)]$ as a τ-smoothed objective. For the simplicity of notation, $f^t(x)$ is the τ^t-smoothed function at iteration t. The ZO-gradient along the direction e is denoted as

$$g_\tau(x, e) = \frac{f(x+\tau e) - f(x)}{\tau} e.$$

Notably, $\mathbb{E}[g_\tau(x,e)] = \nabla f_\tau(x)$, where the expectation is taken over the random direction e.

Now we additionally introduce some useful equations. Let $x, y, \{a_i\}_{i=1}^n \in \mathbb{R}^d$, $\psi, \xi \in \mathbb{R}_+$ are non-negative random variables, φ satisfies Assumption 2.2. Then

$$\begin{aligned}
\phi(y) &\geq \varphi(x) + \langle \nabla\varphi(x), y - x\rangle, && \text{(Conv)}\\
\left\|\sum_{i=1}^n a_i\right\|^2 &\leq n\sum_{i=1}^n \|a_i\|^2, && \text{(CS)}\\
\langle x, y\rangle &\leq \|x\| \cdot \|y\|_*, && \text{(Conj)}\\
\|x+y\| &\leq \|x\| + \|y\|, && \text{(Norm)}\\
\mathbb{E}[\psi\xi] &\leq (\mathbb{E}[\psi^p])^{1/p} (\mathbb{E}[\xi^q])^{1/q}, \; \frac{1}{p} + \frac{1}{q} = 1, && \text{(Hol)}
\end{aligned}$$

C Proofs for Algorithm 1

First, we need to prove a Numerical Lemma that is used further.

Lemma C.1. *Let $\{a^t\}_{t=0}^T$ be a bounded sequence of positive values $\forall t = 1, \dots T: \; a^t \in (0, A)$. Then,*

$$\sum_{t=0}^{T-1} \frac{a^{t+1}}{\sum_{i=0}^t a^i} \leq \frac{2A}{a^0} + 2\log\frac{AT}{a^0}.$$

Proof. Since the sequence is bounded, there exists such index r that $\sum_{i=0}^{r-2} a^i \leq a^{r-1}$, and $\sum_{i=0}^{t} a^i \geq a^{t-1}$ for any $t \geq r-1$. We refer to that property as $(\star)$. First, we divide the sum over the r index:

$$\sum_{t=0}^{T-1} \frac{a^{t+1}}{\sum_{i=0}^{t} a^i} = \sum_{t=0}^{r-1} \frac{a^{t+1}}{\sum_{i=0}^{t} a^i} + \sum_{t=r}^{T-1} \frac{a^{t+1}}{\sum_{i=0}^{t} a^i}. \tag{8}$$

Now, we estimate two sums separately, starting with the first one. We trivially bound $\sum_{i=0}^{t} a^i \geq a^0$ and proceed with

$$\sum_{t=0}^{r-1} \frac{a^{t+1}}{\sum_{i=0}^{t} a^i} \leq \sum_{t=0}^{r-1} \frac{a^{t+1}}{a^0} = \frac{1}{a^0}\left(\sum_{t=0}^{r-2} a^{t+1} + a^{r-1}\right) \overset{(\star)}{\leq} \frac{1}{a^0} \cdot 2a^{r-1} \leq \frac{2A}{a^0}.$$

Now, we turn our attention to the second sum in 8.

$$\begin{aligned}
\sum_{t=r}^{T-1} \frac{a^{t+1}}{\sum_{i=0}^{t} a^i} &= \sum_{t=r}^{T-1} \frac{a^{t+1}}{\frac{1}{2}\sum_{i=0}^{t} a^i + \frac{1}{2}\sum_{i=0}^{t} a^i} \\
&\overset{(\star)}{\leq} \sum_{t=r}^{T-1} \frac{a^{t+1}}{\frac{1}{2}\sum_{i=0}^{t} a^i + \frac{1}{2}a^{t+1}}. \\
&= \sum_{t=r}^{T-1} \frac{a^{t+1}}{\frac{1}{2}\sum_{i=0}^{t+1} a^i} \\
&\leq \sum_{t=0}^{T-1} \frac{2a^{t+1}}{\sum_{i=0}^{t+1} a^i}.
\end{aligned}$$

Next, we introduce $s^t = \sum_{i=0}^{t} a^i$ and use it to rewrite terms of the sum:

$$\frac{2a^{t+1}}{\sum_{i=0}^{t+1} a^i} = 2\frac{s^{t+1} - s^t}{s^{t+1}} = 2\int_{s^t}^{s^{t+1}} \frac{dx}{s^{t+1}} \overset{(i)}{\leq} 2\int_{s^t}^{s^{t+1}} \frac{dx}{x}.$$

Here, in (i), we used $x \leq s^{t+1}$ in the $(s^t; s^{t+1})$ interval. Then, the integrals compose a telescopic sum:

$$\sum_{t=0}^{T-1} \frac{2a^{t+1}}{\sum_{i=0}^{t+1} a^i} \leq \sum_{t=0}^{T-1} 2\int_{s^t}^{s^{t+1}} \frac{dx}{x} = 2\int_{s^0}^{s^T} \frac{dx}{x} = 2\log\left(\frac{s^T}{s^0}\right) \leq 2\log\left(\frac{AT}{a^0}\right).$$

Now, plugging the acquired estimates into 8, we achieve

$$\sum_{t=0}^{T-1} \frac{a^{t+1}}{\sum_{i=0}^{t} a^i} = \frac{2A}{a^0} + 2\log\left(\frac{AT}{a^0}\right). \tag{9}$$

□

The following Lemma is the cornerstone of the proof. It gives a provable bound for the smoothness estimator L^t and suggests the optimal choice of τ^t.

Lemma C.2. *Suppose Assumptions 2.1, 2.2, and 2.3 hold. Then the estimator equation 6 is uniformly bounded:*

$$L^t \leq \frac{L}{c_2 c_{2*}},$$

where c_{2*}, c_2 are norm equivalence factors.

Proof. We begin be using the definition of ZO-gradients.

$$\begin{aligned}
&\|g^t(x^{t+1}, e^t) - g^t(x^t, e^t)\|_* = \\
&\quad = \|\frac{1}{\tau^t}(f(x^{t+1} + \tau^t e^t) - f(x^{t+1}))e^t - \frac{1}{\tau^t}(f(x^t + \tau^t e^t) - f(x^t))e^t\|_*
\end{aligned}$$

$$= \frac{1}{\tau^t} |f(x^{t+1} + \tau^t e^t) - f(x^{t+1}) - f(x^t + \tau^t e^t) - f(x^t)| \cdot \|e^t\|_*.$$

The core of the lemma is to estimate the loss difference. For the simplicity of the notation, we define $z = \frac{1}{2}(x^{t+1} + x^t + \tau^t e^t),\ u = \frac{1}{2}(x^{t+1} - x^t),\ v = \frac{\tau^t}{2} e^t$.
Thus,

$$|f(x^{t+1} + \tau e^t) - f(x^{t+1}) - f(x^t + \tau e^t) + f(x^t)| =$$
$$= |f(z + u + v) - f(z + u - v) - f(z - u + v) + f(z - u - v)|.$$

Now, we introduce four auxiliary functions $\phi_{\pm\pm} : \mathbb{R} \to \mathbb{R},\ \phi_{\pm\pm}(s) = f(z + s(\pm u \pm v))$.
For $\phi_{++}(s) = f(z + s(u + v))$ we write the Tailor series at the point 0 to the second order with the Lagrange remainder ($\xi_{++} \in (0; 1)$):

$$\begin{aligned}
\phi_{++}(1) &= \phi_{++}(0) + (1-0)\phi'_{++}(0) + \frac{1}{2}(1-0)^2 \phi''_{++}(\xi_{++}) \\
&= f(z) + \langle \nabla f(z), u + v \rangle + \frac{1}{2}(u+v)^T \nabla^2 f(z + \xi_{++}(u+v))(u+v) \\
&= f(z) + \langle \nabla f(z), u + v \rangle + \frac{1}{2}(u+v)^T \nabla^2 f(z_{++})(u+v),
\end{aligned}$$

where $z_{\pm\pm} = z + \xi_{\pm\pm}(\pm u \pm v)$.

Analogously, we derive the same series for all functions $\phi_{\pm\pm}(s) = f(z + s(\pm u \pm v))$. Then, we put the acquired equations into the loss difference.

$$\begin{aligned}
& |f(z+u+v) - f(z+u-v) - f(z-u+v) + f(z-u-v)| \\
= \ & |\phi_{++}(1) - \phi_{+-}(1) - \phi_{-+}(1) + \phi_{--}(1)| \\
= \ & \left| \left(f(z) + \langle \nabla f(z), u+v \rangle + \frac{1}{2}(u+v)^T \nabla^2 f(z_{++})(u+v) \right) \right. \\
& - \left(f(z) + \langle \nabla f(z), u-v \rangle + \frac{1}{2}(u-v)^T \nabla^2 f(z_{+-})(u-v) \right) \\
& - \left(f(z) + \langle \nabla f(z), -u+v \rangle + \frac{1}{2}(-u+v)^T \nabla^2 f(z_{-+})(-u+v) \right) \\
& \left. + \left(f(z) + \langle \nabla f(z), -u-v \rangle + \frac{1}{2}(-u-v)^T \nabla^2 f(z_{+-})(-u-v) \right) \right| \\
= \ & \Big| f(z) - f(z) - f(z) + f(z) \\
& + \langle \nabla f(z), (u+v) - (u-v) - (v-u) - (u+v) \rangle \\
& + \frac{1}{2}(u+v)^T \nabla^2 f(z_{++})(u+v) \\
& - \frac{1}{2}(u-v)^T \nabla^2 f(z_{+-})(u-v) \\
& - \frac{1}{2}(-u+v)^T \nabla^2 f(z_{-+})(-u+v) \\
& \left. + \frac{1}{2}(u+v)^T \nabla^2 f(z_{--})(u+v) \right| \\
= \ & \left| \frac{1}{2}(u+v)^T \nabla^2 f(z_{++})(u+v) \right. \\
& - \frac{1}{2}(u-v)^T \nabla^2 f(z_{+-})(u-v) \\
& - \frac{1}{2}(-u+v)^T \nabla^2 f(z_{-+})(-u+v) \\
& \left. + \frac{1}{2}(u+v)^T \nabla^2 f(z_{--})(u+v) \right|.
\end{aligned}$$

Due to the symmetry, zeroth and first order terms vanish. Now, we estimate each term with $(\pm u \pm v)^T \nabla f(z)(\pm u \pm v) \leq L \| \pm u \pm v\|^2$. due to the Assumption 2.1. Therefore

$$|f(z+u+v) - f(z+u-v) - f(z-u+v) + f(z-u-v)|$$

$$
\begin{aligned}
&\leq 4 \cdot \frac{1}{2} \max\{\|u+v\|^2, \|u-v\|^2\} L \\
&\leq 2L(\|u\| + \|v\|)^2 \\
&\overset{(CS)}{\leq} 4L \left(\left\| \frac{\gamma^t \operatorname{lmo}(g^t(x^t, e^t))}{2} \right\|^2 + \left\| \frac{\tau^t e^t}{2} \right\|^2 \right) \\
&\leq \left(\left(\gamma^t \| \operatorname{lmo}(g^t(x^t, e^t))\| \right)^2 + \left(\tau^t \|e^t\|\right)^2 \right) L \\
&\leq ((\gamma^t)^2 + (\tau^t/c_2)^2) L.
\end{aligned}
$$

Here, we put $\| \operatorname{lmo}(g^t(x^t, e^t))\| = 1$, and $\|e^t\| \leq \frac{\|e^t\|_2}{c_2} = \frac{1}{c_2}$ (7). Now, we derive the final bound for the L^t.

$$
\begin{aligned}
L^t &= \frac{\|g^t(x^{t+1}, e^t) - g^t(x^t, e^t)\|_*}{\|x^{t+1} - x^t\|} \\
&= \frac{|f(z+u+v) - f(z+u-v) - f(z-u+v) + f(z-u-v)| \cdot \|e^t\|_*}{\gamma^t \tau^t \| \operatorname{lmo}(g^t(x^t, e^t))\|} \\
&\leq \frac{\|e\|_*}{\gamma^t \tau^t \| \operatorname{lmo}(g^t(x^t, e^t)\|} \cdot L((\gamma^t)^2 + (\tau^t/c_2)^2).
\end{aligned}
$$

In order to acquire a uniform bound, γ^t and τ^t should be linearly related. Plugging $\tau^t = k\gamma^t$, we arrive at

$$
\begin{aligned}
L^t &\leq \frac{\|e\|_*}{\gamma^t \tau^t \| \operatorname{lmo}(g^t(x^t, e^t)\|} \cdot L((\gamma^t)^2 + (k\gamma^t/c_2)^2) \\
&\overset{(7)}{\leq} \frac{L\|e^t\|_2}{k c_{2*}} (1 + k^2/c_2^2).
\end{aligned}
$$

If e is sampled from the $\|\cdot\|_2$-sphere and k is set to the minimizer c_2, we end up with the bound denoted as $\hat{L}$.

$$
L^t \leq \frac{2L}{c_2 c_{2*}} =: \hat{L}.
$$

□

Theorem C.3. ***[Theorem 3.2]*** *Suppose Assumptions 2.1, 2.2, 2.3 hold. Then Algorithm 1 to achieve ε-accuracy needs*

$$
T = \mathcal{O}\left(\frac{C_2^3 \Delta L^3}{c_{2*}^3 \varepsilon^2} \left(1 + \frac{M^2 C_2^2}{\Delta L^{-1}}\right) \mathbb{E}\left[\frac{1}{L^{-1}}\right]^2 \mathbb{E}\left[\log \frac{TL}{L^{-1}}\right]^2 + M C_{2*} \right) \quad \textit{iterations,}
$$

where $\varepsilon = \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\gamma^t}{\sum_{i=0}^{T-1} \gamma^i} \|\nabla f^t(x^t)\|_* \right]$, *and* $\Delta = f(x^0) - \widetilde{f}$.

Proof. We begin with the descent split in two terms: the smoothing difference and the point difference, which we analyze separately:

$$
f^{t+1}(x^{t+1}) - f^t(x^t) = (f^{t+1}(x^{t+1}) - f^t(x^{t+1})) + (f^t(x^{t+1}) - f^t(x^t)).
$$

First, we estimate the function difference. Here, we introduce $\mathbb{E}$ as the expectation over the sample vector e from the sphere in the 2-norm, as in the definition of smoothed function.

$$
\begin{aligned}
f^{t+1}(x^{t+1}) - f^t(x^{t+1}) &= \mathbb{E}_e[f(x^{t+1} + \tau^{t+1} e)] - \mathbb{E}_e[f(x^{t+1} + \tau^t e)] \\
&\overset{(Conv)}{\leq} \mathbb{E}_e \langle \nabla f(x^{t+1} + \tau^{t+1} e), (x^{t+1} + \tau^{t+1} e) - (x^{t+1} + \tau^t e) \rangle \\
&\overset{(Conj)}{\leq} \mathbb{E}_e \left[\|\nabla f(x^{t+1} + \tau^{t+1} e)\|_2 \|(\tau^{t+1} - \tau^t) e\|_2 \right] \\
&\leq \mathbb{E}_e[M \cdot |\tau^{t+1} - \tau^t| \cdot \|e\|_2]
\end{aligned}
$$

$$\leq \quad \mathbb{E}_e[Mc_2(\gamma^t - \gamma^{t+1})].$$

We utilized the definition of $\tau^t = c_2\gamma^t$ and the Assumption 2.3 which implies $\|\nabla f(x)\|_2 \leq M$. Next, we write the definition for step size:

$$\begin{aligned}
f^{t+1}(x^{t+1}) - f^t(x^{t+1}) &\leq \mathbb{E}_e[Mc_2(\gamma^t - \gamma^{t+1})] \\
&= Mc_2\left(\frac{A}{\sqrt{\sum_{i=-1}^{t-1} L^i}} - \frac{A}{\sqrt{\sum_{i=-1}^{t} L^i}}\right) \\
&= Mc_2A\frac{\sqrt{\sum_{i=-1}^{t} L^i} - \sqrt{\sum_{i=-1}^{t-1} L^i}}{\sqrt{\sum_{i=-1}^{t} L^i}\sqrt{\sum_{i=-1}^{t-1} L^i}} \\
&= \frac{Mc_2A\left(\sum_{i=-1}^{t} L^i - \sum_{i=-1}^{t-1} L^i\right)}{\sqrt{\sum_{i=-1}^{t} L^i}\sqrt{\sum_{i=-1}^{t-1} L^i}(\sqrt{\sum_{i=-1}^{t} L^i} + \sqrt{\sum_{i=-1}^{t-1} L^i})} \\
&= \frac{Mc_2}{A} \cdot \frac{\gamma^t\gamma^{t+1}L^t}{\sqrt{\sum_{i=-1}^{t} L^i} + \sqrt{\sum_{i=-1}^{t-1} L^i}}.
\end{aligned}$$

After simple arithmetic, we extract the step size back. Now, we trivially estimate $\sqrt{\sum_{i=-1}^{t} L^i} \geq \sqrt{\sum_{i=-1}^{t-1} L^i}$, and $\gamma^{t+1} \leq \gamma^1 = \frac{A}{\sqrt{L^{-1}}}$:

$$\begin{aligned}
f^{t+1}(x^{t+1}) - f^t(x^{t+1}) &\leq \frac{Mc_2}{A} \cdot \frac{\gamma^t\gamma^{t+1}L^t}{\sqrt{\sum_{i=-1}^{t} L^i} + \sqrt{\sum_{i=-1}^{t-1} L^i}} \\
&\leq \frac{Mc_2\gamma^t\gamma^{t+1}L^t}{2A\sqrt{\sum_{i=-1}^{t-1} L^i}} \\
&= \frac{Mc_2(\gamma^t)^2\gamma^{t+1}L^t}{2A^2} \\
&\leq \frac{Mc_2(\gamma^t)^2\gamma^1L^t}{2A^2} \\
&= \frac{Mc_2(\gamma^t)^2L^t}{2A\sqrt{L^{-1}}}.
\end{aligned}$$

Next, we analyze the point difference term.

$$\begin{aligned}
f^t(x^{t+1}) - f^t(x^t) &\overset{(Conv)}{\leq} \langle \nabla f^t(x^{t+1}), x^{t+1} - x^t\rangle \\
&= \langle g^t(x^t, e^t), x^{t+1} - x^t\rangle \\
&\quad + \langle \nabla f^t(x^{t+1}) - g^t(x^t, e^t), x^{t+1} - x^t\rangle \\
&\overset{(Conj)}{\leq} \gamma^t\langle g^t(x^t, e^t), \mathrm{lmo}(g^t(x^t, e^t))\rangle \\
&\quad + \|\nabla f^t(x^{t+1}) - g^t(x^t, e^t)\|_*\|x^{t+1} - x^t\|.
\end{aligned}$$

Next, we use the main LMO property:

$$\|v\|_* = \max_{\|u\|\leq 1} \langle u, v\rangle = -\langle v, \mathrm{lmo}(v)\rangle,$$

applying it to the first term:

$$\begin{aligned}
f^t(x^{t+1}) - f^t(x^t) &\leq -\gamma^t\|g^t(x^t, e^t)\|_* \\
&\quad + \|\nabla f^t(x^{t+1}) - g^t(x^t, e^t)\|_*\|x^{t+1} - x^t\| \\
&\overset{(Norm)}{\leq} -\gamma^t\|g^t(x^t, e^t)\|_*
\end{aligned}$$

$$
\begin{aligned}
& & &+\gamma^t\|\nabla f^t(x^{t+1}) - g^t(x^{t+1},e^t)\|_*\|\operatorname{lmo}(g^t(x^t,e^t))\| \\
& & &+\|g^t(x^{t+1},e^t) - g^t(x^t,e^t)\|_*\|x^{t+1}-x^t\| \\
&\leq & &-\gamma^t\|\nabla f^t(x^t)\|_* \\
& & &+\gamma^t\|\nabla f^t(x^t) - g^t(x^t,e^t)\|_* \\
& & &+\gamma^t\|\nabla f^t(x^{t+1}) - g^t(x^{t+1},e^t)\|_* \\
& & &+\left(\frac{\|g^t(x^{t+1},e^t) - g^t(x^t,e^t)\|_*}{\|x^{t+1}-x^t\|}\right)\|x^{t+1}-x^t\|^2 \\
&= & &-\gamma^t\|\nabla f^t(x^t)\|_* \\
& & &+\gamma^t\|\nabla f^t(x^t) - g^t(x^t,e^t)\|_* \\
& & &+\gamma^t\|\nabla f^t(x^{t+1}) - g^t(x^{t+1},e^t)\|_* \\
& & &+L^t(\gamma^t)^2.
\end{aligned}
$$

Then, we add together two acquired estimates:

$$
\begin{aligned}
f^{t+1}(x^{t+1}) - f^t(x^t) &\leq \frac{Mc_2(\gamma^t)^2L^t}{2A\sqrt{L^{-1}}} \\
&\quad -\gamma^t\|\nabla f^t(x^t,e^t)\|_* + \gamma^t\|\nabla f^t(x^t) - g^t(x^t,e^t)\|_* \\
&\quad +\gamma^t\|\nabla f^t(x^{t+1}) - g^t(x^{t+1},e^t)\|_* \\
&\quad +L^t(\gamma^t)^2 \\
&= -\gamma^t\|\nabla f^t(x^t,e^t)\|_* \\
&\quad +\gamma^t\|\nabla f^t(x^t) - g^t(x^t,e^t)\|_* \\
&\quad +\gamma^t\|\nabla f^t(x^{t+1}) - g^t(x^{t+1},e^t)\|_* \\
&\quad +L^t(\gamma^t)^2(1+\frac{Mc_2}{2A\sqrt{L^{-1}}}).
\end{aligned}
$$

Summing over $t = 0, \dots T-1$, we arrive at

$$
\begin{aligned}
f^T(x^T) - f^0(x^0) &\leq -\sum_{t=0}^{T-1}\gamma^t\|\nabla f^t(x^t)\|_* \\
&\quad +(1+\frac{Mc_2}{2A\sqrt{L^{-1}}})\sum_{t=0}^{T-1}L^t(\gamma^t)^2 \\
&\quad +\sum_{t=0}^{T-1}\gamma^t\|g^t(x^{t+1},e^t) - \nabla f^t(x^{t+1})\|_* \\
&\quad +\sum_{t=0}^{T-1}\gamma^t\|g^t(x^t,e^t) - \nabla f^t(x^t)\|_*.
\end{aligned}
$$

Now, we take the expectation and rearrange the terms.

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma^t}{\sum_{i=-1}^{T-1}\gamma^i}\|\nabla f^t(x^t)\|_*\right] &\leq (f^0(x^0) - \widetilde{f})\mathbb{E}\left[\frac{1}{\sum_{i=0}^{T-1}\gamma^i}\right] \\
&\quad +\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|g^t(x^{t+1},e^t) - \nabla f^t(x^{t+1})\|_*\right] \\
&\quad +\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|g^t(x^t,e^t) - \nabla f^t(x^t)\|_*\right] \\
&\quad +\left(1+\frac{Mc_2}{2A\sqrt{L^{-1}}}\right)\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{L^t(\gamma^t)^2}{\sum_{i=0}^{T-1}\gamma^i}\right].
\end{aligned}
$$

Here, $f^0(x^0) - f^T(x^T) \leq f^0(x^0) - \widetilde{f}$ as discussed in 3. Now, let us analyze the second term.

$$\begin{aligned}\|g^t(x^{t+1}, e^t) - \nabla f^t(x^{t+1})\|_* &\overset{(Norm)}{\leq} \|g^t(x^{t+1}, e^t)\|_* + \|\nabla f^t(x^{t+1})\|_* \\ &\overset{7}{\leq} C_{2*}(\|g^t(x^{t+1}, e^t)\|_2 + \|\nabla f^t(x^{t+1})\|_2) \\ &\leq 2MC_{2*}.\end{aligned}$$

This bound is uniform and can be applied under the expectation, yielding

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|g^t(x^{t+1}, e^t) - \nabla f^t(x^{t+1})\|_*\right] &\leq \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\right] \cdot 2MC_{2*} \\ &= \mathbb{E}[1] \cdot 2MC_{2*} \\ &= 2MC_{2*}.\end{aligned}$$

Next, we apply Holder inequality (Hol) to the last term with $p = q = 2$:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|\nabla f^t(x^t)\|_1\right] &\leq \Delta\mathbb{E}\left[\frac{1}{\sum_{i=0}^{T-1}\gamma^i}\right] \\ &\quad + 4MC_{2*} \\ &\quad + \left(1 + \frac{Mc_2}{2A\sqrt{L^{-1}}}\right)\left(\mathbb{E}\left[\frac{1}{\sum_{i=0}^{T-1}\gamma^i}\right]^2\right)^{1/2} \\ &\quad \cdot \left(\mathbb{E}\left[\sum_{t=0}^{T-1} L^t(\gamma^t)^2\right]^2\right)^{1/2}.\end{aligned}$$

To estimate the sum with $L^t(\gamma^t)^2$, we apply the Numerical Lemma C.1. It is justified since the bound $\hat{L}$ is derived for L^t in Lemma C.2

$$\sum_{t=0}^{T-1} L^t(\gamma^t)^2 = \sum_{t=0}^{T-1} \frac{A^2 L^t}{\sum_{i=0}^{t-1} L^i} \overset{(LemC.1)}{\leq} 2A^2\frac{\hat{L}}{L^{-1}} + 4A^2 \log\frac{T\hat{L}}{L^{-1}}.$$

Next, considering $\sum_{i=1}^{t}\frac{1}{\sqrt{i}} \geq \frac{\sqrt{t}}{2}$, we estimate

$$\frac{1}{\sum_{i=0}^{T-1}\gamma^i} = \frac{1}{\sum_{i=0}^{T-1}\sqrt{\frac{A}{\sum_{j=0}^{i-1}L^j}}} \overset{(Lem.C.2)}{\leq} \frac{1}{\sum_{i=0}^{T-1}\frac{A}{\sqrt{\hat{L}}\sqrt{i+1}}} \leq \frac{2\sqrt{\hat{L}}}{A\sqrt{T}}.$$

Now, we combine all of the above to derive the following:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|\nabla f^t(x^t)\|_*\right] &\leq \frac{(f^0(x^0) - \widetilde{f})\sqrt{\hat{L}}}{A\sqrt{T}} + 2MC_{2*} \\ &\quad + \frac{2\sqrt{\hat{L}}}{A\sqrt{T}}\left(1 + \frac{Mc_2}{2A\sqrt{L^{-1}}}\right) \\ &\quad \cdot \left(\mathbb{E}\left[2A^2\frac{\hat{L}}{L^{-1}} + 2A^2\log\frac{T\hat{L}}{L^{-1}}\right]^2\right)^{1/2} \\ &\leq \frac{(f^0(x^0) - \widetilde{f})\sqrt{\hat{L}}}{A\sqrt{T}} + 2MC_{2*} \\ &\quad + \frac{4\sqrt{\hat{L}}}{\sqrt{T}}\left(A + \frac{Mc_2}{2\sqrt{L^{-1}}}\right)\left(\mathbb{E}\left[\frac{\hat{L}}{L^{-1}}\right]^2\right)^{1/2}\end{aligned}$$

$$
\begin{aligned}
&+\frac{4\sqrt{\hat{L}}}{\sqrt{T}}\left(A+\frac{Mc_2}{2\sqrt{L^{-1}}}\right)\left(\mathbb{E}\left[\log\frac{T\hat{L}}{L^{-1}}\right]^2\right)^{1/2}\\
\leq\ & \frac{(f^0(x^0)-\widetilde{f})\sqrt{\hat{L}}}{A\sqrt{T}}+2MC_{2*}\\
&+\frac{8\hat{L}^{\frac{3}{2}}}{\sqrt{T}}\left(A+\frac{Mc_2}{2\sqrt{L^{-1}}}\right)\\
&\cdot\left(\mathbb{E}\left[\frac{1}{L^{-1}}\right]^2\right)^{1/2}\cdot\left(\mathbb{E}\left[\log\frac{T\hat{L}}{L^{-1}}\right]^2\right)^{1/2}.
\end{aligned}
$$

Setting $A=\sqrt{f^0(x^0)-\widetilde{f}}$ would lead to the end of the proof. However, $f^0(x^0)$ is inaccessible for the method. Thus, we set $A=\sqrt{f(x^0)-\widetilde{f}}$ and account for the difference:

$$
\begin{aligned}
f^0(x^0)-\widetilde{f} &= \mathbb{E}_e[f(x^0+\tau^0 e)]-\widetilde{f}\\
&= \mathbb{E}_e[f(x^0+\tau^0 e)-f(x^0)]+f(x^0)-\widetilde{f}\\
&\overset{(2.3)}{\leq} M\mathbb{E}_e[\|x^0+\tau^0 e-x^0\|]+f(x^0)-\widetilde{f}\\
&= \tau^0 M\mathbb{E}_e[\|e\|]+f(x^0)-\widetilde{f}\\
&\overset{(7)}{\leq} c_2\gamma^0 M\frac{1}{c_2}+f(x^0)-\widetilde{f}\\
&= \frac{MA}{\sqrt{L^{-1}}}+f(x^0)-\widetilde{f}.
\end{aligned}
$$

We plug the value for A into the equation:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|\nabla f^t(x^t)\|_*\right] &\leq \frac{(f^0(x^0)-\widetilde{f})\sqrt{\hat{L}}}{A\sqrt{T}}+2MC_{2*}\\
&\quad+\frac{8\hat{L}^{\frac{3}{2}}}{\sqrt{T}}\left(A+\frac{Mc_2}{2\sqrt{L^{-1}}}\right)\\
&\quad\cdot\left(\mathbb{E}\left[\frac{1}{L^{-1}}\right]^2\right)^{1/2}\cdot\left(\mathbb{E}\left[\log\frac{T\hat{L}}{L^{-1}}\right]^2\right)^{1/2}\\
&\leq \frac{(\frac{MA}{\sqrt{L^{-1}}}+f(x^0)-\widetilde{f})\sqrt{\hat{L}}}{A\sqrt{T}}+2MC_{2*}\\
&\quad+\frac{8\hat{L}^{\frac{3}{2}}}{\sqrt{T}}\left(A+\frac{Mc_2}{2\sqrt{L^{-1}}}\right)\\
&\quad\cdot\left(\mathbb{E}\left[\frac{1}{L^{-1}}\right]^2\right)^{1/2}\cdot\left(\mathbb{E}\left[\log\frac{T\hat{L}}{L^{-1}}\right]^2\right)^{1/2}\\
&\leq \frac{10\sqrt{(f(x^0)-\widetilde{f})\hat{L}^3}}{\sqrt{T}}\left(1+\frac{Mc_2}{2\sqrt{L^{-1}(f(x^0)-\widetilde{f})}}\right)\\
&\quad\cdot\left(\mathbb{E}\left[\frac{1}{L^{-1}}\right]^2\right)^{1/2}\cdot\left(\mathbb{E}\left[\log\frac{T\hat{L}}{L^{-1}}\right]^2\right)^{1/2}\\
&\quad+2MC_{2*}.
\end{aligned}
$$

The regrouping the terms with $\varepsilon = \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|\nabla f^t(x^t)\|_*\right]$ concludes the proof. □

D ADANAGED IN SPECTRAL NORM

In this context, the norms are related as follows:

$$\|X\| \leq \|X\|_F \leq \|X\|_* \leq \sqrt{\min\{d_1, d_2\}}\|X\|_F.$$

Algorithm 2 Algorithm 2

Input: $X^0 \in \mathbb{R}^{d_1 \times d_2}$, $S^{-1} = L^{-1}, \gamma^0 = \frac{\sqrt{f(X^0)-\widetilde{f}}}{\sqrt{S^{-1}}}, \tau^0 = \gamma^0$.

1: **for** $t = 0, \ldots, T-1$ **do**
2: $E^t \sim S^{d-1}_{\|\cdot\|_F}$ // Sample random seed
3: $G^t(X^t, E^t) = (f(X^t + \tau^t E^t) - f(X^t))\frac{E^t}{\tau^t}$ // Compute ZO-gradient
4: $V^t = \mathrm{NSS}^q(G^t(X^t, E^t))$ // Compute N-S steps
5: $X^{t+1} = X^t + \gamma^t V^t$ // Provide model update
6: $G^t(X^{t+1}, E^t) = (f(X^{t+1} + \tau^t E^t) - f(X^{t+1}))\frac{E^t}{\tau^t}$ // Compute ZO-gradient in new point
7: $L^t = \frac{\|G^t(X^{t+1},E^t) - G^t(X^t,E^t)\|_*}{\|X^{t+1}-X^t\|}$ // Approximate L
8: $S^t = S^{t-1} + L^t, \gamma^{t+1} = \frac{\sqrt{f(x^0)-\widetilde{f}}}{\sqrt{S^t}}, \tau^{t+1} = \gamma^{t+1}$ // Update parameters
9: **end for**
10: **Output:** $\frac{1}{R}\sum_{r=0}^{R-1} X^r$.

Now we provide main convergence result.

Theorem D.1. *Suppose Assumptions 2.1, 2.2, 2.3 hold. Then Algorithm 2 to achieve ε-accuracy needs*

$$T = \mathcal{O}\left(\chi_q^5\left(\frac{\Delta L^3}{c_{2*}^3\varepsilon^2}\left(1 + \frac{M^2}{\Delta L^{-1}}\right)\mathbb{E}\left[\frac{1}{L^{-1}}\right]^2\mathbb{E}\left[\log\frac{TL}{L^{-1}}\right]^2\right) + \chi_q^2 M\right),$$

iterations, where $\varepsilon = \mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|\nabla f^t(X^t)\|_*\right]$, $\Delta = f(X^0) - \widetilde{f}$, and $\chi_q = \frac{1}{1-\varepsilon_q}$. With chosen norms $c_{2*} = \frac{1}{\sqrt{\min\{d_1,d_2\}}}$.

Proof. The fundamentally, the proof stays the same, we only need to account for extra error brought be N-S steps. Following Kim & Oh (2026), we denote

$$\varepsilon_q = \max_{t=1\ldots T}\|V^t - \mathrm{lmo}(G^t)\| = \max_{t=1\ldots T}\|\,\text{N-SS}^q(G^t, q) - \mathrm{lmo}(G^t)\|.$$

Here, LMO in the chosen norms becomes a polar factor derived from SVD. The authors showed that $\varepsilon_q \leq 1$ and converges to 0 at a double exponential rate over q.

Firstly, we recover the bound for L^t. The proof of Lemma C.2 can be followed directly to gain the following:

$$\begin{aligned} & |f(X^{t+1} + \tau^t E^t) - f(X^{t+1}) - f(X^t + \tau^t E^t) + f(X^t)| \\ = \; & |f(Z + U + V) - f(Z + U - V) - f(Z - U + V) + f(Z - U - V)| \\ \leq \; & 4L\left(\|U\|^2 + \|V\|^2\right), \end{aligned}$$

where, similarly, we define $Z = \frac{1}{2}(X^{t+1} + X^t + \tau^t E^t)$, $U = \frac{1}{2}(X^{t+1} - X^t)$, $V = \frac{\tau^t}{2}E^t$.

The error emerges in the $\|U\|$ estimate:
$\|U\| = \frac{\gamma^t}{2}\|V^t\| \leq \frac{\|V^t - \mathrm{lmo}(G^t)\| + \|\mathrm{lmo}(G^t)\|}{2}\gamma^t \leq \frac{\gamma^t(1+\varepsilon_q)}{2}$, $\|V\| = \frac{\tau^t}{2}\|E^t\| \leq \frac{\tau^t}{2}\|E^t\|_F = \frac{\tau^t}{2}$.
Thus, we derive

$$\begin{aligned}
L^t &= \frac{\|G^t(X^{t+1}, E^t) - G^t(X^t, E^t)\|_*}{\|X^{t+1} - X^t\|} \\
&= \frac{|f(Z+U+V) - f(Z+U-V) - f(Z-U+V) + f(Z-U-V)| \cdot \|E^t\|_*}{\gamma^t\tau^t\|V^t\|} \\
&\leq \frac{\|E\|_F}{c_{2*}\gamma^t\tau^t\|V^t\|} \cdot ((\gamma^t)^2(1+\varepsilon_q)^2 + (\tau^t)^2)L, \\
&\overset{(7)}{\leq} \frac{1}{c_{2*}\gamma^t\tau^t(-\|\mathrm{lmo}(G^t) - V^t\| + \|\mathrm{lmo}(G^t)\|)} \cdot ((\gamma^t)^2(1+\varepsilon_q)^2 + (\tau^t)^2)L \\
&\leq \frac{4}{c_{2*}\gamma^t\tau^t(1-\varepsilon_q)} \cdot ((\gamma^t)^2 + (\tau^t)^2)L.
\end{aligned}$$

To maintain the uniform bound, γ^t and τ^t should be linearly related. Choosing $\tau^t = \gamma^t$ results in

$$L^t \leq \frac{2}{c_{2*}(1-\varepsilon_q)} =: \hat{L}.$$

Now, we continue with the proof of the Theorem D.1. The first estimates are exactly the same as in Theorem C.3:

$$f^{t+1}(X^{t+1}) - f^t(X^t) = (f^{t+1}(X^{t+1}) - f^t(X^{t+1})) + (f^t(X^{t+1}) - f^t(X^t)).$$

$$f^{t+1}(X^{t+1}) - f^t(X^{t+1}) \leq \frac{M(\gamma^t)^2 L^t}{2A\sqrt{L^{-1}}}.$$

Now, we analyze the other term.

$$\begin{aligned}
f^t(X^{t+1}) - f^t(X^t) &\overset{(Conv)}{\leq} \langle \nabla f^t(X^{t+1}), X^{t+1} - X^t\rangle \\
&= \langle G^t(X^t, E^t), X^{t+1} - X^t\rangle \\
&\quad + \langle \nabla f^t(X^{t+1}) - G^t(X^t, E^t), X^{t+1} - X^t\rangle \\
&\overset{(Conj)}{\leq} -\gamma^t\langle G^t(X^t, E^t), V^t\rangle \\
&\quad + \|\nabla f^t(X^{t+1}) - G^t(X^t, E^t)\|_*\|X^{t+1} - X^t\| \\
&\overset{(Conj)}{\leq} -\gamma^t\langle G^t(X^t, E^t), \mathrm{lmo}(G^t)\rangle \\
&\quad + \gamma^t\|G^t(X^t, E^t)\|_*\|V^t - \mathrm{lmo}(G^t)\| \\
&\quad + \|\nabla f^t(X^{t+1}) - G^t(X^t, E^t)\|_*\|X^{t+1} - X^t\|.
\end{aligned}$$

Here, we utilize the notation of ε_q to derive

$$\begin{aligned}
f^t(X^{t+1}) - f^t(X^t) &\leq -\gamma^t\|G^t(X^t, E^t)\|_* \\
&\quad + \gamma^t\|G^t(X^t, E^t)\|_*\varepsilon_q \\
&\quad + \|\nabla f^t(X^{t+1}) - G^t(X^t, E^t)\|_*\|X^{t+1} - X^t\| \\
&\leq -\gamma^t(1-\varepsilon_q)\|G^t(X^t, E^t)\|_* \\
&\quad + \|\nabla f^t(X^{t+1}) - G^t(X^{t+1}, E^t)\|_*\|X^{t+1} - X^t\| \\
&\quad + \left(\frac{\|G^t(X^{t+1}, E^t) - G^t(X^t, E^t)\|_*}{\|X^{t+1} - X^t\|}\right)\|X^{t+1} - X^t\|^2 \\
&\overset{(Norm)}{\leq} -\gamma^t(1-\varepsilon_q)\|\nabla f^t(X^t)\|_* \\
&\quad + \gamma^t(1-\varepsilon_q)\|G^t(X^t, E^t) - \nabla f^t(X^t)\|_* \\
&\quad + 2\gamma^t\|\nabla f^t(X^{t+1}) - G^t(X^{t+1}, E^t)\|_*
\end{aligned}$$

$$+4L^t(\gamma^t)^2.$$

Here, in the last transition we estimated

$$\|X^{t+1}-X^t\| = \gamma^t\|V^t\| \overset{(Norm)}{\leq} \gamma^t\|\operatorname{lmo}(G^t)\| + \gamma^t\|V^t - \operatorname{lmo}(G^t)\| \leq \gamma^t(1+\varepsilon_q) \leq 2\gamma^t.$$

Next, we combine the estimates together.

$$\begin{aligned} f^{t+1}(X^{t+1}) - f^t(X^t) &\leq \frac{M(\gamma^t)^2L^t}{2A\sqrt{L^{-1}}} \\ &\quad -\gamma^t(1-\varepsilon_q)\|\nabla f^t(X^t)\|_* \\ &\quad +\gamma^t\|G^t(X^t,E^t)-\nabla f^t(X^t)\|_* \\ &\quad +2\gamma^t\|\nabla f^t(X^{t+1})-G^t(X^{t+1},E^t)\|_* \\ &\quad +4L^t(\gamma^t)^2 \\ &\leq -\gamma^t(1-\varepsilon_q)\|\nabla f^t(X^t)\|_* \\ &\quad +\gamma^t(1-\varepsilon_q)\|G^t(X^t,E^t)-\nabla f^t(X^t)\|_* \\ &\quad +2\gamma^t\|\nabla f^t(X^{t+1})-G^t(X^{t+1},E^t)\|_* \\ &\quad +L^t(\gamma^t)^2\left(4+\frac{M}{2A\sqrt{L^{-1}}}\right). \end{aligned}$$

Summing over $t=1,\dots T$ we derive

$$\begin{aligned} f^T(X^T) - f^0(X^0) &\leq -(1-\varepsilon_q)\sum_{t=0}^{T-1}\gamma^t\|\nabla f^t(X^t)\|_* \\ &\quad +\left(4+\frac{M}{2A\sqrt{L^{-1}}}\right)\sum_{t=0}^{T-1}L^t(\gamma^t)^2 \\ &\quad +\sum_{t=0}^{T-1}\gamma^t\|G^t(X^{t+1},E^t)-\nabla f^t(X^{t+1})\|_* \\ &\quad +2\sum_{t=0}^{T-1}\gamma^t\|G^t(X^t,E^t)-\nabla f^t(X^t)\|_*. \end{aligned}$$

After rearranging the terms,

$$\begin{aligned} &(1-\varepsilon_q)\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|\nabla f^t(x^t)\|_*\right] \\ \leq\ &(f^0(x^0)-\widetilde{f})\mathbb{E}\left[\frac{1}{\sum_{i=0}^{T-1}\gamma^i}\right] \\ &+\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|G^t(X^{t+1},E^t)-\nabla f^t(X^{t+1})\|_*\right] \\ &+2\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|G^t(X^t,E^t)-\nabla f^t(X^t)\|_*\right] \\ &+\left(4+\frac{M}{2A\sqrt{L^{-1}}}\right)\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{L^t(\gamma^t)^2}{\sum_{i=0}^{T-1}\gamma^i}\right]. \end{aligned}$$

We then implement the same estimate for the second and third terms, bounding those with $4MC_{2*}$. Next, (Hol) is applied to the last term:

$$(1-\varepsilon_q)\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|\nabla f^t(X^t)\|_*\right] \leq (f^0(X^0)-\widetilde{f})\mathbb{E}\left[\frac{1}{\sum_{i=0}^{T-1}\gamma^i}\right]$$

$$
\begin{aligned}
&+8MC_{2*} \\
&+\left(4+\frac{M}{2A\sqrt{L^{-1}}}\right) \\
&\cdot\left(\mathbb{E}\left[\frac{1}{\sum_{i=0}^{T-1}\gamma^i}\right]^2\right)^{1/2} \\
&\cdot\left(\mathbb{E}\left[\sum_{t=0}^{T-1}L^t(\gamma^t)^2\right]^2\right)^{1/2}.
\end{aligned}
$$

The following estimates are repeated from Theorem C.3

$$
\begin{aligned}
\sum_{t=0}^{T-1}L_\infty^t(\gamma^t)^2 = A^2\sum_{t=0}^{T-1}\frac{L^t}{\sum_{i=0}^{t-1}L_\infty^i} &\leq 2A^2\frac{\hat{L}}{L^{-1}}+2A^2\log\frac{T\hat{L}}{L^{-1}}, \\
\frac{1}{\sum_{i=0}^{T-1}\gamma^i} &\leq \frac{\sqrt{\hat{L}}}{A\sqrt{T}}.
\end{aligned}
$$

Next, we plug those into the equation, introducing $\chi_q = \frac{1}{1-\varepsilon_q}$.

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|\nabla f^t(X^t)\|_*\right] &\leq \frac{\chi_q(f^0(X^0)-\widetilde{f})\sqrt{\hat{L}}}{A\sqrt{T}}+\chi_q M \\
&\quad+\chi_q\left(1+\frac{M}{2A\sqrt{L^{-1}}}\right)\cdot\frac{\sqrt{\hat{L}}}{A\sqrt{T}} \\
&\quad\cdot\left(\mathbb{E}\left[2A^2\frac{\hat{L}}{L^{-1}}+2A^2\log\frac{T\hat{L}}{L^{-1}}\right]^2\right)^{1/2} \\
&\leq \chi_q\Bigg(\frac{(f^0(X^0)-\widetilde{f})\sqrt{\hat{L}}}{A\sqrt{T}}+8M \\
&\quad+\frac{8\hat{L}^{3/2}}{\sqrt{T}}\left(4A+\frac{M}{2\sqrt{L^{-1}}}\right) \\
&\quad\cdot\left(\mathbb{E}\left[\frac{1}{L^{-1}}\right]^2\right)^{1/2}\cdot\left(\mathbb{E}\left[\log\frac{T\hat{L}}{L^{-1}}\right]^2\right)^{1/2}\Bigg).
\end{aligned}
$$

Same way as in the previous theorem, we need estimate the initial function discrepancy $f(x^0) - f^0(x^0)$ due to inability to compute f^0 directly:

$$
\begin{aligned}
f^0(x^0)-\widetilde{f} &= \mathbb{E}_e[f(x^0+\tau^0 e)]-\widetilde{f} \\
&= \mathbb{E}_E[f(X^0+\tau^0 E)-f(X^0)]+f(X^0)-\widetilde{f} \\
&\overset{(2.3)}{\leq} M\mathbb{E}_E[\|X^0+\tau^0E-X^0\|]+f(X^0)-\widetilde{f} \\
&= \tau^0 M\mathbb{E}_e[\|e\|]+f(X^0)-\widetilde{f} \\
&\overset{(7)}{\leq} \tau^0 M\mathbb{E}_e[\|e\|_F]+f(X^0)-\widetilde{f} \\
&= \gamma^0 M+f(X^0)-\widetilde{f} \\
&= \frac{MA}{\sqrt{L^{-1}}}+f(X^0)-\widetilde{f}.
\end{aligned}
$$

Thus, we derive

$$
\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{\gamma^t}{\sum_{i=0}^{T-1}\gamma^i}\|\nabla f^t(X^t)\|_*\right] \leq \chi_q\left(\frac{(f(X^0)-\widetilde{f}+\frac{MA}{\sqrt{L^{-1}}})\sqrt{\hat{L}}}{A\sqrt{T}}+8M\right.
$$

$$+\frac{8\hat{L}^{3/2}}{\sqrt{T}}\left(4A+\frac{M}{2\sqrt{L^{-1}}}\right)$$
$$\cdot\left(\mathbb{E}\left[\frac{1}{L^{-1}}\right]^2\right)^{1/2}\cdot\left(\mathbb{E}\left[\log\frac{T\hat{L}}{L^{-1}}\right]^2\right)^{1/2}\Bigg).$$

After substituting $A=\sqrt{f(x^0)-\widetilde{f}}$ and $\hat{L}$, we arrive at the Theorem statement. □